Advances in Intelligent Systems and Computing 366

Henry Selvaraj
Dawid Zydek
Grzegorz Chmaj *Editors*

Progress in Systems Engineering

Proceedings of the Twenty-Third International Conference on Systems Engineering

Springer

Advances in Intelligent Systems and Computing

For further volumes:
http://www.springer.com/series/11156

Sponsors

www.unlv.edu

www.aldec.com

www.nvenergy.com

Henry Selvaraj • Dawid Zydek • Grzegorz Chmaj
Editors

Progress in Systems Engineering

Proceedings of the Twenty-Third International Conference on Systems Engineering

Springer

Editors
Henry Selvaraj
University of Nevada at Las Vegas
Las Vegas, Nevada, USA

Grzegorz Chmaj
University of Nevada at Las Vegas
Las Vegas, Nevada, USA

Dawid Zydek
Department of Electrical Engineering
Idaho State University
Pocatello, Idaho, USA

ISSN 2194-5357 ISSN 2194-5365 (electronic)
ISBN 978-3-319-08421-3 ISBN 978-3-319-08422-0 (eBook)
DOI 10.1007/978-3-319-08422-0
Springer Cham Heidelberg Dordrecht London New York

Library of Congress Control Number: 2014945639

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This important series of Conferences arose from technical cooperation among the Technical University of Wroclaw (Poland), Coventry Polytechnic (UK), Wright State University (USA) and the University of Nevada, Las Vegas (USA). Prior to 1980, a series of International Conferences on Systems Science had been held in Wroclaw, Poland. In 1980, it was decided that the Conference would change venue on alternate years and be held in Coventry, UK and, at the same time, change emphasis from Science to Engineering when hosted in UK. Consequently, the first and second International Conferences on Systems Engineering were held in Coventry, UK in 1980 and 1982. In 1982, it was decided that the Engineering series of Conferences would be held in the USA each third year. The Third and Fifth International Conferences on Systems Engineering were held in Dayton, Ohio, USA. The Seventh, Ninth, Eleventh, Thirteenth, Fifteenth, Seventeenth, Nineteenth and Twenty First International Conferences on Systems Engineering were held in Las Vegas, Nevada, USA. This year, 2014, the Twenty Third International Conference on Systems Engineering is held in Las Vegas, Nevada, USA.

Research in the discipline of Systems Engineering is an important concept in the advancement of engineering and information sciences. Systems Engineering attempts to integrate many of the traditional engineering disciplines to solve large complex functioning engineering systems, dependent on components from all the disciplines. The research papers contained in these proceedings reflect the state of the art in systems engineering from all over the world and should serve as vital references to researchers to follow.

We received 273 submissions. Each paper was reviewed by at least two independent reviewers. A total of 129 papers were selected as full papers that gives an acceptance rate 47.2%. The University of Nevada, Las Vegas is very pleased to host this Twenty-Third International Conference on Systems Engineering. We would like to thank all the authors, reviewers, participants and student volunteers for making the conference a success. Our special thanks go to Michael Luby, Merry Stuber, and Lesley Poliner from Springer for their patience and help while organizing and preparing the proceedings. We wish all the participants a fruitful conference and a pleasant stay in Las Vegas.

Las Vegas, Nevada, USA — Drs. Henry Selvaraj
Pocatello, Idaho, USA — Dawid Zydek
Las Vegas, Nevada, USA — Grzegorz Chmaj

Committees

Steering Committee

H. Selvaraj (Chair), University of Nevada, Las Vegas, USA
A. Grzech, Wroclaw University of Technology, Poland
J. Swiatek, Wroclaw University of Technology, Poland
D. J. G. James, Coventry University, UK
K. J. Burnham, Coventry University, UK

General Chair

H. Selvaraj, University of Nevada, Las Vegas, USA
Organizing Committee Co-Chairs
G. Chmaj, University of Nevada, Las Vegas, USA
D. Zydek, Idaho State University, USA

Program Committee

A. V. Balakrishnan, University of California, Los Angeles, USA
G. Borowik, Warsaw University of Technology, Poland
W. N. Burkov, Russian Academy of Sciences, Russia
Z. Chaczko, University of Technology, Sydney, Australia
G. Chmaj, University of Nevada, Las Vegas, USA
H. S. Cho, Korean AIST, Korea
L. Gewali, University of Nevada, Las Vegas, USA
T. F. Gonzalez, University of California at Santa Barbara, USA
G. Guardabassi, Politecnico di Milano, Italy
A. Gunasekaran, University of Massachusetts, USA
L. Hsu, Universidade Federal do Rio de Janeiro, Brazil
A. Kasprzak, Wroclaw University of Technology, Poland
M. H. Kolekar, Indian Institute of Technology Patna, India
L. Koszalka, Wroclaw University of Technology, Poland
N. S. Kumar, Velammal College of Engineering and Technology, India
T. Luba, Warsaw University of Technology, Poland
H. Migliore, Portland State University, USA
S. Muthaly, Royal Melbourne Institute of Technology, Australia
C. N. S. Murthy, Chameli Devi Group of Institutions, India
I. Pozniak-Koszalka, Wroclaw University of Technology, Poland
M. Rawski, Warsaw University of Technology, Poland
H. Selvaraj, University of Nevada, Las Vegas, USA
D. Selvathi, Mepco Schlenk Engineering College, India
P. K. Singh, ABV - Indian Institute of Information Technology & Management Gwalior, India
S. Singh, University of Nevada, Las Vegas, USA

H. Sorenson, University of California, San Diego, USA
B. Steinbach, Freiberg University of Mining and Technology, Freiberg, Germany
P. Stubberud, University of Nevada, Las Vegas, USA
S. Stubberud, Boeing Company, USA
M. Sugisaka, Oita University, Japan
M. Thoma, University of Hannover, Germany
A. Vadivel, National Institute of Technology, Trichy, India
R. Vallee, Universite Paris-Nord, France
R. Venkat, University of Nevada, Las Vegas, USA
L. Wang, Harbin Institute of Technology, China
D. Zydek, Idaho State University, USA

Student Volunteers

L. Abraham, University of Nevada, Las Vegas, USA
A. Alsaraj, Idaho State University, USA
D. Henry, University of Nevada, Las Vegas, USA
A. Khamis, Idaho State University, USA
D. Krol, Idaho State University, USA
B. C. Mummadisetty, University of Nevada, Las Vegas, USA
A. Puri, University of Nevada, Las Vegas, USA
A. Sudhakar, University of Nevada, Las Vegas, USA

Contents

AeroSpace Systems

Titan Science Return Quantification 3
Charles R. Weisbin and William Lincoln

Robust Output Feedback Attitude Control of Spacecraft Using Solar Radiation Pressure 9
Lakshmi Srinivasan, Keum W. Lee, and Sahjendra N. Singh

Online Near-Optimal Path Planning to Back-up Aircraft Mission Capabilities in Emergency Conditions 17
S.M.B. Malaek and Z. Shadram

General Control Systems

Nonlinear Optimal Tracking With Incomplete State Information Using State Dependent Riccati Equation 27
Ahmed Khamis, D. Subbaram Naidu, and Dawid Zydek

Nonlinear Position Control of DC Motor Using Finite-Horizon State Dependent Riccati Equation 35
Ahmed Khamis, D. Subbaram Naidu, and Dawid Zydek

Generalization of the Observer Principle for YOULA-Parametrized Regulators 41
László Keviczky and Csilla Bányász

The Compensation of N-th Order Bilinearity Applied with Model Based Controller 49
Lukasz Gadek, Leszek Koszalka, and Keith J. Burnham

Estimation for Target Tracking Using a Control Theoretic Approach – Part 2 55
Stephen C. Stubberud, Arthur M. Teranishi, and Kathleen A. Kramer

Identification of Fractional Order Models: Application to 1D Solid Diffusion System Model of Lithium Ion Cell 63
Walid Allafi, Ivan Zajic, and Keith J. Burnham

Robust Adaptive Control of the Dynamic Multilinked Object: Control of Robot Manipulator 69
Yuliya Lezhnina, Galina Ternovaya, and Viktoriya Zaripova

Modeling and Identification of a Fractional-Order Discrete-Time Laguerre-Hammerstein System 77
Rafal Stanisławski, Marcin Gałek, Krzysztof J. Latawiec, and Marian Łukaniszyn

A comparative Study of Model-Based and Data-Based Model Order Reduction Techniques for Nonlinear Systems 83
T. Aizad, O. Maganga, M. Sumislawska, and K.J. Burnham

Optimised Job-Shop Scheduling via Genetic Algorithm for a Manufacturing Production System 89
Zhonghua Shen, Keith J. Burnham, and Leonid Smalov

Power Systems

3D CFD Simulation of the Thermal Performance of an Air Channel Solar Heater 95
Samir Moujaes and Jayant Patil

A Concept Study for a Compact High-Speed Rotation Heat Pump 101
Haakon Karlsen and Tao Dong

Experimental Investigation of Developing Spray Boiling on a Flat Flake Surface with Constant Heat Flux 107
Zhongyuan Shi and Tao Dong

Life Cycle Assessment of Circulating Fluidized Bed Combustion with CO_2 Post-Combustion Capture 113
Cristian Dinca, Adrian Badea, Vladimir Tanasiev, and Horia Necula

Suggested Simulation of the First Copper-Chlorine Reactor Step for Solar Hydrogen Generation Process 121
Samir Moujaes and Mohamed Yassin

Voltage Regulation in Resonant Coupled Systems for Near Field Power Transfer 127
Hema Ramachandran and G.R Bindu

Security Breach Possibility with RSS-Based Localization of Smart Meters Incorporating Maximum Likelihood Estimator 133
Mahdi Jamei, Arif I. Sarwat, S.S. Iyengar, and Faisal Kaleem

Active/Reactive Power Control of Three Phase Grid Connected Current Source Boost Inverter Using Particle Swarm Optimization 141
Arman Sargolzaei, Mahdi Jamei, Kang Yen, Arif I. Sarwat, and Mohamed Abdelghani

Anti-Islanding Test Results for Multiple PV Inverter Operations 147
Byunggyu Yu and Youngseok Jung

Intelligent Systems

Presentation of A Fuzzy Control Training and Test System 155
K.-D. Kramer, S. Braune, A. Söchting, T. Stolze, and C. Blankenberg

Web Service Intrusion Detection Using a Probabilistic Framework 161
Hassen Sallay, Sami Bourouis, and Nizar Bouguila

Multi-Agent Reinforcement Learning Control for Ramp Metering 167
Ahmed Fares and Walid Gomaa

Intelligent system concept for high-energy performance and adaptable user comfort 175
Vladimir Tanasiev, Adrian Badea, Cristian Dinca, and Horia Necula

Sparse hidden units activation in Restricted Boltzmann Machine 181
Jakub M. Tomczak and Adam Gonczarek

Accelerated learning for Restricted Boltzmann Machine with Momentum Term 187
Szymon Zaręba, Adam Gonczarek, Jakub M. Tomczak, and Jerzy Świątek

Optimizing Interface Area of Percolated Domains in Two Dimensional Binary Compound: Artificial Neural Network Modeling on Monte Carlo Experiments 193
Yongyut Laosiritaworn and Wimalin Laosiritaworn

Cognitive Science Based Scheduling In Grid Environment 199
N.D. Iswarya, M.A. Maluk Mohamed, and N. Vijaya

Vulnerability evaluation of multiplexing PUF for SVM Attacks 205
S. Kiryu, K. Asahi, and M. Yoshikawa

Autonomous Visualization for Mitigating Lack of Peripheral Vision in Remote Safe Teleoperation 211
J. K. Mukherjee

Improving Multi-Panel Lamination Process Optimization using Response Surface Methodology and Neural Network 221
Wimalin Laosiritaworn

Selecting right questions with Restricted Boltzmann Machines 227
Maciej Zięba, Jakub M. Tomczak, and Krzysztof Brzostowski

A formal approach for identifying assurance deficits in unmanned aerial vehicle software 233
Adrian Groza, Ioan Alfred Letia, Anca Goron, and Sergiu Zaporojan

A Load Optimization Considering Reverse Synergy that May Occur with Mixed Load 241
Yongmin Kim, Munhwan Kim, and Hongchul Lee

Predictability of Firm Financial Sustainability Using Artificial Neural Networks: The Case of Qatar Exchange 245
Farzaneh Amani and Adam Fadlalla

The Periodic Signal Filtration Using the Robust Digital Filter Order Calculation Optimized by Approximation 251
Alexey Sergeev-Horchynskyi and Valeriy Rogoza

A Reasoning System for Predicting Study Level based on User's Watching Behaviors 257
Jeonghyeok Kim, Jaemin Hwang, Sanggil Kang, and Nojeong Heo

Temporal Constraints and Sub-Dimensional Clustering for Fast Similarity Search over Time Series Data. Application to Information Retrieval Tasks. 263
Sidahmed Benabderrahmane

Active Learning based on Random Forest and Its Application to Terrain Classification 273
Yingjie Gu, Dawid Zydek, and Zhong Jin

Classification of Multichannel EEG Signal by Linear Discriminant Analysis 279
Mohammad Rubaiyat Hasan, Muhammad Ibn Ibrahimy, S.M.A. Motakabber, and Shahjahan Shahid

Industrial Automation and Robotics

Virtual Enterprise Process Monitoring: An Approach towards Predictive Industrial Maintenance 285
Filipe Ferreira, Ahm Shamsuzzoha, Americo Azevedo, and Petri Helo

Module-based release management for technical changes 293
Günther Schuh, Sasa Aleksic, and Stefan Rudolf

Trajectory Optimization by Particle Swarm Optimization in Motion Planning 299
Jeong-Jung Kim and Ju-Jang Lee

Cost model for an integrated load carrier design process in the lithium-ion battery production 307
Achim Kampker, Christoph Deutskens, Heiner Hans Heimes, Mathias Ordung, and Andreas Haunreiter

Sensorless Force Estimation for a Two-Link Manipulator Based Upon Linear Dynamics 315
Douglas R. Isenberg

A Joint-Space Parametric Formulation for the Vibrations of Symmetric Gough-Stewart Platforms 323
Behrouz Afzali-Far and Per Lidström

Information and Communication Systems

Software Project Planning Using Agile 333
Jianchao Han and Yan Ma

A Modeling Approach to Support Safety Assurance in the Automotive Domain 339
Yaping Luo, Mark den Br, Luc Engelen, and Martijn Klabbers

Dynamic OD transit matrix estimation: formulation and model-building environment 347
Lídia Montero, Esteve Codina, and Jaume Barceló

Microstrip Spiral Resonator for the UWB Chipless RFID Tag 355
A.K.M.Z. Hossain, S.M.A. Motakabber, and M.I. Ibrahimy

An Evaluation of Intrusion Detection System on Jubatus 359
Tadashi Ogino

System of Conceptual Design Based on Energy-Informational Model 365
Viktorya Zaripova and Irina Petrova

An Algorithm for Multi-Source Geographic Data System 373
Chiang-Sheng Lee, Hsine-Jen Tsai, and Yin-Yih Chang

Methodology and Platform for Business Process Optimization 377
Adam Grzech, Krzysztof Juszczyszyn, and Paweł Świątek

Review and Refined Architectures for Monitoring, Information Exchange, and Control of Interconnected Distributed Resources 383
Y V Pavan Kumar and Bhimasingu Ravikumar

Lossless Compression of Climate Data 391
Bharath Chandra Mummadisetty, Astha Puri, Ershad Sharifahmadian, and Shahram Latifi

Distributed Computer and Computer Networks Systems

Parameter Trade-off And Performance Analysis of Multi-core Architecture 403
Surendra Kumar Shukla, CNS Murthy, and P.K. Chande

Approximation algorithms for utility-maximizing network design problem 411
Maciej Drwal

Network Energy Reduction via an Adaptive Shutdown Algorithm 417
Mohamed K. Watfa, L'emir Bachir Chehab, and Zayed Sulaiman Balbahaith

Improving TCP Performance in Mix Networks 423
Mohamed K. Watfa, Mohamed Diab, and Nikhil Stephen

An Epidemic Routing with Low Message Exchange Overhead for Delay Tolerant Networks 429
Teerapong Choksatid and Sumet Prabhavat

EEIS: an Energy Efficient at Idle Slots MAC layer Protocol for Wireless Sensor Networks 437
Usha Jhadane, Pramod Kumar Singh, and Abhishek Patel

Identification of Redundant Node-Clusters for Improved Face Routing 443
Laxmi Gewali and Umang Amatya

Distributed Processing Applications for UAV/drones: A Survey 449
Grzegorz Chmaj and Henry Selvaraj

UAV Cooperative Data Processing Using Distributed Computing Platform 455
Grzegorz Chmaj and Henry Selvaraj

Analog and Digital Hardware Systems

Implementation of an Efficient Library for Asynchronous Circuit Design with Synopsys 465
Tri Caohuu and John Edwards

A Dynamic System Matching Technique-An Analytical Study 473
Peter Stubberud, Stephen Stubberud, and Allen Stubberud

On the effect of High Power Amplifier Non-linearity on the Ergodic Capacity of Multihop MIMO-OFDM Amplify-and-Forward Relay Network 479
Ishtiaq Ahmad, khaled Ali Abuhasel, and Ateeq Ahmad Khan

Stability Analysis of Continuous Time Sigma Delta Modulators 487
Kyung Kang and Peter Stubberud

An Area Efficient Weighting Coefficient Generation Architecture for Polynomial Convolution Interpolation 495
D. Selvathi and C. John Moses

Biometrics Systems

Counting of water-in-oil droplets for targeted drug delivery systems using capacitive sensing technique 503
Cátia Barbosa and Tao Dong

Privacy Preserving Biometric Voice Authentication System – SIPPA-based Approach 509
Bon K. Sy

Monitoring Urban and Land Use Changes in Al-Kharj Saudi Arabia using Remote Sensing Techniques 515
Osama S. Algahtani, Algahtani S. Salama, Abdullah M. Iliyasu, Belal A. Selim, and K. Kheder

System Engineering Standards, Paradigms, Metrics, Testing, etc

Expert Systems Based Response Surface Models for Multidisciplinary Design Optimization 527
Ramesh Gabbur and K Ramchand

A Survey of Approaches used in parallel architectures and Multi-core Processors, For Performance Improvement 537
Surendra Kumar Shukla, CNS Murthy, and P.K. Chande

Aligning systems engineering and project management standards to improve the management of processes 547
Rui XUE, Claude BARON, Philippe ESTEBAN, and Abd-El-Kader SAHRAOUI

Effect of the groove dimensions and orientation on the static and dynamic performance of non recessed hybrid journal bearing 555
Vijay Kumar Dwivedi, Satish chand, and K. N. Pandey

Understanding Asynchronous Distributed Collaboration in an Enterprise Systems Engineering Context 563
Gary L. Klein, Jill L. Drury, and Sherri L. Condon

A Design Model for Rapid Transit Networks Considering Rolling Stock's Reliability and Redistribution of Services During Disruptions 571
Esteve Codina, Ángel Marín, and Lídia Montero

Management System Architecture for 3D Audio Evaluation Database 579
Jaemin Hwang, Jeonghyuk Kim, and Sanggil Kang

A Generic Metamodel for Context-Aware Applications 587
Imen Jaouadi, Raoudha Ben Djemaa, and Hanene Ben Abdallah

Cost Effectiveness of Coverage-Guided Test-Suite Reduction for Safety-Relevant Systems 595
Susanne Kandl

Towards a Holistic Definition of System Engineering: Paradigm and Modeling Requirements 603
Hycham Aboutaleb and Bruno Monsuez

Migration from Legacy Systems to SOA Applications: A Survey and an Evaluation 609
Sukanya Suwisuthikasem and M.H. Samadzadeh

An Approach to Schedule Production using the Reservation Tables 615
Sergiu Zaporojan, Vasile Moraru, and Adrian Groza

Applying System of Systems Engineering Approach to Build Complex Cyber Physical Systems 621
Lichen Zhang

Model Integration and Model Transformation Approach for Multi-Paradigm Cyber Physical System Development 629
Lichen Lichen

Computer Assisted Medical Diagnostic Systems

2D Multi-Slice and 3D k-Space Simulations using a 3D Quadric Head Phantom with MRI Properties 639
H. Michael Gach

Classification of Lungs Nodule using Hybrid Features from CT Scan Images 645
M. Arfan Jaffar and Eisa Al Eisa

A Smart Carpet Design for Monitoring People with Dementia 653
Osamu Tanaka, Toshin Ryu, Akira Hayashida, Vasily G. Moshnyaga, and Koji Hashimoto

Transportonics Engineering

Rationalisation of the Maintenance Process of Transport Telematics System Comprising two Types of Periodic Inspections 663
Adam Rosinski

An Adaptive Controller of Traffic Lights using Genetic Algorithms 669
Kalum Udagepola, Belal Ali Alshami, Naveed Afzal, and Xiang Li

Parameters Analysis of Satellite Support System in Air Navigation 673
Miroslaw Siergiejczyk, Karolina Krzykowska, and Adam Rosinski

Selected Issues of the Reliability Analysis of GSM-R in Poland 679
Mirosław Siergiejczyk

Speed-Volume Relationship Model for Speed Estimation on Urban Roads in Intelligent Transportation Systems 685
Zilu Liang and Yasushi Wakahara

Superpixel based semantic segmentation forassistance in varying terrain driving conditions 691
Ionut Gheorghe, Weidong Li, Thomas Popham, and Keith J. Burnham

Special Session: Computational Cognitive Science

Emotion Estimation using Geometric Features from Human Lower Mouth Portion 701
P. Shanthi and A. Vadivel

Cognitive Based Sentence Level Emotion Estimation through Emotional Expressions 707
S.G Shaila and A. Vadivel

Hybrid Multilingual Key Terms Extraction System for Hindi and Punjabi Text 715
Vishal Gupta

Sentiment and Emotion Prediction through Cognition: A Review 719
T. Vetriselvi and A. Vadivel

A Short Review for Mobile Applications of Sentiment Analysis on Various Domains 723
M. Sivakumar and U. Srinivasulu Reddy

Human Cognition and Vision Based Earlier Path Determination System for Indoor Mobile Robot Path Planning 727
N. Nithya and D. Tamil Selvi

Special Session: Nature-inspired Computational Methods and Applications

Teaching Learning Based Optimization (TLBO) Based Improved Iris Recognition System 735
Shikha Agrawal, Shraddha Sharma, and Sanjay Silakari

Acceleration based Particle Swarm Optimization (APSO) for RNA Secondary Structure Prediction 741
Jitendra Agrawal and Shikha Agrawal

Performance Analysis of Zone Based Features for Online Handwritten Gurmukhi Script Recognition using Support Vector Machine 747
Karun Verma and R. K. Sharma

Words Are Analogous To Lymphocytes: A Multi-Word-Agent Autonomous Learning Model 755
Jinfeng Yang, Xishuang Dong, and Yi Guan

Agile Rough Set Based Rule Induction to Sustainable Service and Energy Provision 761
Chun-Che Huang, Tzu-Liang (Bill) Tseng, Yu-Sheng Liu, Jun-Wei Chu, and Po-An Chen

Intelligent Web Application Systems Testing through Value Based Test Case Prioritization 765
Abdul Rauf and Adel Ibrahim AlSalem

Iterative Hybrid Identification of Spatial Bilinear Models in the Presence of Uncertainty 769
James E. Trollope and Keith J. Burnham

Special Session: Intelligent Video Surveillance Systems

A Fast Non-searching Algorithm for the High-Speed Target Detection 777
Jibin Zheng, Tao Su, Wentao Zhu, and Qing Huo Liu

A Comparative Study of Video Splitting Techniques 783
Abdul Khader Jilani Saudagar and Habeeb Vulla Mohammed

Trajectory Based Unusual Human Movement Identification for Video Surveillance System 789
Himanshu Rai, Maheshkumar H. Kolekar, Neelabh Keshav, and J.K. Mukherjee

Special Session: From Boolean Problems to the Internet of Everything

Design and Implementation of Novel Algorithms for Frequent Pattern Trees 797
R. Siva Rama Prasad, N.S. Kalyan Chakravarthy, and D. Bujji Babu

Using Symbolic Functional Decomposition to Implement FSMs in Heterogenous FPGAs 805
Piotr Szotkowski, Mariusz Rawski, and Paweł Tomaszewicz

Efficient Functional Decomposition Algorithm Based on Indexed Partition Calculus 809
Mariusz Rawski, Paweł Tomaszewicz, and Piotr Szotkowski

Rule Induction Based on Logic Synthesis Methods 813
Grzegorz Borowik, Andrzej Kraśniewski, and Tadeusz Łuba

Simpler Functions for Decompositions 817
Bernd Steinbach

Node Demand Reverse Deduction (DRD) Technology for Water Supply Networks 825
Ronghe Wang, Zhixun Wang, Junhui Ping, Jilong Sun, and Chaohong Xiao

Generalized Spring Tensor Model: A New Improved Load Balancing Method in Cloud Computing 831
Shahrzad Aslanzadeh and Zenon Chaczko

Middleware Solution for Cross-Site Data Transfer 837
Zenon Chaczko, Shahrzad Aslanzadeh, and Mehdi Soltani

Autonomous Model of Software Architecture for Smart Grids 843
Zenon Chaczko, Shahrzad Aslanzadeh, and Alqarni Lulwah

Specification and Design Method for Big Data Driven Cyber Physical Systems 849
Lichen Zhang

Simulating Active Interference Cancellation in Cognitive Radio 859
Zenon Chaczko, Grzegorz Borowik, and Philip Hsieh

A Development Study on Performance of a Real-Time Interface Device 865
Anıl Güçlü, Yağmur Atay, and Yasin Genç

Task Allocation within Mesh Networks: Influence of Architecture and Algorithms 869
Aleksandra Postawka and Iwona Poźniak Koszałka

An Overview of Chip Multi-Processors Simulators Technology 877
Malik Al-Manasia and Zenon Chaczko

A Survey on Design and Implementation of Floating Point Adder in FPGA 885
Luka Daoud, Dawid Zydek, and Henry Selvaraj

Hybrid GPU/CPU Approach to Multiphysics Simulation 893
Dawid Krol, Jason Harris, and Dawid Zydek

Erratum 1 . . . . . E1

Erratum 2 . . . . . E3

Titan Science Return Quantification

Charles R. Weisbin and William Lincoln

1 Introduction

Part of the procedure for submitting a proposal to NASA for a mission or an instrument to be included in a mission is to complete a Science Traceability Matrix (STM), which is intended to show that what is being proposed would contribute to satisfying one or more of the agency's top-level science goals. But the information included in this document is not traditionally in the form of quantities that can be used directly to compute anticipated results, and so evaluations and comparisons are often more subjective than might ideally be desired.

We added quantitative elements to NASA's Science Traceability Matrix and developed a software tool to process the data. We then applied this methodology to evaluate a group of competing concepts for a proposed mission to Saturn's moon, Titan.

Although we recognize that our methodology does not completely eliminate subjectivity since it relies on estimates based on the experience and knowledge of relevant scientists and engineers, it has the merit of clarifying and focusing attention on specific numerical values of significant importance and stimulating discussion about them.

2 Augmenting the Science Traceability Matrix

The STM that NASA has been using asks proposers to provide the following information, which forms a chain from top-level science goals down to the instruments that would be required to furnish the information needed to meet those goals:

1. Relevant NASA science goals (big, general questions such as those set forth in the planetary science decadal survey published by the National Research Council in 2011 [1])
2. Science subgoals (sets of specific questions, the answers to which would lead to answering each of the science-goal questions)
3. Science objectives (sets of more-detailed questions designed to answer each of the subgoal questions—e.g., whether a particular isotope exists on Titan)
4. Scientific measurement requirements (parameters of the measurements needed to answer the science-objective questions)
   a. Observables
   b. Physical parameters
5. Instrument performance requirements (specific measurements needed to meet the measurement requirements)
   a. Measurements
   b. Range and sensitivity of each measurement
6. Projected instrument performance (instruments proposed to make the required measurements)
7. Mission requirements (general description of the types of instruments and functions that would be required)

Our augmented STM adds two significant features to the standard STM:

- It enables a quantitative estimation of the impact of each key link in the STM chain on the next-highest link. Integrating over the entire chain serves to quantify, for each proposed measurement of each proposed mission, its contribution to a set of top-level science goals.
- It calls for listing the events that could degrade the value of the measurements, such as failure of one or more instruments or missing the targeted landing zone, and estimating the probability of each event occurring. Each combination of possible events (e.g., no instruments fail, only instrument #1 fails, only instrument #2 fails, both instruments #1 and #2 fail, etc.) constitutes an "event scenario" as discussed below.

C.R. Weisbin (✉) • W. Lincoln
NASA Jet Propulsion Laboratory, Pasadena, California, USA
e-mail: charles.r.weisbin@jpl.nasa.gov; william.lincoln@jpl.nasa.gov

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_1,

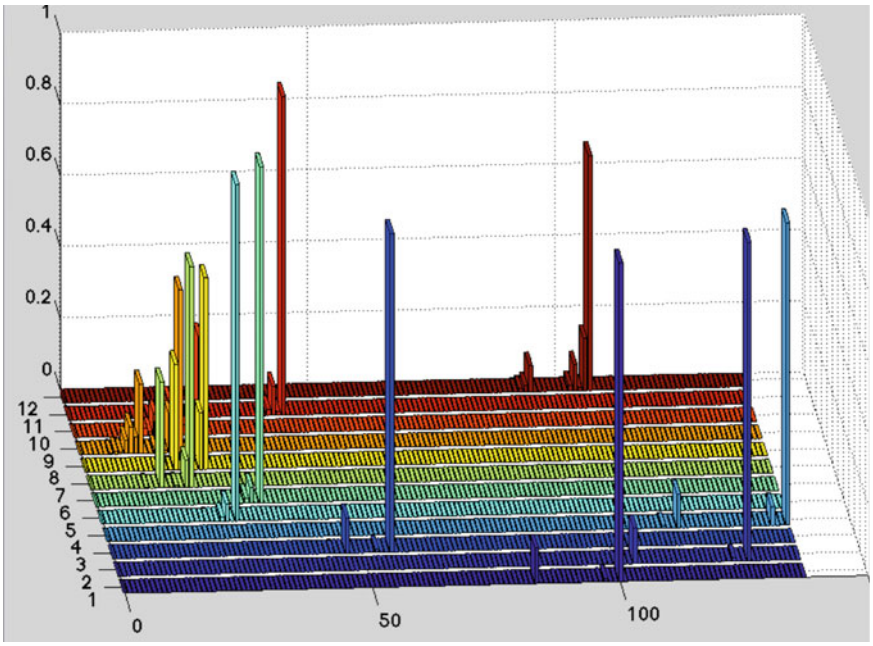


Fig. 1 Relative science value of 12 mission concepts (numbered 1 through 12 on the Z axis) given a particular set of facts and assumptions about each concept. Each bar represents the percentage of minimum acceptable science return (location along the X axis) that a mission concept achieves given a certain combination of conditions, such as unexpected environment or failure of one or more instruments, and the probability of that combination of conditions occurring (the height of the bar on the Y axis). Mission concept #1 is arbitrarily chosen as the baseline, which produces 100% of the minimum acceptable science return if nothing goes wrong

We believe that this augmented process enhances the ability of decision-makers to compare competing concepts objectively and to see clearly the inputs that lead to any given results. For the proposers of mission concepts, it also has the advantage of providing a forum for scientists and engineers to reach consensus on important numerical values, and encouraging the provision of specific information that enables decision-makers to evaluate their proposals in the clearest light.

Our additions to the STM call for the following quantitative information:

A. Relative value of each subgoal towards a top-level goal (i.e., estimate of the percentage of a particular top-level goal that will be satisfied if the subgoal is fully met).
B. Estimate of the percent completion of each subgoal, given full satisfaction of all contributing objectives.
C. Estimate of the percent completion of each objective, given full satisfaction of all contributing observables.
D. Capability-reducing events
   1) Event name
   2) Estimate of event probability
   3) Expected observational loss due to event occurring, from 0 for no loss to 1 for complete loss

Integrating over the quantities provided for A, B and C enables projection of the science return a candidate mission concept would produce, assuming it works perfectly as intended (which we report as a percentage of a baseline mission concept that produces the minimum acceptable science return). This performance is then reduced within a series of event scenarios (i.e., specific combinations of capability-reducing events) by factoring in the probability of the event occurring (D2) and its associated loss to science return (D3) to provide a probabilistic risk-adjusted performance for that specific scenario.

This provides a data point, which can be seen as any one of the bars in Fig. 1. Integrating over all the data points for a given mission concept gives the probability that the mission concept will produce a greater science return than the baseline mission concept.

We created a software tool called SCORE (Science Concept Optimization and Results Evaluator) to process the data collected in the augmented STM. It is capable of calculating the science return to be expected under each of the thousands of possible combinations of mission scenarios.

3 Example: A Study of Potential Missions to Titan

Titan is Saturn's largest moon and one of the most desired targets for further study among planetary scientists, who are eager to follow up on the discoveries made by the Huygens probe in 2005 and the Cassini spacecraft that is currently exploring Saturn and its moons.

The science team participating in this study identified 12 concepts for missions to Titan. Eight are landers, two would float balloons in Titan's atmosphere, and two would conduct their measurements while orbiting Titan. Within each of these three categories, the mission concepts differ from one another in terms of the instruments they would carry and the measurements they are designed to make.

Our study was directed toward analyzing how each proposed experiment of each proposed mission would help to answer one or more of the following questions (top-level science goals) set forth in the 2011 planetary science decadal survey [1]:

1. How did the giant planets and their satellite systems accrete, and is there evidence that they migrated to new orbital positions?
2. What were the primordial sources of organic matter, and where does organic synthesis continue today?
3. Beyond Earth, are there modern habitats elsewhere in the solar system with the necessary conditions, organic matter, water, energy and nutrients to sustain life, and do organisms live there now?
4. Can understanding the roles of physics, chemistry, geology and dynamics in driving planetary atmospheres lead to a better understanding of climate change on Earth?
5. How have the myriad chemical and physical processes that shaped the solar system operated, interacted and evolved over time?

The team of scientists and engineers participating in this study estimated all of the values required by our augmented STM as described above. This process was carried out for every measurement envisioned for every instrument proposed for each of the 12 mission-concept candidates. SCORE was used to process the thousands of combinations and to calculate the science return of each scenario (i.e., each combination of each mission concept's instruments and capability-reducing events) relative to the mission concept chosen to serve as the baseline.

A small excerpt of the augmented STM Excel data sheet follows. It has been modified from the spreadsheet format to increase legibility in this paper. Items in red show what we added to the standard STM in our augmentation process.

NASA science goals priority question 1: How did the giant planets and their satellite systems accrete, and is there evidence that they migrated to new orbital positions?

Science subgoal 1: Characterize the exchange process between the interior and the atmosphere; relate them to the original composition of Titan "satellitesimals," comets, other satellites.

Relative value of subgoal toward goal: 100%

Measurements: Mass spectroscopy

Science Objective question 1: Is there 22Ne on Titan, unlike other places?

Percent completion of subgoal given full satisfaction of all contributing objectives: 10%

Observables: Confirm tentative detection of 22Ne by Huygens GCMS. Physical parameters: 22Ne to ± 10% accuracy down to 10-20 ppb.

Percent completion of Objective given full satisfaction of all contributing observables: 50%

Science Objective question 2: Is there selective trapping of Kr and Xe relative to 36Ar by Titan processes (aerosols, clathrates, dissolution)?

Percent completion of subgoal given full satisfaction of all contributing objectives: 10%

Observables: 36Ar/Kr, 36Ar/Xe, Kr/Xe rations in atmosphere and surface liquids/solids. Physical parameters: Ar, Kr, Xe to ± 25% accuracy for mole fractions down to 0.1 ppb (in atmosphere) and 10 ppb (liquid/ground moisture/solids).

Percent completion of Objective given full satisfaction of all contributing observables: 95%

Science Objective question 3: Do noble gas abundances correspond to direct deposition, trapping in amorphous ice, clathration in the solar nebula, or a combination of processes?

Percent completion of subgoal given full satisfaction of all contributing objectives: 20%

Observables: 36Ar/Kr, 36Ar/Xe, Kr/Xe rations in atmosphere and surface liquids/solids. Physical parameters: Ar, Kr, Xe to ± 25% accuracy for mole fractions down to 0.1 ppb (in atmosphere) and 10 ppb (liquid/ground moisture/solids).

Percent completion of Objective given full satisfaction of all contributing observables: 30%

Science Objective question 4: What is the influence of atmospheric processes (photochemistry, escape) on carbon, nitrogen, and hydrogen isotopes?

Percent completion of subgoal given full satisfaction of all contributing objectives: 20%

Observables: 13C/12C, 15 N/14 N, and D/H profile in atmospheric volatiles, condensates, and surface liquids or solids. Physical parameters: Carbon, nitrogen, and hydrogen isotopes in nitrogen, hydrocarbons, etc. to ± 5%.

Percent completion of Objective given full satisfaction of all contributing observables: 50%

Science Objective question 5: Is Titan's methane primordial or formed by internal or nebula processes from CO/CO2?

Percent completion of subgoal given full satisfaction of all contributing objectives: 20%

Observables: C isotopes in CH4, CO, CO2. Physical parameters: 13C/12C to ± 5% accuracy down to 1 ppm total CO/CO2.

Percent completion of Objective given full satisfaction of all contributing observables: 30%

Potential Capability-Reducing Events

Event: Poorer detection threshold or calibration issue.
Event probability: 0.1
Expected observation loss: 0.2 (where 0 = no loss and 1 = complete loss)

Measurements: Images

Science Objective question 6: Do materials currently vent from Titan's surface?

Percent completion of subgoal given full satisfaction of all contributing objectives: 20%

Observables: Search for ground fogs correlated with fractures or vents, if any. Physical parameters: Surface feature, morphology, mists.

Percent completion of Objective given full satisfaction of all contributing observables: 20%

Potential Capability-Reducing Events

Event 1: Interference from low-lying hazes
Event probability: 0.1
Expected observation loss: 0.5
Event 2: Lack of wind results in lack of movement
Event probability: 0.1
Expected observation loss: 0.5

Science subgoal 2: Determine origin and evolution of Titan's atmosphere.

Relative value of subgoal toward goal: 10%

Measurements: Mass spectrograms

Science Objective question 1: What are the escape rates of N2 and CH4 from Titan?

Percent completion of subgoal given full satisfaction of all contributing objectives: 100%

Observables: Isotopic ratios in N, C, and H. Physical parameters: 14 N/15 N in N2, 12C/13C in HCN, and D/H in CH4 to 2%.

Percent completion of Objective given full satisfaction of all contributing observables: 75%

Potential Capability-Reducing Events

Event 1: Issue with data acquisition during aerocapture
Event probability: 0.1
Expected observation loss: 0.1

The STM also includes the range, sensitivity and instrument projected to be used for each measurement. Our enhanced version indicates which measurements apply to which of the candidate mission concepts.

4 Results

One of the lander concepts (#1 in Fig. 1 and Table 1) was designated the baseline mission, which by definition would produce 100% of the minimum acceptable science return if nothing goes wrong. Two of the other lander concepts (#2 and #4) were found to exceed the baseline science return under all of the scenarios that were considered, while one of the orbiter concepts (#12) was found to have a 91% probability of doing so. The remaining eight mission concepts were found to fall short of the baseline and would not need to be considered unless further analysis showed that the cost and/or risk of the baseline and three higher-scoring concepts were unacceptable.

Participants in this study reported that the quantification process provided clarity and focus to the science discussion of mission concepts.

We conducted this study with a neutral attitude toward risk. However, the formulas could be adjusted to reward or penalize missions that have higher probabilities of exceeding the baseline under some scenarios and of falling short of the baseline under other scenarios, depending on the preference of the decision-maker.

As noted above, the analysis cited here used the science goals set forth in the 2011 planetary science decadal survey as the top-level science goals. The same methodology would apply to an analysis based on any set of top-level science goals, but of course the results might differ.

Table 1 Results calculated when nothing goes wrong (column 2) and when projected science return for each mission concept is integrated across all of the capability-reducing events that were considered in this study (column 3). The results shown in red are for mission concepts that meet or exceed the baseline science return

Mission Concept	% baseline science return achieved with no capability-reducing events	Probability of exceeding baseline, integrated across capability-reducing events
1 (baseline)	**100**	**0.50**
2	**127**	**1.00**
3	56	0.00
4	**137**	**1.00**
5	27	0.00
6	33	0.00
7	20	0.00
8	24	0.00
9	20	0.00
10	26	0.00
11	43	0.00
12	**106**	**0.91**

5 Next Steps

It would be a straightforward matter to conduct sensitivity analyses of the various inputs to identify which ones are most significant in determining a mission concept's science return. Those inputs could be subjected to greater scrutiny if desired.

We note that cost was not taken into account in this study. However, a comparative cost-benefit analysis of the various mission concepts could easily be done. Ultimately, the mission concepts could be compared in a science-cost-risk space for the most meaningful relative ranking.

Acknowledgments Technical data inputs to the study were provided by Titan team scientists led by Dr. Mathieu Choukroun. This study was sponsored by the JPL Solar System Exploration Program Office and was conducted at NASA's Jet Propulsion Laboratory, which is operated by the California Institute of Technology.

Reference

1. National Research Council. *Vision and Voyages for Planetary Science in the Decade 2013-2022*. Washington, DC: The National Academies Press, 2011.

Robust Output Feedback Attitude Control of Spacecraft Using Solar Radiation Pressure

Lakshmi Srinivasan, Keum W. Lee, and Sahjendra N. Singh

1 Introduction

Spacecraft and interplanetary probes, orbiting beyond the Earth's detectable atmosphere, experience physical pressure caused by impinging solar radiation. Researchers have considered use of solar radiation pressure (SRP) exerted on control surfaces mounted on satellites for the purpose of control. Researchers have proposed a variety of control surface configurations including trailing cone, reflector-collector, weathervane tail, mirror arrays, solar paddles, and solar sails for deriving solar control forces [PV1]. In the Mariner IV mission [S1] and the OTS-2 mission of the European Space Agency [R1] solar vanes and flaps were employed for the control of geostationary communication satellites.

In the past, a variety of control systems for the attitude control of satellites using solar radiation pressure have been developed. A time optimal control law for pitch angle control has been designed [R1]. The control of an Earth-pointing satellite has been also considered [PV2]. Optimal control laws for inertially-fixed attitude control have been designed [PV3, VP]. A control law for large angle maneuver has been proposed [Ve]. Joshi and Kumar designed attitude control systems for satellites orbiting in elliptic orbits [JK1]. A nonlinear feedback linearizing attitude control law has been developed [SY2]. Authors have also considered design of attitude control systems for satellite models in the presence of uncertainties. The variable structure control systems [PKB1, PKB2] and adaptive sliding mode control systems [VK1, V1] have been proposed. The solar pressure adaptive controllers for attitude control have also been developed [SY2, LS1, LS2, LS3]. Recently, an $\mathcal{L}_1$ adaptive pitch angle controller using SRP has been designed [LS3]. Also a solar attitude controller [SL1] for a finite-time regulation based on a higher-order sliding mode control technique, has been developed. But for the synthesis of control law in [SL1], measurements of the first and second derivatives of the pitch angle are required. The adaptive laws developed in [PKB1, PKB2, VK1, LS2] also assumed availability of the complete state vector. Certainly, it is desirable to use fewer sensors for measurement. As such it is of interest to develop adaptive SRP attitude control systems for satellite models with unmodelled dynamics which require only the pitch angle measurement for feedback.

The contribution of this paper lies in the derivation of a robust output feedback control system for large pitch angle control of satellites in elliptic orbits using the SRP, despite uncertainties. The nonaffine-in-control model of the satellite includes unmodelled nonlinear functions, unknown inertial and solar parameters and time-varying disturbance input. The satellite is equipped with two rotating reflective control surfaces (solar flaps) for the purpose of control. It is assumed that only the pitch angle is measured for synthesis. The control torque derived from the SRP is an implicit function of the deflection angles of the solar flaps. A robust nonlinear feedback linearizing control law is designed for large angle rotational maneuvers of the satellite in the pitch plane. The control system includes a high-gain observer to obtain the estimates of the derivatives of the pitch angle and the lumped unmodelled nonlinearities in pitch dynamics for synthesis. Simulation results are presented which show that in the closed-loop system precise pitch angle trajectory control of the spacecraft moving in an elliptic orbit is accomplished, in spite of large parameter uncertainties, unmodelled nonlinearites and external disturbance moment in the model.

L. Srinivasan (✉) • S.N. Singh
Department of Electrical and Computer Engineering,
University of Nevada, Las Vegas, NV 89154-4026, USA
e-mail: Sahjendra.Singh@unlv.edu

K.W. Lee
Division of Electronic Information and Communication, Kwandong University, Gangwon 210-701, South Korea
e-mail: kwlee@kd.ac.kr

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_2,

2 Dynamics of Spacecraft

Fig. 1 shows an unsymmetrical satellite with its center of mass S rotating in an elliptic orbit about the Earth's center O. The chosen inertial (XYZ), rotating orbital ($X_0Y_0Z_0$) and body-fixed ($X_bY_bZ_b$) coordinate systems are also shown in the figure. (The axes Z, Z_0 and Z_b normal to the orbital plane are not shown in the figure.) The solar aspect angle is denoted by ϕ, and ω and θ are the argument of perigee and true anomaly, respectively. The pitch angle α is equal to $\lambda + \theta$, where λ is the angle between the body-fixed axis X_b and the local vertical axis X_0. The solar radiation torque is produced by two identical, highly reflective, lightweight control surfaces P_1 and P_2 mounted on the satellite. The center of pressure of each control surface lies on the X_b axis. The rotation angles of the two flaps measured from the axis X_b are δ_1 and δ_2. Since the radiation forces on these control surfaces are directed along the surface normals, only the rotation of the satellite about the axis normal to the orbital plane is produced by the solar radiation pressure.

The second-order differential equation describing the pitch attitude of the spacecraft is described by [PV2]

$$I_z \frac{d^2\alpha}{dt^2} = M_g + M_s + M_d(t) \tag{1}$$

where M_s is the net solar torque, M_g is the gravitational torque, and $M_d(t)$ denotes the external time-varying disturbance torque. Of course, the chosen model is valid under the assumption that the roll and yaw angles of the satellite are controlled by means of additional solar flaps and actuators so that its axis Z_b remains normal to the orbit. The net solar torque produced by the control surfaces is a nonlinear function of δ_i. It has been shown in [PV1, PV2] that it is given by

$$\begin{aligned} M_s &= C_s'\sigma_s(\phi)[\sin^2(\alpha + \beta_s(\phi) + \delta_1)\Delta_1 \cos\delta_1 \\ &\quad - \sin^2(\alpha + \beta_s(\phi) + \delta_2)\Delta_2 \cos\delta_2] \\ &\doteq C_s'\sigma_s\psi(\alpha, \beta_s, \delta) \end{aligned} \tag{2}$$

where $\Delta_i = sgn(\sin(\alpha + \beta_s + \delta_i)), i = 1, 2$ and $\delta = (\delta_1, \delta_2)^T$

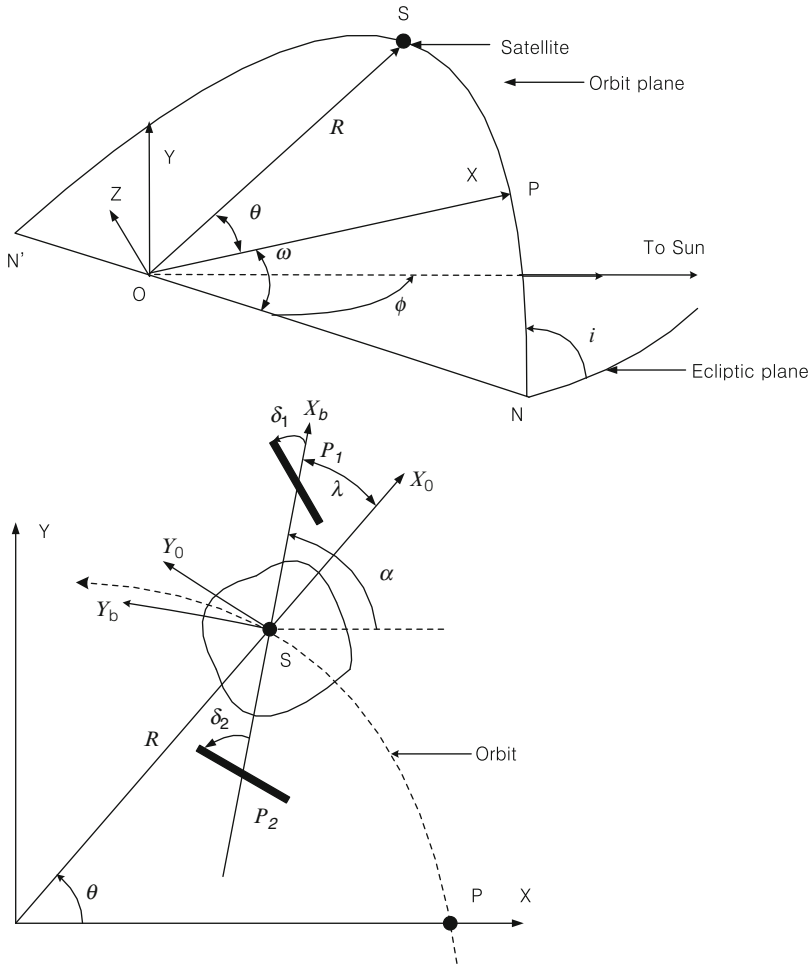


Fig. 1 Orbital and satellite coordinate systems

and the nonlinear function ψ is defined in equation (1). The functions σ_s and β_s are

$$\begin{aligned}\sigma_s(\phi) &= 1 - \sin^2\phi \sin^2 i; \beta_s(\phi) \\ &= \omega - \tan^{-1}(\tan\phi \cos(i))\end{aligned} \tag{3}$$

The solar aspect angle varies from 0 to 2π radians in a year; and therefore, it is a slowly varying function of θ.

The parameter $C_s^{'}$ is $C_s^{'} = 2\rho_s p A_s l$, where A_s is the surface area of the solar flap exposed to impinging photons, p is the nominal SRP constant, ρ_s is the fraction of impinging photons specularly reflected, and l is the distance between the center of pressure on the solar flap and the system center of mass. The gravity gradient torque M_g acting on the spacecraft is given by

$$M_g = -\frac{3\mu}{R^3(\theta)}(I_x - I_y)\sin\lambda\cos\lambda \tag{4}$$

where I_x, I_y and I_z are the moments of inertia of the satellite about the body-fixed axes (X_b, Y_b, Z_b) and $R(\theta)$ is the distance of the satellite center of mass from the Earth's center.

For the satellite moving in an elliptic orbit, $R(\theta)$ and the orbital angular velocity are given by [PKB2]

$$\begin{aligned}R(\theta) &= \frac{a(1-e^2)}{1+e\cos\theta} = \frac{\mu^{1/3}(1-e^2)}{\Omega^{2/3}(1+e\cos(\theta))} \\ \frac{d\theta}{dt} &= \frac{\sqrt{\mu a(1-e^2)}}{R^2}\end{aligned} \tag{5}$$

where e is the eccentricity, a denotes the semi-major axis of the orbit, and the mean orbital rate is $\Omega = (\mu/a^3)^{1/2}$. Now instead of the time t, the true anomaly θ is treated as an independent variable. For simplicity in notation, the derivatives of functions with respect to θ will be denoted by overdots. Using Eqs.(1) and (5), it can be shown that the derivative of the pitch angle with respect to θ satisfies [PKB2, LS2]

$$\begin{aligned}(1+e\cos\theta)\ddot{\alpha} = &-1.5K\sin 2(\alpha-\theta) + 2e\dot{\alpha}\sin\theta \\ &+ C_s\sigma_s M_{sn}(\alpha,\theta,\beta_s(\phi),\delta) + M_{dn}(\theta)\end{aligned} \tag{6}$$

where $K = (I_x - I_y)I_z^{-1}$, $C_s = C_s^{'}(I_z\Omega^2)^{-1}$, and

$$\begin{aligned}M_{sn} &= \left(\frac{1-e^2}{1+e\cos\theta}\right)^3\psi; \\ M_{dn} &= M_d\frac{(1-e^2)^3}{(1+e\cos\theta)^3 I_z\Omega^2}\end{aligned} \tag{7}$$

Solving for $\ddot{\alpha}$, Eq. (6) gives

$$\ddot{\alpha} = f_0(\alpha,\dot{\alpha},\theta) + C_s v(\alpha,\theta,\delta) \tag{8}$$

where the nonlinear functions f_0 and v are

$$\begin{aligned}f_0(\alpha,\dot{\alpha},\theta) &= (1+e\cos\theta)^{-1}[2e\sin\theta\dot{\alpha} - 1.5K\sin 2(\alpha-\theta) + M_{dn}] \\ v(\alpha,\theta,\delta) &= (1+e\cos\theta)^{-1}\sigma_s M_{sn}\end{aligned} \tag{9}$$

Note that the disturbance input M_d is included in the nonlinear function f_0. For the design of the controller, it is assumed that the nonlinear function f_0 as well as the solar parameter C_s are not known.

Suppose that α_r is a given reference pitch angle trajectory. The objective is to design a robust control law such that the pitch angle α asymptotically converges to the reference trajectory α_r, despite the presence of disturbance input. Furthermore, controller is to be synthesized using only the pitch angle α.

3 Robust Feedback Linearizing Control Law

In this section, a feedback linearizing control system is designed. Because the solar torque is an implicit function of the solar plate deflection angles, it will be convenient to treat $\dot{\delta}$ as control input vector. Differentiating Eq. (8), it can be shown that the third derivative of the pitch angle with respect to θ satisfies

$$\begin{aligned}\dddot{\alpha} &= \dot{f}_0(\alpha,\dot{\alpha},\ddot{\alpha},\theta) + C_s\left[\frac{\partial v}{\partial\alpha}\dot{\alpha} + \frac{\partial v}{\partial\theta}\right] + C_s B_s u \\ &\doteq f_a(x,\theta,\delta) + C_s B_s u\end{aligned} \tag{10}$$

where $x = (\alpha,\dot{\alpha},\ddot{\alpha})^T$, $u = \dot{\delta}\in R^2$ and the input matrix is

$$B_s = \left[\frac{\partial v}{\partial\delta_1}, \frac{\partial v}{\partial\delta_2}\right] \tag{11}$$

For the derivation of the control law, it is assumed that the function f_a is represents an unstructured nonlinear function and the solar parameter C_s is not known. We are interested in the region Ω_s of the state space in which the rank of $B_s(\alpha, \theta, \delta)$ is 1. For the derivation of control law, the unknown nonlinear function $f_a(x, \theta, \delta)$ and the unknown parameter C_s are decomposed as

$$f_a = f_a^* + \Delta f_a; C_s = C_s^* + \Delta C_s \tag{12}$$

where functions with '*' are the nominal values and the unknown parts are denoted with $\Delta(.)$ Then one can write Eq.(10) as

$$\frac{d^3\widetilde{\alpha}}{d\theta^3} = -\ddot{\alpha}_r + f_a^* + \Delta f_a + (C_s^* + \Delta C_s)B_s(\alpha,\theta,\delta)u \quad (13)$$

where $\dot{\alpha} = \alpha - \alpha_r$ is the tracking error. Because the vector function B_s is known, consider a new input signal

$$u_a = B_s(x,\theta,\delta)u, u_a \in R \quad (14)$$

Define a lumped nonlinear function $\eta \in R$ as

$$\eta = \Delta f_a + \Delta C_s u_a \quad (15)$$

Then one can write Eq.(13) as

$$\frac{d^3\widetilde{\alpha}}{d\theta^3} = -\ddot{\alpha}_r + f_a^* + \eta + C_s^* u_a \quad (16)$$

In view of Eq. (16), a feedback linearizing control law is chosen as

$$u_a = (C_s^*)^{-1}\left(\ddot{\alpha}_r - f_a^* - \eta - p_3\ddot{\widetilde{\alpha}} - p_2\dot{\widetilde{\alpha}} - p_1\widetilde{\alpha} - p_0 x_s\right); \quad \dot{\alpha} = \widetilde{\alpha} \quad (17)$$

where p_i are the feedback gains. Substituting the control law (17) in (16) gives

$$(s^4 + p_3 s^3 + p_2 s^2 + p_1 s + p_0)\widetilde{\alpha} \doteq \lambda(s)\widetilde{\alpha} = 0 \quad (18)$$

where s denotes the Laplace variable. The feedback gains are chosen such that the roots of $\lambda(s) = 0$ are stable. For such a choice of feedback gains, $\dot{\alpha}$ converges to zero. But the control law (17) is not implementable because the nonlinear function η and the derivatives of α are not known.

Let z_1, z_2, z_3 and z_4 be the estimates of $\dot{\alpha}$, $\ddot{\widetilde{\alpha}}$, $\dddot{\widetilde{\alpha}}$ and η, respectively. Then, in view of Eq. (17), one chooses a modified feedback linearizing control law as

$$u_a = (C_s^*)^{-1}(\ddot{\alpha}_r - f_a^* - z_4 - p_3 z_3 - p_2 z_2 - p_1 z_1 - p_0 x_s) \quad (19)$$

Using Eq. (19) for u_a, now $\dot{\delta} = u$ can be obtained as

$$u = B_s^{*T}(B_s^*(B_s^*)^T)^{-1}u_a \quad (20)$$

Note that if the estimation errors $(\dot{\alpha} - z_1)$, $(\ddot{\widetilde{\alpha}} - z_2)$, $(\dddot{\widetilde{\alpha}} - z_3)$ and $(\eta - z_4)$ are zero, then the control law Eq. (19) becomes the exact feedback linearizing control law Eq. (17).

4 Estimator Design

In this section, the design of an estimator is considered. The structure of the estimator is based on the results of [A1, EK1, KE]. Here the nonlinear function η is treated as a state variable. Differentiating η gives

$$\dot{\eta} = \frac{d}{d\theta}[\Delta f_a + \Delta C_s u_a] f_\eta \quad (21)$$

Note that the derivative of u_a can be obtained by using Eq. (19). The nonlinear function f_η is not known. For the derivation of the estimator, consider a set of equations

$$\frac{d}{dt}\begin{bmatrix}\dot{\alpha}\\ \ddot{\widetilde{\alpha}}\\ \dddot{\widetilde{\alpha}}\\ \eta\end{bmatrix} = \begin{bmatrix}\ddot{\widetilde{\alpha}}\\ \dddot{\widetilde{\alpha}}\\ f_a^* - \ddot{\alpha}_r + \eta + C_s^* u_a\\ f_\eta\end{bmatrix} \quad (22)$$

For obtaining estimates (z_1, z_2, z_3, z_4) of $(\widetilde{\alpha}, \dot{\widetilde{\alpha}}, \ddot{\widetilde{\alpha}}, \eta)$, a high-gain estimator is designed. The advantage of this estimator is that the estimation error converges to zero in a very short period. In view of Eq. (22), the observer is selected as \openup 12pt

$$\begin{aligned}\dot{z}_1 &= z_2 + \epsilon^{-1}d_1(\widetilde{\alpha} - z_1)\\ \dot{z}_2 &= z_3 + \epsilon^{-2}d_2(\widetilde{\alpha} - z_1)\\ \dot{z}_3 &= -\ddot{\alpha}_r + f_a^* + C_s^* u_a + z_4 + \epsilon^{-3}d_3(\widetilde{\alpha} - z_1)\\ \dot{z}_4 &= \epsilon^{-4}d_4(\widetilde{\alpha} - z_1)\end{aligned} \quad (23)$$

where d_i, $(i = 1, 2, 3, 4)$, are real numbers, and $\epsilon > 0$ is a small parameter. The parameters d_i are selected so that the roots of

$$s^4 + d_1 s^3 + d_2 s^2 + d_3 s + d_4 = 0 \quad (24)$$

are stable.

Define the estimation errors as $e_1 = \dot{\alpha} - z_1$, $e_2 = \dot{\alpha} - z_2$, $e_3 = \ddot{\widetilde{\alpha}} - z_3$ and $e_4 = \eta - z_4 = \widetilde{\eta}$. Subtracting Eq. (23) from (22), one obtains the dynamics of the estimation error as

$$\begin{aligned}\dot{e}_1 &= e_2 - \epsilon^{-1}d_1 e_1\\ \dot{e}_2 &= e_3 - \epsilon^{-2}d_2 e_1\\ \dot{e}_3 &= e_4 - \epsilon^{-3}d_3 e_1\\ \dot{e}_4 &= -\epsilon^{-4}d_4 e_1 + f_\eta\end{aligned} \quad (25)$$

Introduce a change of variables as $(i = 1, 2, 3, 4)$

$$\xi_i = e_i \epsilon^{i-4} \quad (26)$$

Using the definition of ξ_i, Eq. (25) can be written as

$$\epsilon\dot{\xi} = A_0\xi + (0,0,0,1)^T\epsilon f_\eta \tag{27}$$

where $\xi = (\xi_1,\xi_2,\xi_3,\xi_4)^T \in R^4$ and the stable matrix A_0 is

$$A_0 = \begin{pmatrix} -d_1 & 1 & 0 & 0 \\ -d_2 & 0 & 1 & 0 \\ -d_3 & 0 & 0 & 1 \\ -d_4 & 0 & 0 & 0 \end{pmatrix} \tag{28}$$

Equation (27) is in a singularly perturbed form. It has been shown in [A1,EK1,KE] that for sufficiently small ε, the error ξ_i converges to zero in a short time. For convergence analysis, one may follow the steps in the derivation of [A1, EK1]; and therefore, it is not repeated here. As the estimation error converges to zero, the control law Eq. (19) becomes a feedback linearizing control law, and in the closed-loop system including the high-gain observer Eq. (23), the performance of the deterministic feedback controller is recovered after a very short transient process. It is pointed out that in contrast to parameter adaptive systems, here the lumped unstructured nonlinear function is adapted using the dynamic estimator.

5 Simulations results

This section presents the results of digital simulation. The complete closed-loop system including the satellite model Eq. (8), the control law Eq. (19) and the high-gain observer Eq. (23) with and without external disturbance moment is simulated for a set of values of K, C_s, eccentricity e, orbit inclination i and solar aspect angle ϕ. The solar aspect angle ϕ is a slowly varying function. The function ϕ given by

$$\phi(\theta) = \phi_0 + (\partial\phi/\partial\theta)(\theta - \theta_0)$$

is used here for computation, where $\phi_0 = \phi(\theta_0)$. The inclination of the orbital plane of the geosynchronous satellite is $i = 23.\ 5^o$. The semi-major axis is $a = 42,\ 241$ km and I_z is 500 kg.m^2. The initial conditions of the spacecraft are chosen as $\theta_o = 0, \alpha(\theta_o) = 100^o$ and $\dot{\alpha}(\theta_o) = 0$. The initial values of the flap deflections are $\delta_1(\theta_o) = 0^o$ and $\delta_2(\theta_o) = 0^o$. The nominal parameter $C_s{}^*$ is set to 6.2. The nonlinear nominal function $f_a{}^*(x,\ \theta,\ \delta)$ is assumed to be zero for simplicity in implementation; that is, $f_a(x,\ \theta,\ \delta) = \Delta f_a$. Apparently, such a choice of Δf_a represents a large uncertainty in the model. The reference pitch angle trajectory is generated by a fifth-order reference generator given by

$$\frac{d^4\alpha_r}{d\theta^5} = -p_{r3}\frac{d^3\alpha_r}{d\theta^3} - p_{r2}\frac{d^2\alpha_r}{d\theta^2} - p_{r1}\frac{d\alpha_r}{d\theta} - p_{r0}(\alpha_r - \alpha^*) \tag{29}$$

where α^* is the target pitch angle. The initial conditions are $\alpha_r(0) = 100^o$ and $d^j\alpha_r(0)/d\theta^j = 0, j = 1,\ 2.\ .\ .\ 4$. The poles of the reference generator are at -1, -1.5, -2.5, and -2. The roots of $\lambda(s)$ in Eq.(18) are set at -2.5, -3, -4, and -3.5. The roots of Eq. (24) for the observer are -1.5, -2.5, -2, and -1; and the ε is selected to be 0.005. These controller parameters have been selected by observing the simulated responses.

Robust attitude control despite sinusoidal, random and pulse disturbance input $M_d:\ K = 0.5, C_s = 5,\ e = 0.2,\ i = 23.5^o,\ \phi_0 = 45^o,\ \alpha^* = 0^o$

Simulation is done to examine the performance of the adaptive controller in the presence (i) sinusoidal, (ii) random and (iii) pulse type disturbance inputs, shown in the left, center, and right column in Fig. 2, respectively. The random disturbance is generated by passing a white noise with unit variance through a transfer function $F(s) = 5 \times 10^{-10}/(s+5)$. The initial value is $\alpha(0) = 0^o$, and it is desired to control the pitch angle to zero. Note that the nominal values $f_a{}^*$ and $C_s{}^*$ are zero and 6.2, respectively. It is observed in Fig. 2 that the controller achieves the regulation of the pitch angle to the target value in the presence of each disturbance input in about one orbit time. In the steady-state, it is observed that flap deflection is a periodic function in the presence of sinusoidal disturbance (Fig. 2, left column). The maximum value of control surface deflection is about (23, 22) (deg). The control signal $C_s v$ is also shown in the figure.

Extensive simulation has been performed for several values of the solar aspect angle, the eccentricity e of the orbit, the orbit inclination i, and the model parameters K and C_s. These results showed that the designed control law accomplishes robust regulation of the pitch angle trajectory, even in the presence of disturbance input.

6 Conclusions

The design of a robust output feedback adaptive control system for the pitch angle control of spacecraft, orbiting in elliptic orbits, using solar radiation pressure was considered. The parameters of the nonaffine-in-control spacecraft model were assumed to be unknown, and external disturbance input was assumed to be acting on the satellite. It was assumed that only the pitch angle is measured for feedback. A robust feedback linearizing control law was designed for the tracking of reference pitch angle trajectory. For the synthesis of the control law, a high-gain estimator was designed for the estimation of the pitch angle derivatives as well as the lumped unmodelled nonlinear function in the pitch dynamics. In the closed-loop system, the controller accomplished precise pitch attitude control, despite uncertainties and disturbance input.

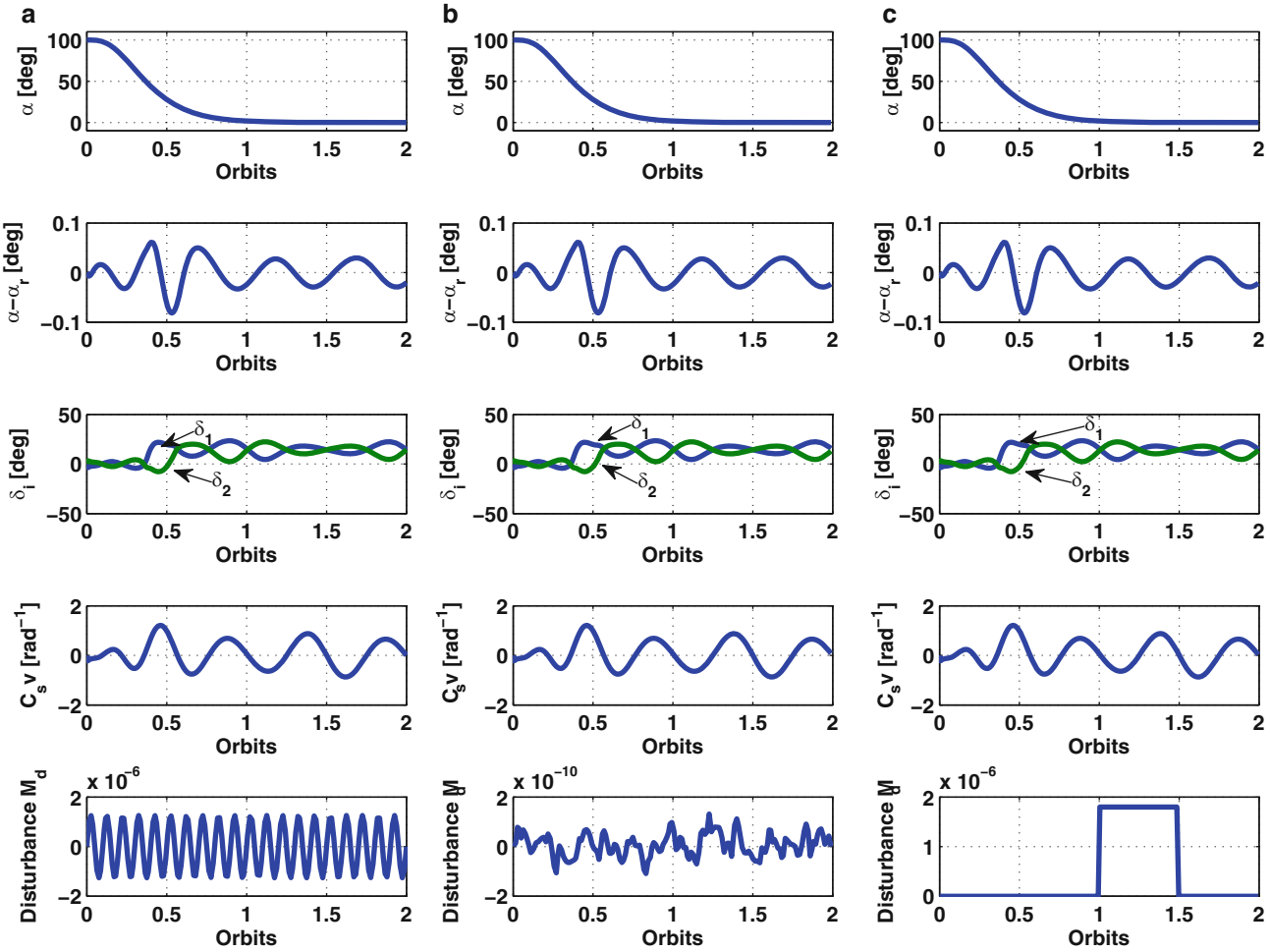


Fig. 2 $C_s = 5, K = 0.5, i = 23.5, e = 0.2, \phi_0 = 45^o, \alpha^* = 0^o$; (**a**) response for sinusoidal disturbance, (**b**) random disturbance (**c**) pulse type disturbance

References

1. Alvarez-Ramirez,J.: Adaptive Control of Feedback Linearizable Systems: A Modeling Error Compensation Approach. Int. J. Robust Nonlin. Contr. 9 (1999) 361–377
2. Esfandiari, F., Khalil, H. K.: Output Feedback Stabilization of Fully Linearizable Systems. Int. J. Control. 56(5) (1992) 1007–1037
3. Joshi, V. K., Kumar, K.: New solar attitude control approach for satellites in elliptic orbits. J. Guidance, Control, and Dynamics. 3(1) (1980) 42–47
3. Khalil, H.K., Esfandiari,F.: Semiglobal Stabilization of a Class of Nonlinear Systems Using Output Feedback. IEEE Tr. Automat. Control. 38(9) (1993) 1412–1415
4. Lee, K.W., Singh, S.N.: Non-certainty-equivalent adaptive satellite attitude control using solar radiation pressure. Proc. IMechE. Part G: J. Aerospace Eng. 223 (2009) 977–988
5. Lee, K.W., Singh, S.N.: Attractive manifold-based adaptive solar attitude control of satellites in elliptic orbits. Acta Astronautica 68 (1-2) (2011) 185–196
6. Lee, K.W., Singh, S.N.: $\mathcal{L}_1$ adaptive attitude control of satellites in elliptic orbits using solar radiation pressure. Proc. IMechE. Part G: J. Aerospace Eng. 228 (2014) 611–626
7. Pande, K.C., Davies, M.S., Modi, V.J.: Time-optimal pitch control of satellites using solar radiation pressure. J. Spacecraft and Rockets. 11(8) (1974) 601–603
8. Patel, T.R., Kumar, K.D., Behdinan, K.: Satellite attitude control using solar radiation pressure based on non-linear sliding mode control. Proc. IMechE. Part G: J. Aerospace Eng. 222 (2008) 379–392
9. Patel, T.R., Kumar, K.D, Behdinan, K.: Variable structure control for satellite attitude stabilization in elliptic orbits using solar radiation pressure. Acta Astronautica. 64(2-3) (2009) 359–373
10. Pande, K.C., Venkatachalam, R.: Semipassive pitch attitude control of satellites by solar radiation pessure. IEEE Tr. Aerosp. Electron. Syst. 15(2) (1979) 194–198
11. Pande, K.C., Venkatachalam, R.: Solar pressure attitude stabilization of earth-pointing spacecraft. IEEE Tr. Aerosp. Electron Syst. 17(6) (1981) 748–756
12. Pande, K.C., Venkatachalam, R.: Optimal solar pressure attitude control of spacecraft - I: inertially-fixed attitude stabilization. Acta Astronautica. 9(9) (1982) 533–540
13. Renner, U.: Attitude control by solar sailing-a promising experiment with OTS-2. European Space Agency Journal. 3 (1979) 35–40
14. Scull, J.R.: Mariner IV revisited, or the tale of the ancient mariner. Proc. 20th International Astronautical Federation Congress, Argentina. (1969) 747–758

15. Srinivasan, L., Lee, K.W., Singh, S.N.: Finite-time control of satellites in elliptic orbits despite uncertainties using solar radiation pressure. AIAA SciTech (2014)
16. Singh, S. N., Yim, W.: Feedback linearization and solar pressure satellite attitude control. IEEE Tr. Aerosp. Electron. Syst. 32(2) (1996) 732–741
17. Singh, S.N., Yim, W.: Nonlinear Adaptive Spacecraft attitude control using solar radiation pressure. IEEE Tr. Aerosp. Electron. Syst. 41(3) (2005) 770–779
18. Varma, S.: Control of satellites using environmental forces: aerodynamic drag/ solar radiation pressure. Ph. D. Thesis, Ryerson University, Toronto, Canada (2011)
19. Venkatachalam, R.: Large angle pitch attitude maneuver of a satellite using solar radiation pressure. IEEE Tr. Aerosp. Electro. Syst. 29(4) (1993) 1164–1169
20. Varma, S., Kumar, K.D.: Fault tolerant satellite attitude control using solar radiation pressure based on nonlinear adaptive sliding mode. Acta Astronautica. 66 (2010) 486–500
21. Venkatachalam, R., Pande, K.C.: Optimal solar pressure attitude control of spacecraft - II:large-angle attitude maneuvers. Acta Astronautica. 9(9) (1982) 541–545

Online Near-Optimal Path Planning to Back-up Aircraft Mission Capabilities in Emergency Conditions

S.M.B. Malaek and Z. Shadram

1 Introduction

Based on system theory, the operational architecture can be seen as both functional and physical architecture. Today's point of view on safety and safe aircraft is originating from physical architecture by looking through health and functional components. In this research we want to introduce a novel point of view relies on functional architecture which is looking for the question that is the airplane capable of performing its mission or not. Thus, based on theorizing the management process of aircraft after occurrence of an undesirable event, first the failure severity should be analyzed, and then based on the result, the possibility of accomplishing the mission should be computed and if required, the alternative mission such as new destination and the appropriate trajectory should be determined.

One of the advantage of implementing such a system is to prevent unnecessary emergency declarations which lead to direct and indirect costs for the whole aviation industry, such as costs related to facilities required to emergency landing, increasing workload of the alternative airport crew and the pressure on the flight crew; these increasing pressure and workload may induce another errors and accident to other aircrafts nearby; and indirectly reducing the stakeholders' trust to the airline and aerial transportation.

Accordingly, the highest purpose of this research is an intelligent aircraft which can recognize its problems and capabilities during the flight and manage the flight by itself. Such a system will be very advantageous in Unmanned Aerial Vehicles (UAVs). Although this class of aircrafts is spreading wider each day, their reliability is low. Thus, developing a systematic solution for different damage scenarios such as malfunction in control systems or propulsion system becomes more important.

Today's advancement in airplane design technologies, make airplane a highly complex system in such a way that is impossible to allocate the whole control and management tasks on the human pilot. Hence, the level of automation [1] is increasing by the advancement of design complexities. Yet, the human power of decision making in emergency conditions is a special characteristic make him Irreplaceable. Hence, in this research we are trying to introduce an expert system which can aid the pilot in management of the complex conditions during flight.

The task of this expert system is to determine the aircraft functionality and suggest the pilot how to continue the mission or even abort it. Any expert system is expected to have a "knowledge base" and an "interface engine". Here the knowledge base is consisting of FTA and FMEA documents as offline provided knowledge about how an error will threaten the safe flight and online provided knowledge of what is happening by sensors and indicators measurements and identification tools. The interface engine, on the other hand, is consisting of a simulation subsystem which assessing the knowledge and predicting the upcoming events and the aircraft abilities to accomplish the mission and based on the result it will suggest the future strategies to the pilot.

Other researches in the field of intelligent aircrafts or based on the safety issues and the concept of increasing post failure flight survivability. Thus, the concentration of these researches is on Fault Tolerant Flight Control Systems, Ref 2 and 3. Their operational concept is to provide each aircraft a system that whenever a failure occurred, maintains flight control, and suggest an optimal trajectory to nearest landing location. Ref 4 is describing the system which NASA introduced to fulfill this task, and Ref 5 and Ref 6 explain more details about the system.

S.M.B. Malaek (✉) • Z. Shadram
Aerospace Engineering Department, Sharif University of Technology, Tehran, Iran
e-mail: Malaek@sharif.edu; Shadram@ae.sharif.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_3,

In this paper we will introduce the architecture of our proposed system, and describe more explicitly about its simulation subsystem which was the concentration of this phase of the research. Then the application of the system will presented in an incident which was one of the motivation of the research.

2 System Overview

The purpose of this research is to present the conceptual design of a system which can be used to manage flight while different unexpected events happen. The approach to the idea of this system has some different origins. The first origin is the question that whether it is possible to use the airplane design documents FTA, Fault Tree Analysis, and FMEA, Failure Mode and Effect Analysis, as an online source to manage events during flight. FTA presents that each event or error in each component affects the safety of flight under which scenario. It shows how an error in a component could lead to risk the airplane safety. On the other hand, FMEA presents error modes in each component and the effect that is expected by occurrence of the event. Hence, if there is a possibility to analyze these documents, by the occurrence of an element of FTA, any path which commence with this element will be under suspicion as the probability of occurrence of the path will increase. Then, according to the data included in FMEA, the effects related to the occurrence of the element will be anticipated. As, by the use of system identification and parameter estimation tools, it is possible to identify the problematic element in the airplane system, then it is possible to locate the error on the FTA document and to understand the forgoing situations, the certain and possible effects of the error. Thus, based on the knowledge which is a combination of what identification system and FTA and FMEA analysis will result in, the system will achieve a set of flight conditions and procedures which are the consequences of the events occurred during the flight and affect the airplane systems, and the procedures which are defined by the offline analysis of aviation industry and defined for certain conditions.

When an error occurs there are more effects than what FMEA will present, these effects are due to real flight conditions, so to consider the current flight conditions and to find out how the event affects the mission, there will be a Simulation subsystem to predict what is going to happen during the flight. In other word, based on the simulation results it will be possible to know the flight states and trajectories vs. time and the destination that would be achieved will be identified. These results can update the FTA/FMEA process and will reveal more effects and considerations about the flight. Figure 1 shows a schematic view of how the recognition subsystem and simulation subsystem work together.

This process should be done periodically, to provide a list of reachable destinations including the primary desired landing airport. Then based on the existing conditions, the list should be prioritized. It is clear that while no unexpected event happens, the desired primary landing airport is on top of the list. In figure 2 the flowchart of the subsystem which task is to analysis the simulation results and provides the prioritized list of destinations is presented.

Figure 3 presents the overall architecture of the proposed system. In summary the sensors' data go through the recognition subsystem which includes system identification tools, and online analysis of FTA and FMEA documents. After recognizing the flight conditions, the simulation will be run based on the conditions to predict the future trajectories and states vs. time. At the end, the simulation results will be analyzed to determine the prioritized destination list, which will be announced to the pilot.

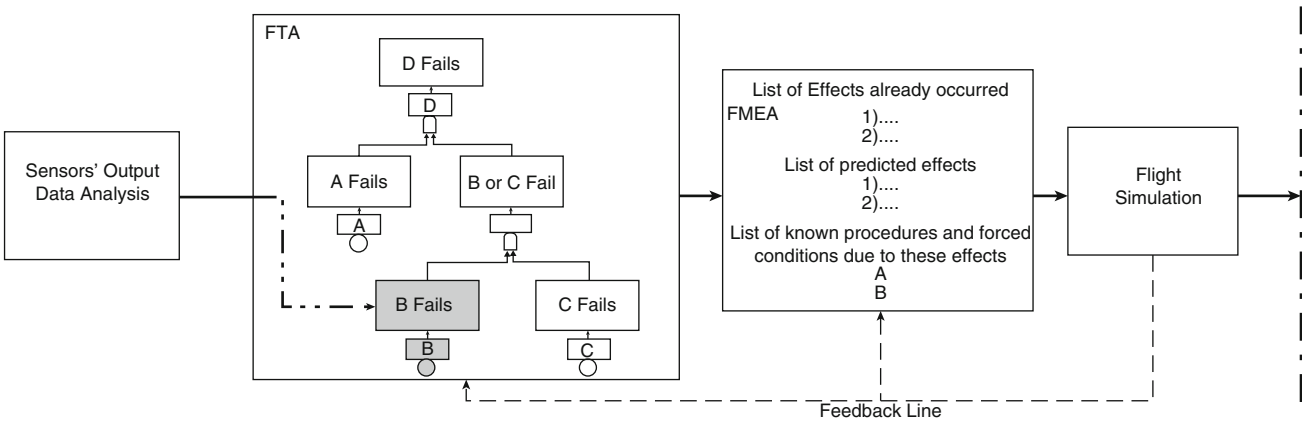


Fig. 1 Architecture of recognition subsystem

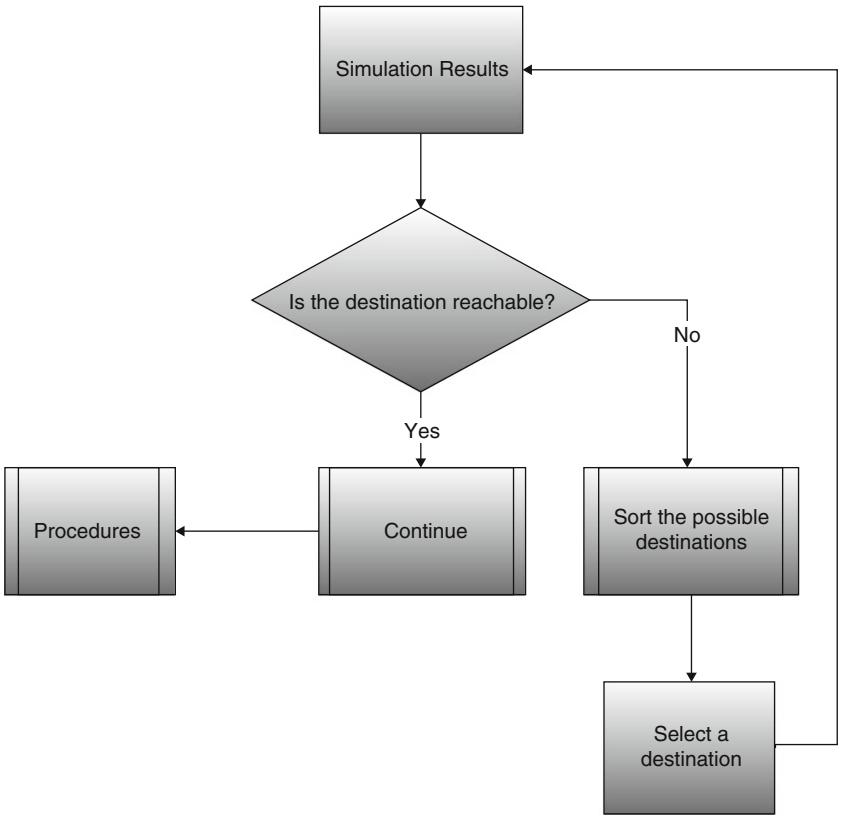


Fig. 2 the flowchart of final processor subsystem

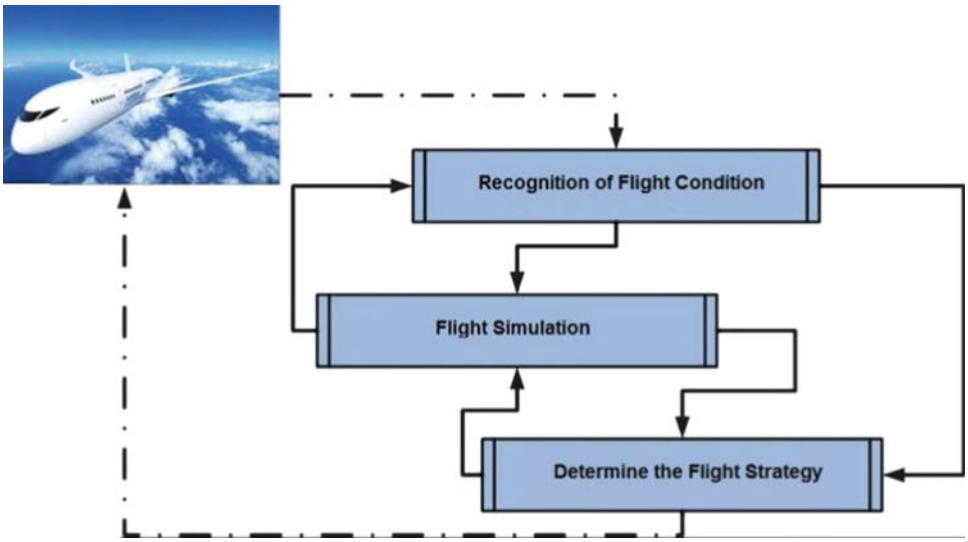


Fig. 3 Overall system architecture

3 Modeling and Simulation

In this phase, the focus of the research was on developing a simulator program which can include damages, such as malfunction in control systems, propulsion system, or airframe structure, and perform the simulation in the least time. Thus, we develop a program which is called Advanced Aircraft Simulator 1.0, which can meet these requirements. To reduce the computation time the linear model of aerodynamic forces are used. To make the simulator capable of simulating damages in airframe structure the equations of motion are written about a reference point which is not center of gravity, therefore if the damage in airframe structure change the center of gravity, the equations are also valid.

The simulator must has all the three loop of pure dynamic, control and stabilization, and navigation system, to be appropriate for the purpose of the system, as it can

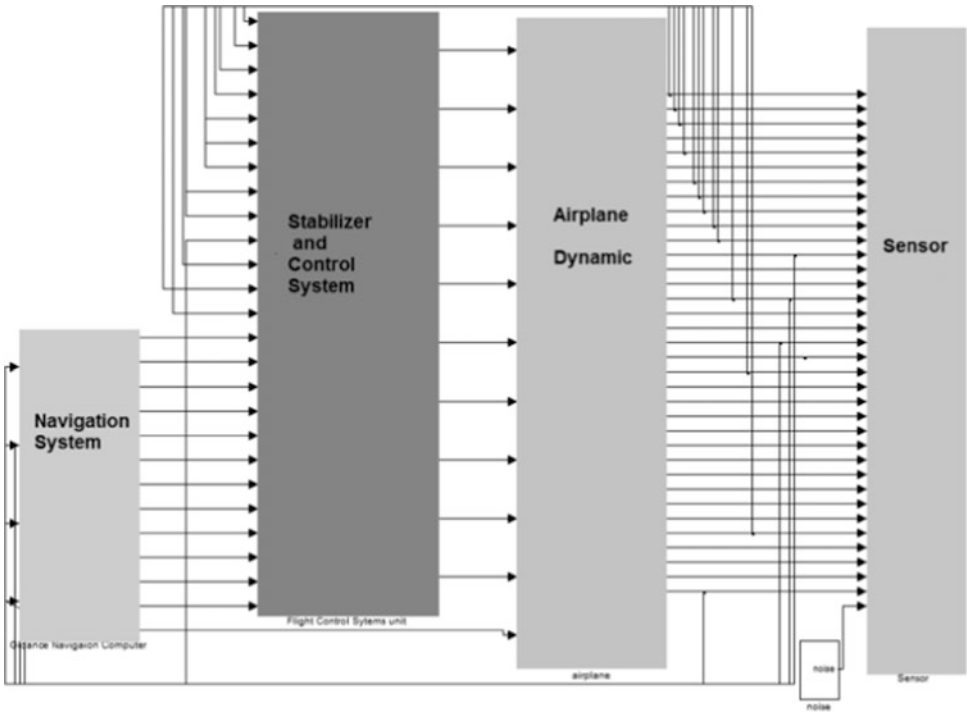


Fig. 4 Simulator Architecture

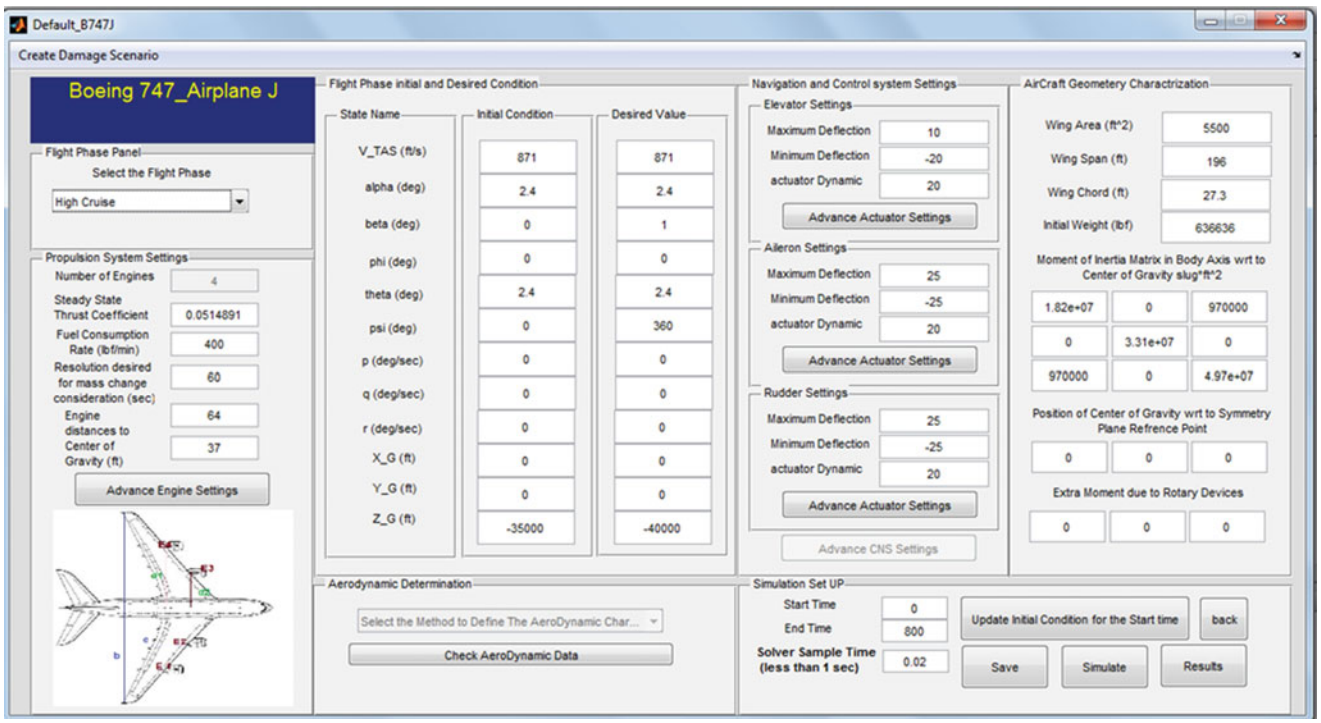


Fig. 5 The simulator program layout

predict the trajectory of the aircraft. Figure 4 shows the simulator architecture.

The assumptions for the modeling and simulation in the program are: subsonic flight, spherical model of the earth for gravity acceleration, solid aircraft. In addition it is assumed that the change in mass due to damage happens such as step function, and the mass change due to fuel consumption is modeled with ramp function. It means mass is an independent variable in the equations.

Figure 5 shows the layout of the simulator program which is developed during the research. It has the ability to make changes in control systems, propulsion system and the

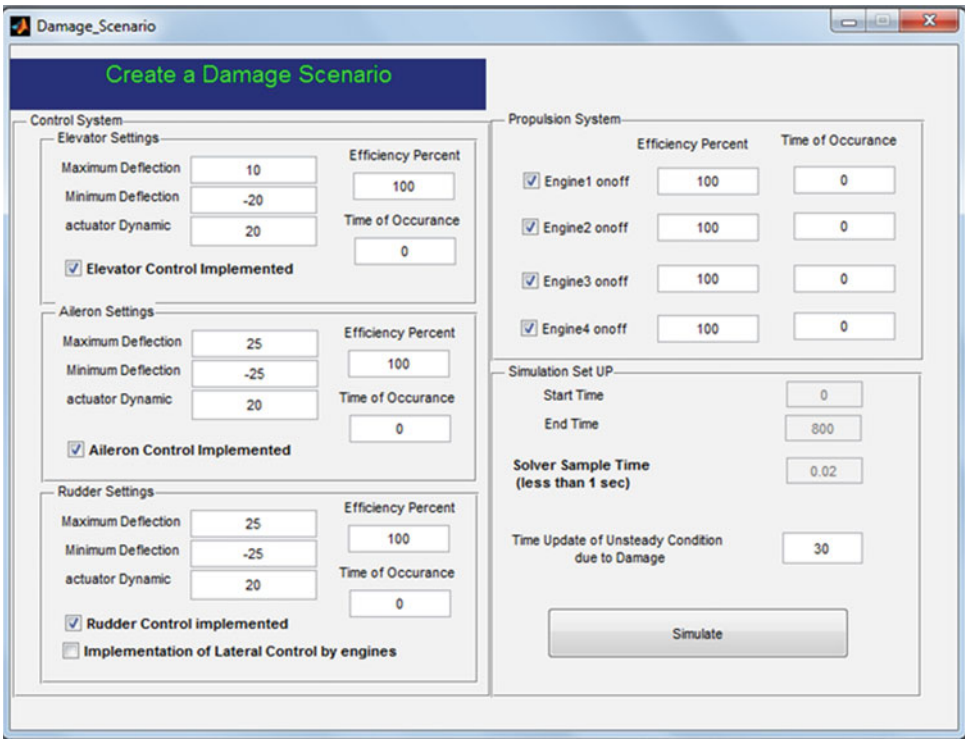


Fig. 6 The module of Create Damage Scenario

airframe weight or the location of center of gravity. The goal of developing this program is to test the theory that if we could identify the errors by the use of FTA and FMEA and identification tools, is that possible to anticipate the trajectories with the simulation, in the incidents which their surviving interpreted as chance or extraordinary abilities of the pilot. So this program has a module which is called "Create Damage Scenario" which is developed to create the incidents scenario for the simulation. Figure 6 shows the layout of this module.

An exclusive innovation in this simulation program is the trim time interval. As we mentioned before, to minimize the computation time in the simulation phase, it is better to use linear aerodynamic model for forces and moment, yet this assumption required to be in a trim manner to validate the perturbation theory; on the other hand, in some damage scenario being in a full trim manner is impossible, for instant in an all engines out scenario there is no speed trim available. To manage the dilemma we assume that we can define an artificial trim condition which should be updated in appropriate time intervals, and use the linear model based on the condition. So we tried this modeling assumption in the simulation process and compare the result with the reality.

The aircraft data that is used in this research as the case is based on aerodynamic derivatives and coefficients related to the Boeing B-747 in high cruise flight condition; Ref 7. It is declared that a B-747 with no engine power will glide with a ratio of 1:15. Thus we try to find a time interval for trim update to produce this glide ratio for the all engine out scenario. Figure 7 presents the predicted trajectories after all engines are shutting down in a cruise condition at 37000 Ft. flight. It shows that in 5 seconds updates the simulation can predict the 1 to 15 glide ratio. In this phase we suggest a simple neural net, to set this simulation settings based on the errors detected. Offline analysis and experiments can find the appropriate settings for discrete scenarios, so for online use, the neural net can choose the settings based on the knowledge base we provide for it in advance based on those experiments.

4 Air Canada: Flight 143

On July 23, 1983, Air Canada Flight 143, a Boeing 767-233 jet, ran out of fuel at an altitude of 41,000 feet, about halfway through its flight originating in Montreal to Edmonton; Ref 8. When they understood there was no propulsive power any more, they had decided to have an emergency landing in Winnipeg Airport. At 39 miles near the airport they found out they would not make this distance and they had decided to land in a retired Gimli Airport in 12 miles of their location, which had no instrument and was not prepared for their landing, they arrived there with a speed more that appropriate, so the pilot used the slip forward maneuver which is familiar for glider pilots, to reduce the speed for performing the landing phase in that runway.By chance he was a glider pilot and surviving through this incident was recognized due to the pilot extra ordinate abilities. Here we show that if our

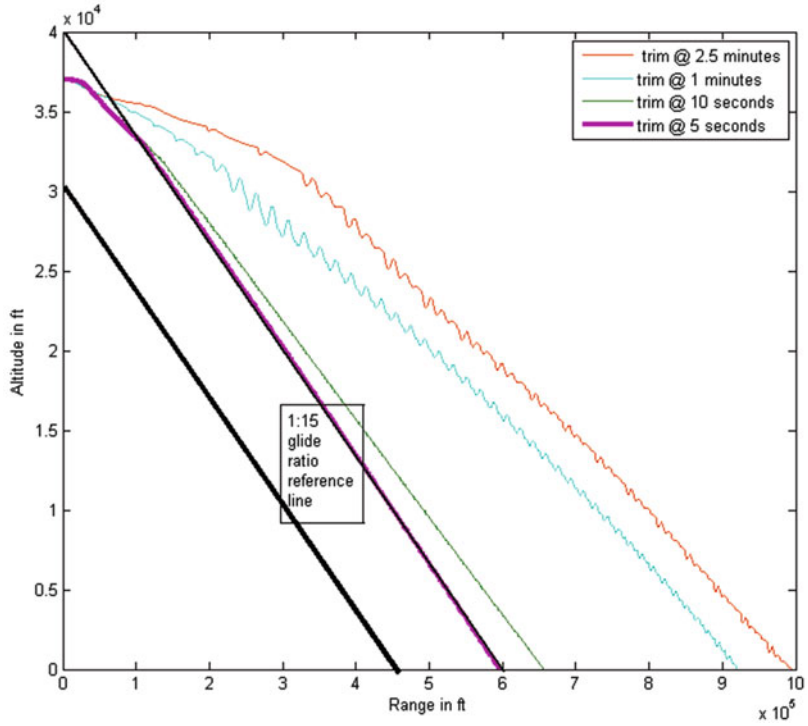


Fig. 7 Altitude vs. Range for All EnginesOut Scenario

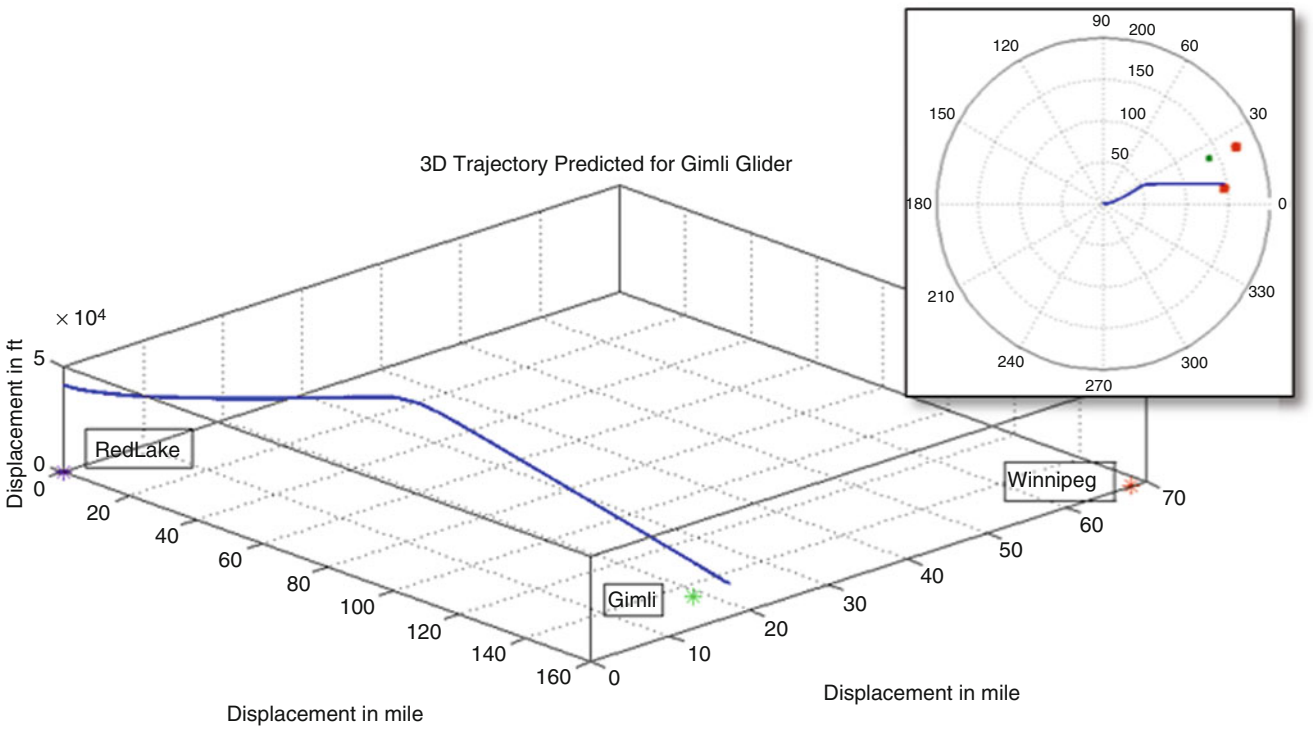


Fig. 8 The predicted trajectory which shows that Winnipeg was an unreachable airport and the farthest airport they could arrive was Gimli, the thumbnail shows changes in heading during the predicted flight trajectory

system were used in that flight, we could anticipate that Winnipeg Airport was an unreachable destination, and the farthest airport they could make was Gimli, and as it was not an in-use airport they would choose another destination or the ground crew would facilitate there for an emergency landing. Besides, they would start the landing maneuver at a point which they could make it without any extra effort. Figure 8 presents the trajectory predicted by the developed

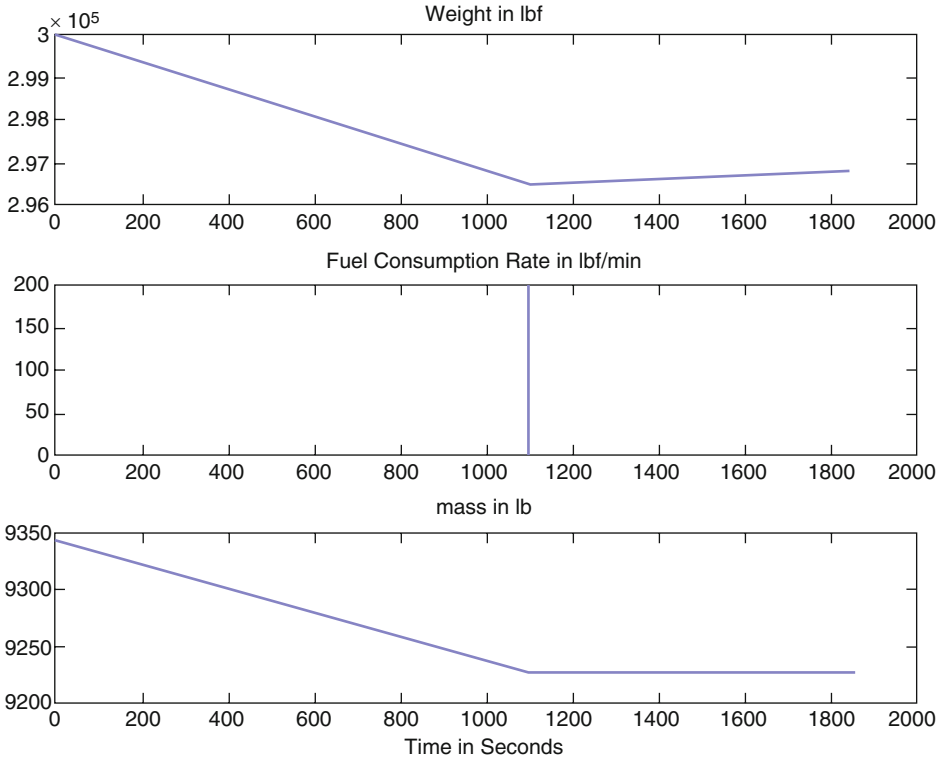


Fig. 9 Weight, mass and fuel consumption rate change in the damage scenario

simulator. The aircraft was a Boing B-767, the simulated scenario is shutting down the first engine at t = 0, and the second engine is shutting down about 20 minutes later.

5 Discussion and Conclusion

In this work, we have effectively demonstrated the suitability of an "Expert System" based on already available documents; known as "FTA" and "FMEA". Current stage of the work, however, has been devoted to prove the viability of such concept, via forensic investigation of some known major incidents. The most interesting point of the proposed conceptis the fact that we do not need to conduct any further engineering work; as most aircraft manufacturers prepare a great-deal of case-studies and documents as part of their effort to develop 'FTA" as well as "FMEA" and as their standard practice during anyaircraft design process. The remaining task is just to enhance or re-architecture such documents to be suitable for the aircraft "Flight Management System" computer. Fortunately, the computing power of existing computers together with affordable hard-disks makes it quite possible to do that with no extra burden. Obviously, whenever a new level of automation is proposed for commercial aircraft is proposed; liability and legal concerns are raised; which require careful investigations. Nonetheless, we expect no serious legal concerns as the aircraft has already been certified based on the same "FTA" and "FMEA" concepts.

We also note that, the main technical aspect of the work would be just the timely identification the flying aircraft. As far as proper "Identification Methods", we are interested in fast identification technique with minimum processing time, which allows identifying and analyzing false sensors data as well as "Sensor mal-functions".

Further studies might also be necessary to expand the existing concept of "Destination" and "Destination Alternate". That is, it might be quite necessary to prepare different "Destination Alternates" based on off-line simulations for any mission between any two possible airports. Fortunately, existing commercial aerial routes are known and so are the available airports.

Moreover, our existing knowledge about geographical elevation-maps is easily accessible; which are quite helpful managing the flight. The knowledge base should be improved by means such as elevation maps of the flight location in addition to climatic situation awareness which can help the system to analyze the possible strategies to continue the flight. This is possible by the data which can be obtained simultaneously from satellites such as GPS. Projecting the predicted trajectory due to the physical constraints on the aircraft on the elevation map of the region the trajectory can be assessed. For instance in the Jakarta Incident, on 24 June 1982, a B_747, Ref 9, flight BA_009 entered a volcanic ash resulting in the failure of all four engines. The aircraft was

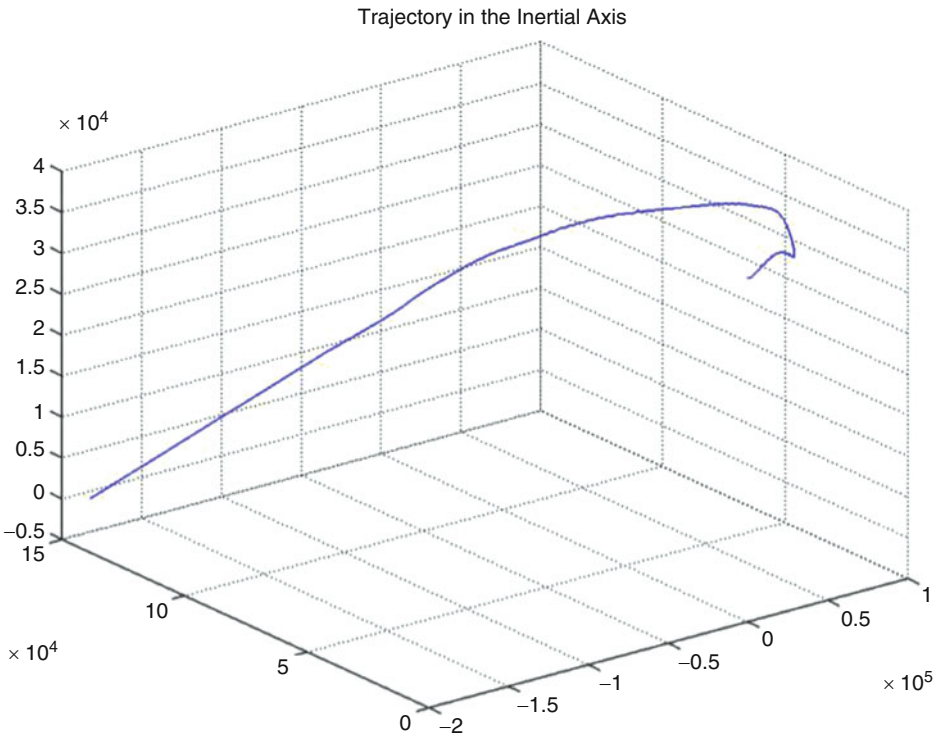


Fig. 10 Flight BA-009 Predicted Trajectory

gliding, they had turned to land in an alternate airport, yet there was Java mountains in the way which they have no idea whether they could make the require altitude for clearing the mountain or not. If the predicted trajectory shown in figure 10 were projected on the elevation map of the region, they could decide more precisely about continuing the glide path, or ditching in the ocean which was their second option for the time that it became obvious that is impossible to clear the mountain. Actually when they lost altitude, they passed the ash region and they became able to restart the engines, and that time they did not encounter the dilemma that critically. Yet, if the engines could not be restarted, awareness of what are their real abilities helps them to decide better and choose better trajectories. If they knew they should ditch in the ocean, it would better to control their glideslope for the best touch down path. Also, nowadays that we know by passing the ashes the engines will be restarted, our expert system should suggest the trajectory which will exit the ash cloud optimally.

The proposed concept can also be expanded by the so called "Aircraft Energy State Formulation" to consider air-traffic around any given aircraft. This is powerful concept that allows us to prioritize alternate trajectories in congested routes. In fact, we could use aircraft fuel energy consumption rate as a measure to prioritize, in case of multiple emergencies.

References

Sheridan,T.B.;Verplank,W."Human and Computer Control of Undersea Tele-operators" Cambridge, MA: Man-Machine Systems Laboratory, Department of Mechanical Engineering, MIT. 1978

NhanNguyan; KalmanjeKrishnakumar; John Kaneshige; Pascal Nespeca; "Dynamic and Adaptive Control for Stability Recovery of Damaged Asymmetric Aircraft", *AIAA Guidance, Navigation and Control*, Colorado, 2006.

T.J.J. Lombaert,; Q.P. Chu; J.A. Mulder;D.A. Joosten; "Real Time Damaged Aircraft Model Identification for Reconfiguring Flight Control", AIAA Atmospheric Flight Mechanics, South Carolina, 2007

Liang Tang; Michael Roemer; JianhuaGe; Agamemnon Crassidis; J.V. R. Prasad; Christine Belcastro; "Methodologies for Adaptive Flight Envelope Estimation and Protection", AIAA Guidance, Navigation, and Control Conference Chicago, Illinois, Aug. 2009.

Zhang, X., Polycarpou, M. M., and Parisini, T., Design and analysis of a fault isolation scheme for a class of uncertain nonlinear systems, IFAC Annual Reviews in Control, Volume 32, Issue 1, pp. 107-121, 2008.

3Nguyen, N., Krishnakumar, K., Kaneshige, J., and Nespeca, P., Flight Dynamics and Hybrid Adaptive Control of Damaged Aircraft, Journal of Guidance, Control, and Dynamics, Vol. 31, No. 3, Pp 751-764, May–June 2008.

Jan Roskam; Airplane Flight Dynamics And Automatic Flight Controls (Part I, II), 3rd Edition, DARcorporation, 2001.

Flight Safety Foundation; Aviation Safety Network; http://aviation-safety.net/database/record.php?id = 19830723-0

Stewart, Stanley; *Emergency: Crisis on the Flight Deck*, 2nd Edition, The Crowood Press, 2002.

Nonlinear Optimal Tracking With Incomplete State Information Using State Dependent Riccati Equation

Ahmed Khamis, D. Subbaram Naidu, and Dawid Zydek

1 Introduction

Kalman filter is an effective minimum variance linear state estimator that estimates the unmeasured system states corrupted with white process and measurement noise. The standard Kalman filter is limited only to linear systems. Most real-world systems are nonlinear, in which case standard Kalman filters are not applicable [15]. Therefore, it becomes necessary to use some other nonlinear filter techniques. The extended Kalman filter (EKF) is the most widely applied state estimation algorithm for nonlinear systems. The EKF is used to estimate the unmeasured states of nonlinear systems. The EKF relies on linearization of the nonlinear system using Taylor series expansion near the operating point [12]. In linearization, we assume that the range of operation is small. Consequently, the EKF will only be effective in the small neighborhood of the operating points, and the accuracy of this technique will decrease for large operating range of nonlinear systems.

The need to improve performance in control systems requires more and more accurate modeling [10]. However, if a model is a good representation of the real system over a wide range of operating points, it is most often nonlinear [3]. There exist many nonlinear control design techniques, each has benefits and flaws. Selecting the suitable control technique for nonlinear system usually requires consideration of different factors, e.g. performance, optimality, and cost. One of the highly promising and rapidly developing techniques for nonlinear optimal controllers is the State Dependent Riccati Equation (SDRE) technique [13]. The SDRE has become a very attractive tool for the systematic design of nonlinear controllers, very common within the control community over the last decade. The SDRE is an extremely effective algorithm for nonlinear feedback control design by allowing nonlinearities in the system states while additionally offering great design flexibility through design matrices [3].

Inspired by the great potential of the SDRE for infinite horizon optimal control of nonlinear systems [6], this paper offers a new technique for optimal tracking of nonlinear stochastic systems. This is accomplished by integrating the Kalman filter with the SDRE technique. Kalman filter is used to estimate the unmeasured states which are corrupted with noises in the nonlinear model.

The structure of the paper is as follows: Section II presents a brief overview of SDRE. Section III discuses the idea of standard Kalman filter. Section IV presents the continuous time tracking for optimal nonlinear stochastic systems. Finally, conclusions of the paper are in Section V.

2 Infinite-Horizon SDRE Technique

SDRE, which is also referred to as the Frozen Riccati Equation (FRE) [7], first proposed by Pearson (1962) and later expanded by Wernli & Cook (1975), and studied by Mracek & Cloutier (1998) [5]. SDRE has become as a very attractive tool for the systematic design of nonlinear controllers, very common within the control community over the last decade, providing an extremely effective algorithm for nonlinear feedback control design by allowing nonlinearities in the system states while additionally offering great design flexibility through design matrices [1]. The method involves factorization of the nonlinear dynamics into product of a matrix-valued function. Thus, the SDRE algorithm captures the nonlinearities of the system, transforming the original nonlinear system to a linear-like structure with state dependent coefficient (SDC) matrices, and minimizing a nonquadratic performance index with a quadratic-like structure [3]. The Riccati equation using the SDC matrices is then solved online to give the sub optimum control law. Moreover, with enough sample points, the suboptimal solution

A. Khamis (✉) • D.S. Naidu • D. Zydek
Department of Electrical Engineering, Idaho State University, Pocatello, ID, USA
e-mail: khamahme@isu.edu; naiduds@isu.edu; zydedawi@isu.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_4,

can be made to be very close to optimal solution. The coefficients of this Riccati equation vary with the each point in state space. The algorithm thus involves solving, at a given point in state space, a SDRE whose point wise stabilizing solution during state evolution yields the SDRE nonlinear feedback control law. As the SDRE depends only on the current state, the computation can be carried out online, in which case the SDRE is defined along the state trajectory. In addition, a primary advantage offered by SDRE to the control designer is the opportunity to make tradeoffs between control effort and state errors by tuning the SDC.

The computation aspect, applied here to SDRE can be solved in many ways. One of the approaches passes the processing and other possible data to the ground infrastructure of computational units [11, 16]. These units may be of different scale and location, but due to flexible architecture of distributed processing systems it is perceived as the single entity. The example of the system that applies for SDRE processing is the P2P computing system presented in [4]. Authors provided the complete solution for geographically spread processing with extended results delivery mechanisms. This way, the results can be effectively delivered to the SDRE unit.

SDRE control theory has only been developed for the infinite-time non-linear optimal regulation (stabilization) problem, for which the reference signal $\mathbf{z}(t) = 0$ and $\mathbf{C}(\mathbf{x}) = I_{n\mathbf{x}n}$. This is because the method needs solving the infinite-time algebraic Riccati equation. Unluckily, the undeveloped theory of the infinite-time LQ optimal tracking problem has hindered its application for solving non-linear trajectory tracking problems, unless an integral servomechanism is used [3], which increases the number of states and thus the computation time required for solving algebraic Riccati equations. Regardless of the undeveloped theory of infinite-time LQ optimal tracking control, a good approximation can be developed for excessively large terminal time [2]. The derived results are approximate in nature and are valid for very large values of the terminal time.

The state-feedback optimal controller is obtained in the form

$$\mathbf{u}(\mathbf{x}) = -\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}'(\mathbf{x})[\mathbf{P}(\mathbf{x})\mathbf{x}(t) - \mathbf{g}(\mathbf{x})]. \tag{1}$$

P(x) is a positive-definite solution of the continuous-time SDRE

$$\begin{aligned}\mathbf{P}(\mathbf{x})\mathbf{A}(\mathbf{x}) + \mathbf{A}'(\mathbf{x})\mathbf{P}(\mathbf{x}) - \mathbf{P}(\mathbf{x})\mathbf{B}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}'(\mathbf{x})\mathbf{P}(\mathbf{x}) \\ + \mathbf{C}'(\mathbf{x})\mathbf{Q}(\mathbf{x})\mathbf{C}(\mathbf{x}) = 0,\end{aligned} \tag{2}$$

and **g(x)** is a solution of the continuous-time State Dependent non-homogeneous equation

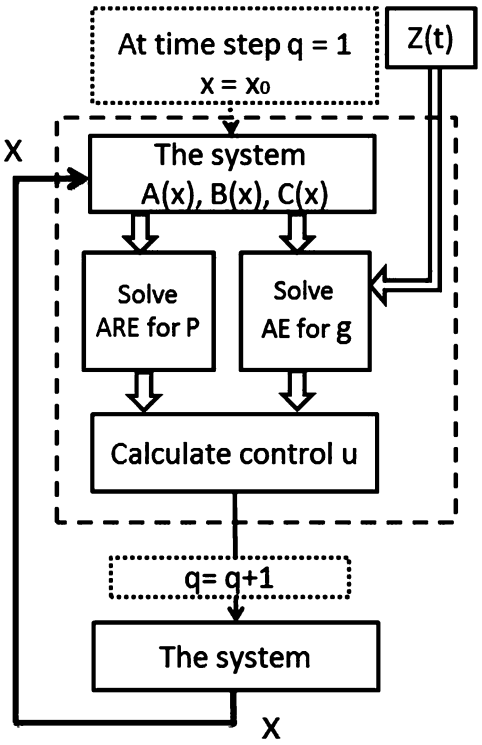


Fig. 1 Process of Infinite-Horizon SDRE Tracking

$$\mathbf{g}(\mathbf{x}) = -\left([\mathbf{A}(\mathbf{x}) - \mathbf{B}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}'(\mathbf{x})\mathbf{P}(\mathbf{x})]'\right)^{-1}\mathbf{C}'(\mathbf{x})\mathbf{Q}(\mathbf{x})\mathbf{z}(\mathbf{x}). \tag{3}$$

The resulting SDRE-controlled trajectory is the solution of the closed-loop dynamics

$$\begin{aligned}\dot{\mathbf{x}}(t) = [\mathbf{A}(\mathbf{x}) - \mathbf{B}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}'(\mathbf{x})\mathbf{P}(\mathbf{x})]\mathbf{x}(t) \\ + \mathbf{B}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}'(\mathbf{x})\mathbf{g}(\mathbf{x}).\end{aligned} \tag{4}$$

Fig. 1 shows a process of the infinite-horizon SDRE tracking [9]. At each sample time, the following procedure is accomplished. First, the current state vector $\mathbf{x}(t)$ is used to calculate numerical values for **A(x)**, **B(x)**, and **C(x)**. Then, using the LQT equations, **P(x)** and **g(x)** are calculated. Control input **u(x)** is then calculated and applied to the system. This procedure is then repeated at the next sample time. Because of its approximating nature, the SDRE technique is considered a suboptimal solution. However, with the proper choices for the **A(x)**, **B(x)**, and **C(x)** matrices, and with the proper amount of sample times, the SDRE technique can provide a very adequate optimal solution.

3 Standard Kalman Filter

The Kalman filter was developed by Rudolf E. Kalman in 1960 [8]. The Kalman filter can be used to estimate the states of continuous-time or discrete-time linear systems.

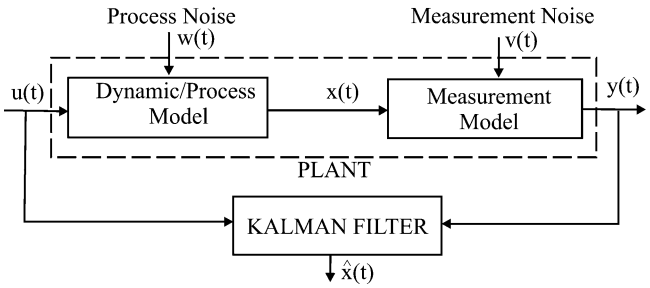


Fig. 2 Linear Continuous-Time Kalman Filter

Here, a brief overview of continuous-time Kalman filter for linear systems, which is needed in later sections, is given.

Consider the linear, continuous-time, stochastic system with dynamic model:

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{B}_w(t)\mathbf{w}(t),\\ \mathbf{y}(t) &= \mathbf{C}(t)\mathbf{x}(t) + \mathbf{v}(t),\end{aligned} \tag{5}$$

where, $\mathbf{w}(t)$ and $\mathbf{v}(t)$ are process, and measurement (white, Gaussian) random noises with zero mean (i.e., $\bar{\mathbf{w}}(t) = \bar{\mathbf{v}}(t) = 0$) and covariances $\mathbf{Q}_w(t)$ and $\mathbf{R}_v(t)$, respectively, and assumed to be *uncorrelated*, (see Fig. 2).

The estimated state $\hat{\mathbf{x}}(t)$ is given by:

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{A}(t)\hat{\mathbf{x}}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{K}_e(t)[\mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t)], \tag{6}$$

$$\dot{\hat{\mathbf{x}}}(t) = [\mathbf{A}(t) - \mathbf{K}_e(t)\mathbf{C}(t)]\hat{\mathbf{x}}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{K}_e(t)\mathbf{y}(t), \tag{7}$$

where, $\mathbf{K}_e(t)$ is the estimator gain, $\hat{\mathbf{x}}(t)$ is the *state estimate* with initial value

$$\mathscr{E}\{\mathbf{x}(t=0)\} = \bar{\mathbf{x}}(t_0) = \hat{\mathbf{x}}(t_0), \tag{8}$$

where, $\mathscr{E}$ stands for *expected, average or mean value* and considered intuitively equal to the *estimate*.

Let us define the error $\mathbf{e}(t)$ between the true or actual state $\mathbf{x}(t)$ and the state estimate $\hat{\mathbf{x}}(t)$ as

$$\mathbf{e}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t); \tag{9}$$

$$\dot{\mathbf{e}}(t) = \dot{\mathbf{x}}(t) - \dot{\hat{\mathbf{x}}}(t), \tag{10}$$

Substituting (5) and (7) in (10)

$$\dot{\mathbf{e}}(t) = \mathbf{A}(t)\mathbf{e}(t) + \mathbf{K}_e(t)\mathbf{C}(t)\mathbf{e}(t) + \mathbf{B}_w(t)\mathbf{w}(t) - \mathbf{K}_e(t)\mathbf{v}(t), \tag{11}$$

$$\dot{\mathbf{e}}(t) = [\mathbf{A}(t) - \mathbf{K}_e(t)\mathbf{C}(t)]\mathbf{e}(t) + \mathbf{B}_{wk}(t)\mathbf{z}_{wk}(t), \tag{12}$$

where

$$\mathbf{B}_{wk}(t) = [\mathbf{B}_w(t) - \mathbf{K}_e(t)]; \qquad \mathbf{z}_{wk} = [\mathbf{w}(t)\mathbf{v}(t)]', \tag{13}$$

using the results from [14] on propagation of state vector

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}_w(t)\mathbf{w}(t), \tag{14}$$

and the corresponding state estimate error covariance $\mathbf{P}_e(t)$ which can be calculated from

$$0 = \mathbf{A}(t)\mathbf{P}_e(t) + \mathbf{P}_e(t)\mathbf{A}'(t) + \mathbf{B}_w(t)\mathbf{Q}_w(t)\mathbf{B}_w'(t). \tag{15}$$

Now, using the result (15) for the error dynamics (12)

$$\begin{aligned}0 = {} & [\mathbf{A}(t) - \mathbf{K}_e(t)\mathbf{C}(t)]\mathbf{P}_e(t) + \mathbf{P}_e(t)[\mathbf{A}(t) - \mathbf{K}_e(t)\mathbf{C}(t)]'\\ & + [\mathbf{B}_w(t) - \mathbf{K}_e(t)]\begin{bmatrix}\mathbf{Q}_w(t)\\ \mathbf{R}_v(t)\end{bmatrix}[\mathbf{B}_w(t) - \mathbf{K}_e(t)]',\end{aligned} \tag{16}$$

$$\begin{aligned}0 = {} & [\mathbf{A}(t) - \mathbf{K}_e(t)\mathbf{C}(t)]\mathbf{P}_e(t) + \mathbf{P}_e(t)[\mathbf{A}(t) - \mathbf{K}_e(t)\mathbf{C}(t)]'\\ & + [\mathbf{B}_w(t)\mathbf{Q}_w(t)\mathbf{B}_w'(t) + \mathbf{K}_e(t)\mathbf{R}_v(t)\mathbf{K}_e'(t),\end{aligned} \tag{17}$$

where, $\mathbf{P}_e = \mathbf{P}_e(t) = \mathscr{E}\{[\mathbf{x}(t)\text{-}\hat{\mathbf{x}}(t)][\mathbf{x}(t)\text{-}\hat{\mathbf{x}}(t)]'\}$ is to be solved in *forward* direction with initial condition

$$\mathbf{P}_{e0} = \mathbf{P}_e(t_0) = \mathscr{E}\{[\mathbf{x}(t_0) - \hat{\mathbf{x}}(t_0)][\mathbf{x}(t_0) - \hat{\mathbf{x}}(t_0)]'\}. \tag{18}$$

We have the condition on $\mathbf{K}_e(t)$ for minimum error variance as

$$\mathbf{K}_e(t) = \mathbf{P}_e(t)\mathbf{C}'(t)\mathbf{R}_v^{-1}(t). \tag{19}$$

Using the optimal Kalman gain (19) in the covariance relation (17), we get

$$\begin{aligned}0 = {} & \mathbf{A}(t)\mathbf{P}_e(t) + \mathbf{P}_e(t)\mathbf{A}'(t) - \mathbf{P}_e(t)\mathbf{C}'(t)\mathbf{R}_v^{-1}(t)\mathbf{C}(t)\mathbf{P}_e(t)\\ & + \mathbf{B}_w(t)\mathbf{Q}_w(t)\mathbf{B}_w'(t),\end{aligned} \tag{20}$$

with initial condition $\mathbf{P}_e(t=0) = \mathbf{P}_{e0}$. This is called the matrix continuous, algebraic Riccati equation (CARE) arising in optimal state estimation.

Fig. 3 shows a structure of the standard linear continuous-time Kalman filter.

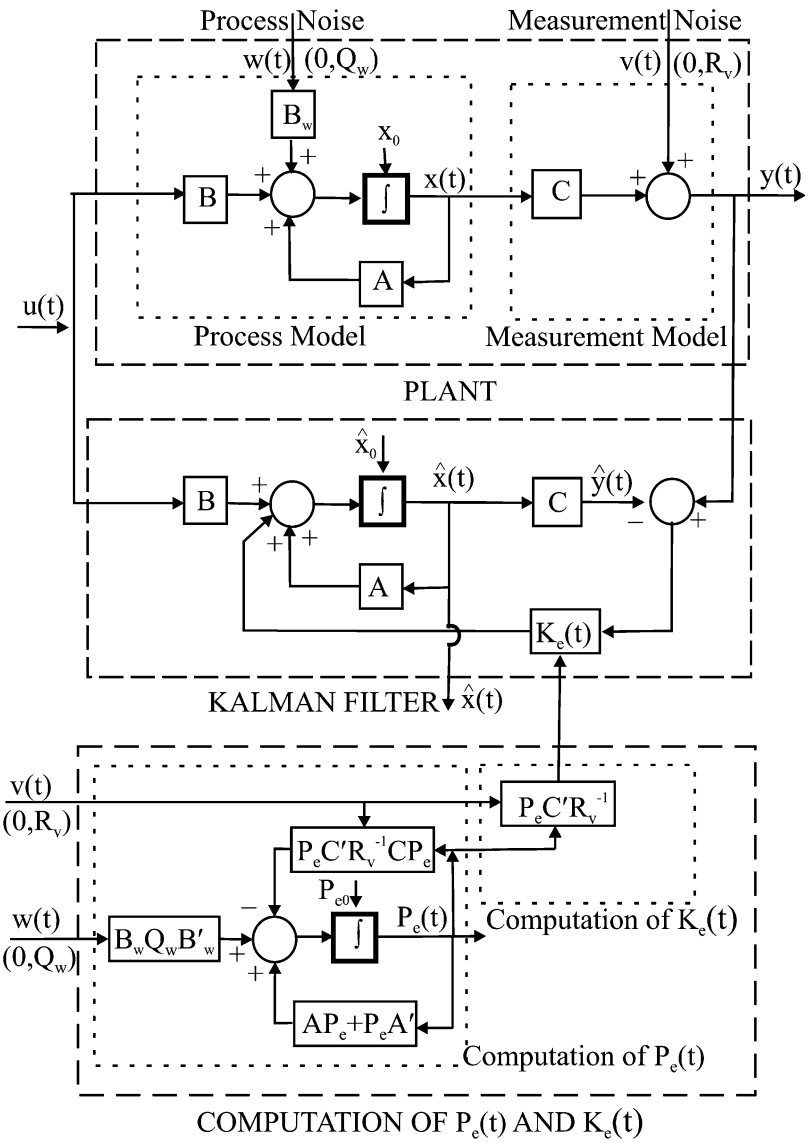


Fig. 3 Standard Linear Continuous-Time Kalman Filter

4 The Continuous-Time Tracking for Nonlinear Stochastic Systems

4.1 Optimal Estimation

Let us reproduce the nonlinear system with noises in state dependent form

$$\dot{\mathbf{x}}(t) = \mathbf{A}(\mathbf{x})\mathbf{x}(t) + \mathbf{B}(\mathbf{x})\mathbf{u}(t) + \mathbf{B}_w(t)\mathbf{w}(t), \tag{21}$$

$$\mathbf{y}(t) = \mathbf{C}(\mathbf{x})\mathbf{x}(t) + \mathbf{v}(t). \tag{22}$$

In order to find the best estimate $\widehat{\mathbf{x}}(t)$ and the corresponding covariance matrix $\mathbf{P}_e(\widehat{\mathbf{x}}, t)$, we use the results of Sec. 3. At each time step, the estimate equations are

$$\begin{aligned} \dot{\widehat{\mathbf{x}}}(t) &= \mathbf{A}(\widehat{\mathbf{x}})\widehat{\mathbf{x}}(t) + \mathbf{B}(\widehat{\mathbf{x}})\mathbf{u}(t) + \mathbf{K}_e(\widehat{\mathbf{x}}, t)[\mathbf{y}(t) - \mathbf{C}(\widehat{\mathbf{x}})\widehat{\mathbf{x}}(t)]; \\ \widehat{\mathbf{x}}(t_0) &= \overline{\mathbf{x}}(t_0), \end{aligned} \tag{23}$$

where, $\mathbf{K}_e(\widehat{\mathbf{x}}, t)$, the optimal Kalman estimator gain, is obtained as as

$$\mathbf{K}_e(\widehat{\mathbf{x}}, t) = \mathbf{P}_e(\widehat{\mathbf{x}}, t)\mathbf{C}'(\widehat{\mathbf{x}})\mathbf{R}_v^{-1}(t), \tag{24}$$

and $\mathbf{P}_e(\widehat{\mathbf{x}}, t)$ is the solution of the matrix algebraic Riccati equation

$$\begin{aligned} 0 = {} & \mathbf{A}(\widehat{\mathbf{x}})\mathbf{P}_e(\widehat{\mathbf{x}}, t) + \mathbf{P}_e(\widehat{\mathbf{x}}, t)\mathbf{A}'(\widehat{\mathbf{x}}) + \mathbf{B}_w(t)\mathbf{Q}_w(t)\mathbf{B}_w'(t) \\ & - \mathbf{P}_e(\widehat{\mathbf{x}}, t)\mathbf{C}'(\widehat{\mathbf{x}})\mathbf{R}_v^{-1}(t)\mathbf{C}(\widehat{\mathbf{x}})\mathbf{P}_e(\widehat{\mathbf{x}}, t), \end{aligned} \tag{25}$$

where $\mathbf{P}_e(\widehat{\mathbf{x}}, t_0) = \mathbf{P}_{e0}$.

The minimization of J is equivalent to minimization of

$$J_a(\mathbf{x},u)=\mathscr{E}\left\{\frac{1}{2}\int_0^{\infty}[\widehat{\mathbf{e}}'(t)\mathbf{Q}(\widehat{\mathbf{x}})\widehat{\mathbf{e}}(t)+\mathbf{u}'(\widehat{\mathbf{x}},t)\mathbf{R}(\widehat{\mathbf{x}})\mathbf{u}(\widehat{\mathbf{x}},t)dt\right\}, \tag{26}$$

where $\widehat{\mathbf{e}}(t)=\mathbf{z}(t)\text{-}\mathbf{C}(\widehat{\mathbf{x}})\widehat{\mathbf{x}}(t)$.

4.2 Optimal Tracking

At each time step, using the results of nonlinear tracking obtained in Sect. 2 except that the state is now the optimal estimate $\widehat{\mathbf{x}}(t)$.

$$\mathbf{u}(\widehat{\mathbf{x}},t)=-\mathbf{R}^{-1}(\widehat{\mathbf{x}})\mathbf{B}'(\widehat{\mathbf{x}})[\mathbf{P}_c(\widehat{\mathbf{x}},t)\widehat{\mathbf{x}}(t)-\mathbf{g}(\widehat{\mathbf{x}},t)], \tag{27}$$

where, $\mathbf{P_c(x)}$ is a positive-definite solution of the continuous-time State Dependent Riccati Equation

$$\mathbf{P_c}(\widehat{\mathbf{x}})\mathbf{A}(\widehat{\mathbf{x}})+\mathbf{A}'(\widehat{\mathbf{x}})\mathbf{P_c}(\widehat{\mathbf{x}})-\mathbf{P}(\widehat{\mathbf{x}})_{\mathbf{c}}\mathbf{B}(\widehat{\mathbf{x}})\mathbf{R}^{-1}(\widehat{\mathbf{x}})\mathbf{B}'(\widehat{\mathbf{x}})\mathbf{P_c}(\widehat{\mathbf{x}})+\mathbf{C}'(\widehat{\mathbf{x}})\mathbf{Q}(\widehat{\mathbf{x}})\mathbf{C}(\widehat{\mathbf{x}})=0, \tag{28}$$

and $\mathbf{g(x)}$ is a solution of the continuous-time State Dependent non-homogeneous equation

$$\mathbf{g}(\widehat{\mathbf{x}})=-\left([\mathbf{A}(\widehat{\mathbf{x}})-\mathbf{B}(\widehat{\mathbf{x}})\mathbf{R}^{-1}(\widehat{\mathbf{x}})\mathbf{B}'(\widehat{\mathbf{x}})\mathbf{P_c}(\widehat{\mathbf{x}})]'\right)^{-1}\mathbf{C}'(\widehat{\mathbf{x}})\mathbf{Q}(\widehat{\mathbf{x}})\mathbf{z}(\widehat{\mathbf{x}}), \tag{29}$$

4.3 Simulation Results

For numerical simulation and analysis, the developed estimation and optimal tracking technique is implemented for noise cancellation for inverted pendulum controlled by DC motor, as shown in Fig. 4.

The dynamic equations for system under concern are:

$$V(t) = L\frac{di(t)}{dt}+Ri(t)+k_b\frac{d\theta(t)}{dt}, \tag{30}$$

$$ml^2\frac{d^2\theta(t)}{dt^2} = -mglsin(\theta(t))-k_m i(t), \tag{31}$$

where, V is the control voltage, L is the motor inductance,i is the current through the motor winding, R the motor winding resistance,k_b the motor's back electro magnetic force constant, θ the angle of pendulum, m the mass of pendulum, l the length of rod, g the gravitational constant, and k_m the damping (friction) constant.

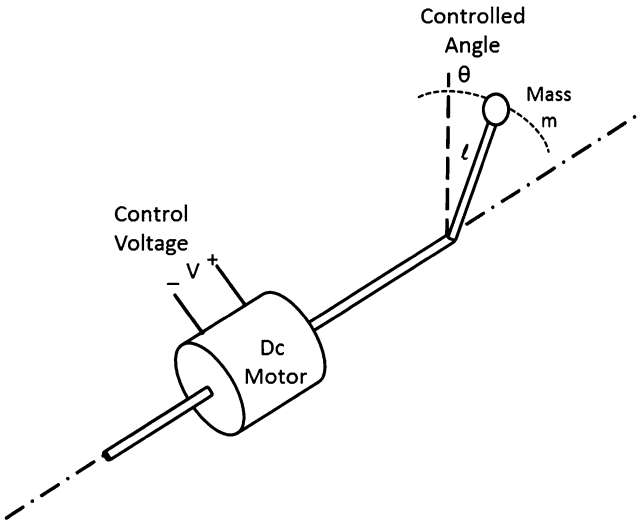


Fig. 4 Inverted Pendulum Controlled by DC Motor

The system nonlinear state equations can be written in the state dependent form:

$$\begin{bmatrix}\dot{x_1}\\ \dot{x_2}\\ \dot{x_3}\end{bmatrix}=\begin{bmatrix}0 & 1 & 0\\ \frac{g}{l}\frac{sin(x_1)}{x_1} & 0 & \frac{k_m}{ml^2}\\ 0 & -\frac{k_b}{L} & \frac{R}{L}\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}+\begin{bmatrix}0\\ 0\\ \frac{1}{L}\end{bmatrix}u. \tag{32}$$

where: $\theta=x_1$, $\dot{\theta}=x_2$, $i=x_3$, $V=u$.

Let the selected weighted matrices are

$$\mathbf{Q}=diag(100,0,0),\mathbf{R}=0.07,\mathbf{F}=diag(1,1,1). \tag{33}$$

The covariances of the noises have been taken as

$$Q_w=diag(0.2,0.2,0.2),R_v=100. \tag{34}$$

The simulations are performed for 300 time steps and the resulting angle trajectories is shown in Fig. 5, where the dash-dot line denotes the *reference* angle trajectory, the dashed line denotes the *actual* angle, and the solid line denotes the *estimated* angle. The optimal control voltage is shown in Fig. 6, where the solid line denotes the estimated optimal control and the doted line denotes the actual optimal control signal.

Comparing these trajectories in Fig. 5, it's clear that the propose algorithm gives very good results as the estimated optimal angle is making a very good tracking to the reference angle, and also the estimated trajectory is very close to the actual output trajectory, the average error for this example is 0.02%.

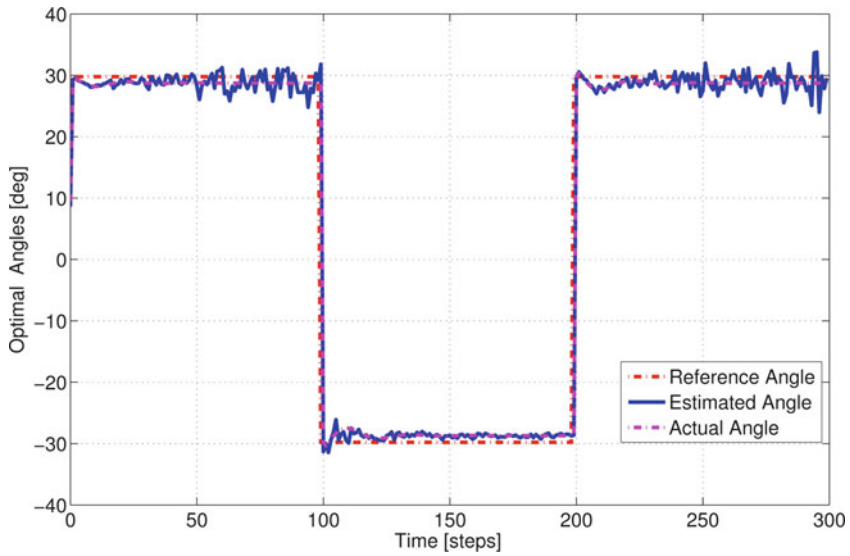


Fig. 5 Angle Trajectories for Inverted Pendulum Controlled by DC Motor

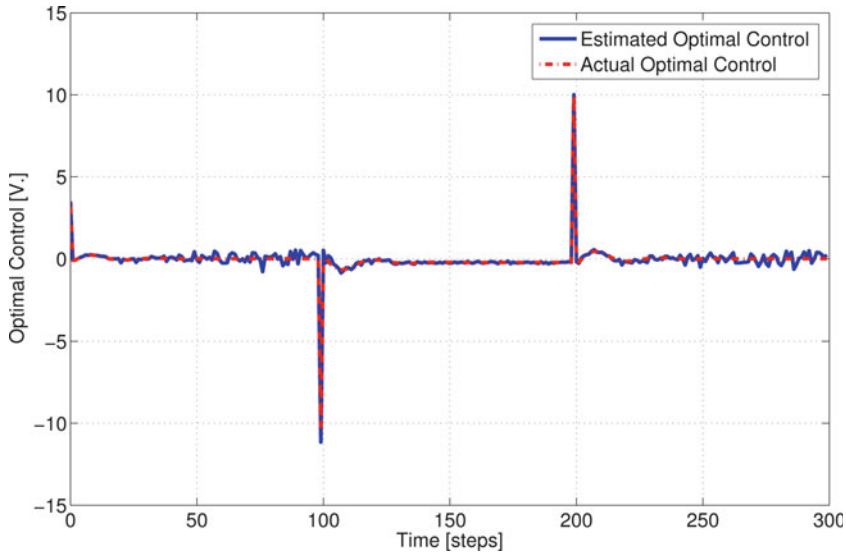


Fig. 6 Optimal Control for Inverted Pendulum Controlled by DC Motor

5 Conclusions

This paper presents a new efficient online technique used for nonlinear stochastic regulator and tracking problems. The idea of the proposed technique is to integrate the Kalman filter algorithm and the SDRE technique. Unlike the ordinary methods which deal with the linearized system, this technique will estimate the unmeasured states of the nonlinear system directly, and this will make the proposed technique effective for wide range of operating points.

References

1. Çimen, T.: State-dependent Riccati equation (SDRE) control: a survey. Proc. of the 17th IFAC World Congress pp. 3761–3775 (2008)
2. Çimen, T.: Development and validation of a mathematical model for control of constrained nonlinear oil tanker motion. Mathematical and Computer Modeling of Dynamical Systems 15(1), 1749 (2009)
3. Çimen, T.: Survey of state-dependent Riccati equation in nonlinear optimal feedback control synthesis. AIAA Journal of Guidance, Control, and Dynamics 35(4), 1025–1047 (2012)
4. Chmaj, G., Walkowiak, K.: A p2p computing system for overlay networks. Future Generation Computer Systems 29(1), 242–249 (2013)

5. Cloutier, J.R., Stansbery, D.: Nonlinear, hybrid bank-to-turn/skid-to-turn autopilot design. Proceedings of the AIAA Guidance, Navigation, and Control Conference (2001)
6. Cloutier, J.: State-dependent Riccati equation techniques: An overview. Proc. American Control Conference 2, 932–936 (1997)
7. Doyle, J., Huang, Y., Primbs, J., Freeman, R., Murray, R., Packard, A., Krstic, M.: Nonlinear control: Comparisons and case studies. In Notes from the Nonlinear Control Workshop conducted at the American Control Conference, Albuquerque, New Mexico (1998)
8. Kalman, R.: A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82(1), 35–45 (1960)
9. Khamis, A., Naidu, D.: Nonlinear optimal tracking using finite-horizon state dependent Riccati equation (SDRE). Proceedings of the 4th International Conference on Circuits, Systems, Control, Signals (WSEAS) pp. 37–42 (August 2013), valencia, Spain
10. Khamis, A., Zydek, D., Borowik, G., Naidu, D.: Control system design based on modern embedded systems. 14th International Workshop on Computer Aided Systems Theory - EUROCAST 2013 pp. 346–347 (2013), las Palmas de Gran Canaria, Spain
11. Liu, B., Zydek, D., Selvaraj, H., Gewali, L.: Accelerating high performance computing applications: Using cpus, gpus, hybrid cpu/gpu, and fpgas. In: Parallel and Distributed Computing, Applications and Technologies (PDCAT), 2012 13th International Conference on. pp. 337–342. IEEE (2012)
12. Rigatos, G., Tzafestas, S.: Extended Kalman filtering for fuzzy modeling and multi-sensor fusion. Mathematical and Computer Modeling of Dynamical Systems: Methods, Tools and Applications in Engineering and Related Sciences 13(3), 251–266 (2007)
13. Shawky, A., Zydek, D., Elhalwagy, Y., Ordys, A.: Modeling and nonlinear control of aflexible-linkmanipulator. Applied Mathematical Modeling (2013)
14. Simon, D.: Optimal State Estimation: Kalman, H-infinity, and Nonlinear Approaches. John Wiley & Sons (2006)
15. Simon, D.: Using nonlinear Kalman filtering to estimation signals. Embedded Systems Design 19(7), 38–53 (Dec 2006)
16. Zydek, D., Chmaj, G., Chiu, S.: Modeling computational limitations in h-phy and overlay-noc architectures. The Journal of Supercomputing pp. 1–20 (2013)

Nonlinear Position Control of DC Motor Using Finite-Horizon State Dependent Riccati Equation

Ahmed Khamis, D. Subbaram Naidu, and Dawid Zydek

1 Introduction

The need to improve performance in control systems requires more and more accurate modeling [10]. However, if a model is a good representation of the real system over a wide range of operating points, it is most often nonlinear [2]. Traditional technique to control nonlinear systems is to linearize the nonlinear system in a small region around the operating point and then design linear controllers. These controllers with constant gains can be expected to perform satisfactorily in the neighborhood of the operating point. However, they may not be capable of dealing with a situation over large range of operating points. To overcome this drawback, one approach is to use extended linearization design [16]. This approach is to design several linear controllers matching to several operating points that may cover the whole dynamic region of the system. Then these linear controllers are pieced together to obtain a nonlinear controller, which is known as gain scheduling with respect to constant operating equilibrium points [17]. But the major limitation of gain scheduling is that stability properties of the closed-loop system can be guaranteed only in a vicinity of the equilibrium manifold and under an assumption of slow variation of signals [14].

In contrast, recent methods for nonlinear control consider the nonlinearities fully into account. The main advantage of such approaches is that for large regions of the state-space, the controller does not need to be gain scheduled according to the operating point. Furthermore, this control is achieved even for large state deviations. One of the highly promising and rapidly emerging techniques for nonlinear optimal controllers designing is the State Dependent Riccati Equation (SDRE) technique [6]. The SDRE technique can be used for regulation or tracking of infinite-horizon nonlinear systems [3, 18].

A primary advantage offered by the SDRE to the control designer is the opportunity to make tradeoffs between the control effort and the state errors, that tradeoffs can be accomplished by accurate tuning of the state dependent coefficients (SDC) matrices. Also, as the SDRE depends only on the current state, the SDRE computation can be carried out online, in which case the SDRE is defined along the state trajectory. The SDRE computation can be solved in many ways. One of the approaches which passes the processing and other possible data to the ground infrastructure of computational units is discussed in [5, 11, 19].

In this paper, we address the position control of a permanent magnet DC motor based on the nonlinear motor system dynamics. A novel technique for tracking of finite-horizon nonlinear systems is used. This technique is based on the change of variable [15], that converts the nonlinear differential Riccati equation (DRE) to a linear differential Lyapunov equation [13], which can be solved in real time at each time step [9].

The reminder of this paper is organized as follows: the nonlinear mathematical model of the DC motor is discussed in Section II. Section III presents the nonlinear finite-horizon tracking technique via SDRE. Simulation results are discussed in Section IV. Finally, conclusions of this paper are given in Section V.

2 DC Motor Nonlinear Mathematical Modeling

In this section, the modeling approach adopted to identify a nonlinear mathematical model for the motor is demonstrated. The DC motor used in this paper is carbon-brush permanent magnet 12v DC motor. The mathematical model is shown in Fig. 1, where R is the armature resistance, L is the armature inductance, v is the voltage applied to the motor, i is the current through the motor, e is the back emf

A. Khamis (✉) • D.S. Naidu • D. Zydek
Department of Electrical Engineering, Idaho State University, Pocatello, ID, USA
e-mail: khamahme@isu.edu; naiduds@isu.edu; zydedawi@isu.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_5,

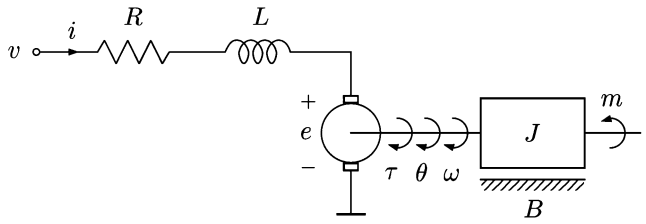


Fig. 1 Permanent magnet DC motor system model

voltage, J is the moment of inertia of the load, B is the viscous friction coefficient, τ is the torque generated by the motor, θ is the angular position of the motor,and ω is the angular velocity of the motor.

The dynamic equations for the DC motor are:

$$\omega(t) = \frac{d\theta(t)}{dt}, \tag{1}$$

$$v(t) = L\frac{di(t)}{dt} + Ri(t) + k_w\omega(t), \tag{2}$$

$$k_i i(t) = J\frac{d\omega(t)}{dt} + B\omega(t) + Csgn(\omega), \tag{3}$$

where k_w is the back emf constant, k_i is the torque constant of the motor, and C is the motor static friction, and the signum function $sgn(\omega)$ is defined as

$$sgn(\omega) = \begin{cases} -1 & \text{for} \quad \omega < 0, \\ 0 & \text{for} \quad \omega = 0, \\ 1 & \text{for} \quad \omega > 0, \end{cases} \tag{4}$$

or it can be written in this form:

$$sgn(\omega) = \frac{|\omega|}{\omega}. \tag{5}$$

The system nonlinear state equations can be written in the form:

$$\dot{x_1} = x_2, \tag{6}$$

$$\dot{x_2} = \left(\frac{-B}{J} - \frac{-C}{J|x_2|}\right)x_2 + \frac{k_i}{J}x_3, \tag{7}$$

$$\dot{x_3} = -\frac{k_w}{L}x_2 - \frac{R}{L}x_3 + \frac{1}{L}u, \tag{8}$$

$$y = x_1, \tag{9}$$

where: $\theta = x_1$, $\dot{\theta} = x_2$, $i = x_3$, $v = u$.

3 Nonlinear Tracking Using Finite-Horizon Differential SDRE

3.1 Problem Formulation

The nonlinear system considered in this paper is assumed to be in the form:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})\mathbf{u}(t), \tag{10}$$

$$\mathbf{y}(t) = \mathbf{h}(\mathbf{x}). \tag{11}$$

That nonlinear system can be expressed in a state-dependent like linear form, as:

$$\dot{\mathbf{x}}(t) = \mathbf{A}(\mathbf{x})\mathbf{x}(t) + \mathbf{B}(\mathbf{x})\mathbf{u}(t), \tag{12}$$

$$\mathbf{y}(t) = \mathbf{C}(\mathbf{x})\mathbf{x}(t), \tag{13}$$

where $\mathbf{f(x)} = \mathbf{A(x)x}(t)$, $\mathbf{B(x)} = \mathbf{g(x)}$, $\mathbf{h(x)} = \mathbf{C(x)x}(t)$.

Let $\mathbf{z}(t)$ be the desired output. The goal is to find a state feedback control law that minimizes a cost function given by [12]:

$$\mathbf{J}(\mathbf{x},\mathbf{u}) = \frac{1}{2}\mathbf{e}'(t_f)\mathbf{F}\mathbf{e}(t_f) + \frac{1}{2}\int_{t_0}^{t_f}[\mathbf{e}'(t)\mathbf{Q}(\mathbf{x})\mathbf{e}(t) + u'(\mathbf{x})\mathbf{R}(\mathbf{x})\mathbf{u}(\mathbf{x})]\,dt, \tag{14}$$

where $\mathbf{e}(t) = \mathbf{z}(t) - \mathbf{y}(t)$, $\mathbf{Q}(\mathbf{x})$ and $\mathbf{F}$ are symmetric positive semi-definite matrices, and $\mathbf{R}(\mathbf{x})$ is a symmetric positive definite matrix. Moreover, $\mathbf{x}'(t)\mathbf{Q}(\mathbf{x})\mathbf{x}(t)$ is a measure of control accuracy and $\mathbf{u}'(\mathbf{x})\mathbf{R}(\mathbf{x})\mathbf{u}(\mathbf{x})$ is a measure of control effort [4].

3.2 Solution for Finite-Horizon Differential SDRE Tracking

To minimize the above cost function (14), a feedback control law can be given as

$$\mathbf{u}(\mathbf{x}) = -\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}'(\mathbf{x})[\mathbf{P}(\mathbf{x})\mathbf{x}(t) - \mathbf{g}(\mathbf{x})], \tag{15}$$

where $\mathbf{P}(\mathbf{x})$ is a symmetric, positive-definite solution of the Differential State Dependent Riccati Equation, strictly speaking it could be called State Dependent Differential Riccati Equation (SDDRE), of the form

$$-\dot{\mathbf{P}}(\mathbf{x}) = \mathbf{P}(\mathbf{x})\mathbf{A}(\mathbf{x}) + A'(\mathbf{x})\mathbf{P}(\mathbf{x}) - \mathbf{P}(\mathbf{x})\mathbf{B}(\mathbf{x})\mathbf{R}^{-1}\mathbf{B}'(\mathbf{x})\mathbf{P}(\mathbf{x}) + \mathbf{C}'(\mathbf{x})\mathbf{Q}(\mathbf{x})\mathbf{C}(\mathbf{x}), \tag{16}$$

with the final condition

$$\mathbf{P}(\mathbf{x}, t_f) = \mathbf{C}'(t_f)\mathbf{F}\mathbf{C}(t_f). \tag{17}$$

The resulting SDRE-controlled trajectory becomes the solution of the state-dependent closed-loop dynamics

$$\dot{\mathbf{x}}(t) = [\mathbf{A}(\mathbf{x}) - \mathbf{B}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}'(\mathbf{x})\mathbf{P}(\mathbf{x})]\mathbf{x}(t) + \mathbf{B}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}'(\mathbf{x})\mathbf{g}(\mathbf{x}), \tag{18}$$

where $\mathbf{g(x)}$ is a solution of the state-dependent non-homogeneous vector differential equation

$$\dot{\mathbf{g}}(\mathbf{x}) = -[\mathbf{A}(\mathbf{x}) - \mathbf{B}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}'(\mathbf{x})\mathbf{P}(\mathbf{x})]'\mathbf{g}(\mathbf{x}) - \mathbf{C}'(\mathbf{x})\mathbf{Q}(\mathbf{x})\mathbf{z}(\mathbf{x}), \tag{19}$$

with the final condition

$$\mathbf{g}(\mathbf{x}, t_f) = \mathbf{C}'(t_f)\mathbf{F}\mathbf{z}(t_f). \tag{20}$$

As the SDRE function of ($\mathbf{x}$, t), we do not know the value of the states ahead of present time step. Consequently, the state dependent coefficients cannot be calculated to solve (16) with the final condition (17) by backward integration from t_f to t_0. To overcome this problem, an approximate analytical approach is used [8, 13, 15]. which converts the original nonlinear Ricatti equation to a linear differential Lyapunov equation. At each time step, the Lyapunov equation can be solved in closed form. In order to solve the DRE (16), one can follow the following steps:

– Solve Algebraic Riccati Equation (ARE) to calculate the steady state value $\mathbf{P}_{ss}\mathbf{(x)}$

$$\mathbf{P}_{ss}(\mathbf{x})\mathbf{A}(\mathbf{x}) + \mathbf{A}'(\mathbf{x})\mathbf{P}_{ss}(\mathbf{x}) - \mathbf{P}_{ss}(\mathbf{x})\mathbf{B}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}'(\mathbf{x})\mathbf{P}_{ss}(\mathbf{x}) + \mathbf{Q}(\mathbf{x}) = 0. \tag{21}$$

– Use change of variables technique and assume that

$$\mathbf{K}(\mathbf{x}, t) = [\mathbf{P}(\mathbf{x}, t) - \mathbf{P}_{ss}(\mathbf{x})]^{-1}. \tag{22}$$

– Calculate the value of $\mathbf{A}_{cl}\mathbf{(x)}$ as

$$\mathbf{A}_{cl}(\mathbf{x}) = \mathbf{A}(\mathbf{x}) - \mathbf{B}(\mathbf{x})\mathbf{R}^{-1}\mathbf{B}'(\mathbf{x})\mathbf{P}_{ss}(\mathbf{x}). \tag{23}$$

– Calculate the value of $\mathbf{D}$ by solving the algebraic Lyapunov equation [7]

$$\mathbf{A}_{cl}\mathbf{D} + \mathbf{D}\mathbf{A}_{cl}' - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}' = \mathbf{0}. \tag{24}$$

– Solve the differential Lyapunov equation

$$\mathbf{x}(\mathbf{x}, \mathbf{t}) = \mathbf{K}(\mathbf{x}, \mathbf{t})\mathbf{A}_{cl}'(\mathbf{x}) + \mathbf{A}_{cl}(\mathbf{x})\mathbf{K}(\mathbf{x}, \mathbf{t}) - \mathbf{B}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}'(\mathbf{x}). \tag{25}$$

The solution of (25), as shown by [1], is given by

$$\mathbf{K}(\mathbf{t}) = \mathbf{e}^{\mathbf{A}_{cl}(\mathbf{t}-\mathbf{t}_f)}(\mathbf{K}(\mathbf{x}, \mathbf{t}_f) - \mathbf{D})\mathbf{e}^{\mathbf{A}_{cl}'(\mathbf{t}-\mathbf{t}_f)} + \mathbf{D}. \tag{26}$$

– Calculate the value of $\mathbf{P}(\mathbf{x}, t)$ from the equation

$$\mathbf{P}(\mathbf{x}, \mathbf{t}) = \mathbf{K}^{-1}(\mathbf{x}, \mathbf{t}) + \mathbf{P}_{ss}(\mathbf{x}). \tag{27}$$

– Calculate the steady state value $\mathbf{g}_{ss}\mathbf{(x)}$ from the equation

$$\mathbf{g}_{ss}(\mathbf{x}) = [\mathbf{A}(\mathbf{x}) - \mathbf{B}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}'(\mathbf{x})\mathbf{P}_{ss}(\mathbf{x})]'^{-1}\mathbf{C}'(\mathbf{x})\mathbf{Q}(\mathbf{x})\mathbf{z}(\mathbf{x}). \tag{28}$$

– Use change of variables technique and assume that $\mathbf{K}_g(\mathbf{x}, t) = [\mathbf{g}(\mathbf{x}, t) - \mathbf{g}_{ss}(\mathbf{x})]$.

– Solve the differential equation

$$\mathbf{K}_g(t) = \mathbf{e}^{-(\mathbf{A}-\mathbf{B}\mathbf{R}^{-1}\mathbf{B}'\mathbf{P})'(\mathbf{t}-\mathbf{t}_f)}[\mathbf{g}(\mathbf{x}, t_f) - \mathbf{g}_{ss}(\mathbf{x})]. \tag{29}$$

– Calculate the value of $\mathbf{g}(\mathbf{x}, t)$ from the equation

$$\mathbf{g}(\mathbf{x}, t) = \mathbf{K}_g(\mathbf{x}, t) + \mathbf{g}_{ss}(\mathbf{x}). \tag{30}$$

– Calculate the value of the optimal control $\mathbf{u}(\mathbf{x}, t)$ as

$$\mathbf{u}(\mathbf{x}, t) = -\mathbf{R}^{-1}(\mathbf{x})\mathbf{B}'(\mathbf{x})[\mathbf{P}(\mathbf{x}, t)\mathbf{x}(t) - \mathbf{g}(\mathbf{x}, t)]. \tag{31}$$

Fig. 2 summarized the overview of the process of finite-horizon SDDRE tracking technique

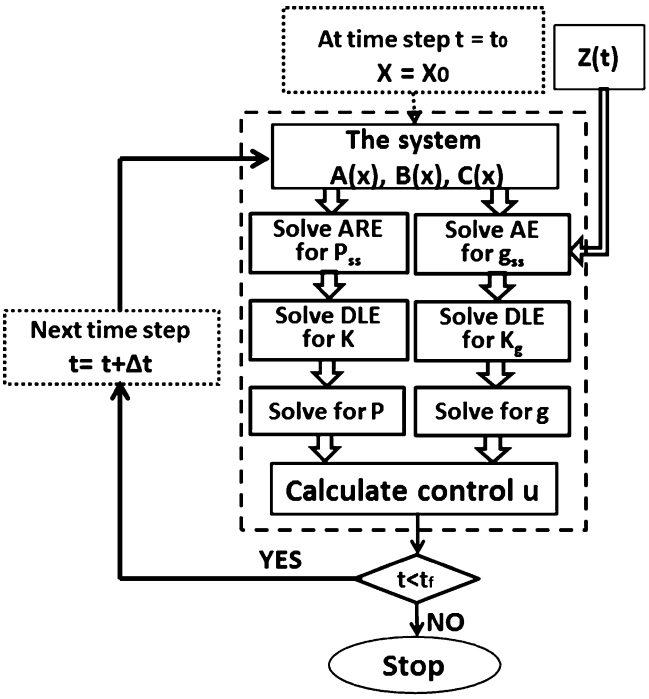


Fig. 2 Overview of The Process of Finite-Horizon Differential SDRE Tracking

Note: It is easily seen that this technique with finite-horizon differential SDRE can be used for linear systems and the resulting differential SDRE becomes the standard DRE [12].

4 Simulation Results

The system nonlinear state equations of the DC motor (6-8) can be rewritten in state dependent form:

$$\begin{bmatrix} \dot{x_1} \\ \dot{x_2} \\ \dot{x_3} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & \left(\frac{-B}{J} - \frac{-C}{J|x_2|}\right) & \frac{k_i}{J} \\ 0 & -\frac{k_w}{L} & -\frac{R}{L} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{L} \end{bmatrix} u. \quad (32)$$

where: $\theta = x_1$, $\dot{\theta} = x_2$, $i = x_3$, $v = u$.

Let the reference angle is

$$z(t) = 90^o, \quad (33)$$

and the selected weighted matrices are

$$Q = diag(60, 0, 0), R = 0.7, F = diag(1, 1, 1). \quad (34)$$

The simulations are performed for final time of 10 seconds and the resulting output trajectories is shown in Fig. 3, where the dash-dot line denotes the *reference* angle trajectory, and the solid line denotes the *actual* trajectory. The optimal control is shown in Fig. 4.

Comparing these trajectories in Fig. 3, it's clear that the propose algorithm gives very good results as the actual optimal angle is making a very good tracking to the reference angle with average error of 0.05%.

From these results, it can be seen that the developed algorithm is able to solve the DC motor nonlinear tracking problem.

5 Conclusions

The paper presented a new finite-horizon tracking technique for nonlinear systems. This technique based on change of variables that converts the nonlinear differential Riccati equation to a linear Lyapunov equation.The Lyapunov equation is solved in a closed form at the given time step. Simulation results for DC motor are included to demonstrate the effectiveness of the developed technique.

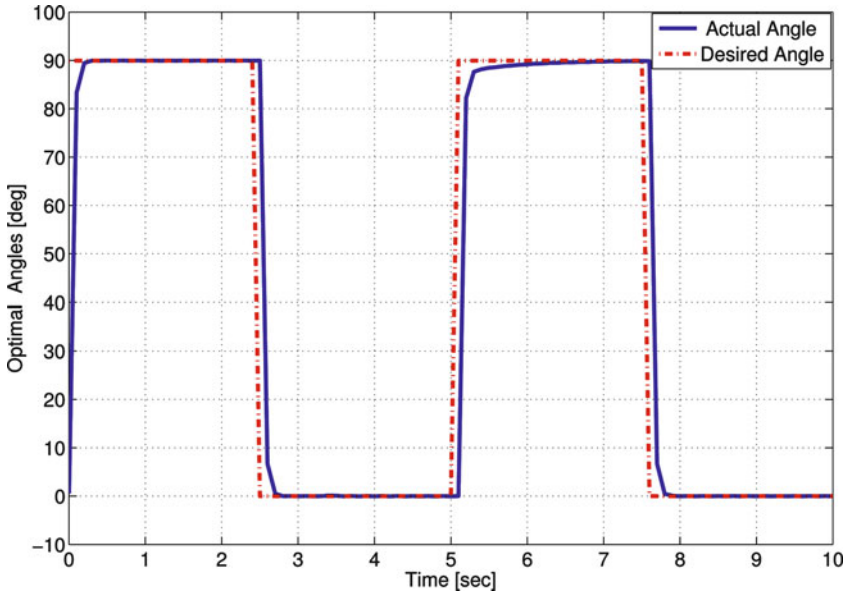


Fig. 3 Optimal angle tracking for the DC motor

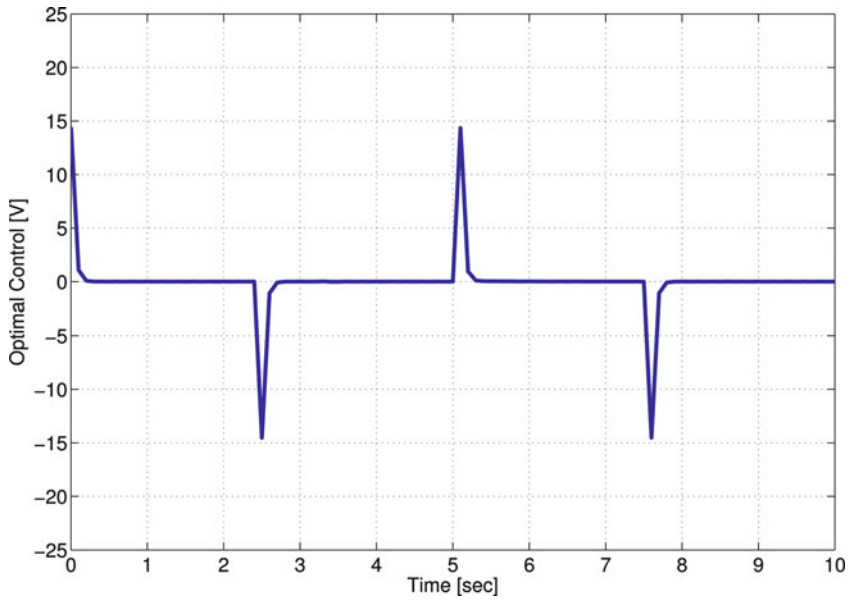


Fig. 4 Optimal control voltage for the DC motor

References

1. Barraud, A.: A new numerical solution of xdot=a1*x+x*a2+d, x (0)=c. IEEE Transaction on Automatic Control 22(6), 976–977 (Dec 1977)
2. Çimen, T.: Recent advances in nonlinear optimal feedback control design. Proceedings of the 9th WSEAS International Conference on Applied Mathematics, Istanbul, Turkey pp. 460–465 (May 2006)
3. Çimen, T.: Development and validation of a mathematical model for control of constrained nonlinear oil tanker motion. Mathematical and Computer Modeling of Dynamical Systems 15(1), 1749 (2009)
4. Çimen, T.: Survey of state-dependent Riccati equation in nonlinear optimal feedback control synthesis. AIAA Journal of Guidance, Control, and Dynamics 35(4), 1025–1047 (2012)
5. Chmaj, G., Walkowiak, K.: A p2p computing system for overlay networks. Future Generation Computer Systems 29(1), 242–249 (2013)
6. Cloutier, J.: State-dependent Riccati equation techniques: An overview. Proc. American Control Conference 2, 932–936 (1997)
7. Gajic, Z., Qureshi, M.: The Lyapunov matrix equation in system stability and control. New York: Dover Publications (2008)
8. Heydari, A., Balakrishnan, S.: Path planning using a novel finite-horizon suboptimal controller. Journal of Guidance, Control, and Dynamics pp. 1–5 (2013)
9. Khamis, A., Naidu, D.: Nonlinear optimal tracking using finite-horizon state dependent Riccati equation (SDRE). Proceedings of the 4th International Conference on Circuits, Systems, Control, Signals (WSEAS) pp. 37–42 (August 2013), valencia, Spain
10. Khamis, A., Zydek, D., Borowik, G., Naidu, D.: Control system design based on modern embedded systems. 14th International Workshop on Computer Aided Systems Theory - EUROCAST 2013 pp. 346–347 (2013), las Palmas de Gran Canaria, Spain
11. Liu, B., Zydek, D., Selvaraj, H., Gewali, L.: Accelerating high performance computing applications: Using cpus, gpus, hybrid cpu/gpu, and fpgas. In: Parallel and Distributed Computing, Applications and Technologies (PDCAT), 2012 13th International Conference on. pp. 337–342. IEEE (2012)
12. Naidu, D.: Optimal Control Systems. CRC Press (2003)
13. Nazarzadeh, J., Razzaghi, M., Nikravesh, K.: Solution of the matrix Riccati equation for the linear quadratic control problems. Mathematical and Computer Modelling 27(7), 51–55 (1998)
14. Neamtu, A., Stoica, A.: A bumpless transfer method for automatic flight control switching. UPB Scientific Bulletin, Series D 74 (2012)
15. Nguyen, T., Gajic, Z.: Solving the matrix differential Riccati equation: a Lyapunov equation approach. IEEE Trans. Automatic Control 55(1), 191–194 (2010)
16. Ramirez, H.: Design of pi controllers for DC-to-DC power supplies via extended linearization. International Journal of Control 51(3), 601–620 (1990)
17. Rugh, W., Shamma, J.: Research on gain scheduling. Automatica 36(10), 1401–1425 (2000)
18. Shawky, A., Zydek, D., Elhalwagy, Y., Ordys, A.: Modeling and nonlinear control of aflexible-linkmanipulator. Applied Mathematical Modeling (2013)
19. Zydek, D., Chmaj, G., Chiu, S.: Modeling computational limitations in h-phy and overlay-noc architectures. The Journal of Supercomputing pp. 1–20 (2013)

Generalization of the Observer Principle for YOULA-Parametrized Regulators

László Keviczky and Csilla Bányász

1 Introduction, the State Feedback (*SF*)

It is a well known methodology to use the state variable representations (*SVR*) of linear time invariant (*LTI*) single input - single output (*SISO*) systems [1]. The *SVR* proved to be excellent tool to implement both *LQR* (Linear system - Quadratic criterion - Regulator) control and pole placement design. The practical applicability required to introduce the observers, which make this methodology widely applied even for large scale and higher dimension plants [3]. Thousands of theoretical considerations mostly concentrate on the irregularities and special structures in the *SVR* appearing and much less publications deal with the model error properties of these systems.

It is possible to find a proper new way to discuss and investigate the the special properties and limitations of the classical state-feedback (*SF*), state-feedback/observer (*SFO*) topologies if someone replaces the *SVR* by their transfer function representations (*TFR*) [2].

Consider a *SISO* continuous time (t) *LTI* dynamic plant described by the *SVR*

$$P = \frac{B}{A} \tag{1}$$

Here P is the *TFR* of the open-loop system with the numerator and denominator polynomials

$$B(s) = s^n + b_1 s^{n-1} + \cdots + b_{n-1} s + b_n \tag{2}$$

$$A(s) = s^n + a_1 s^{n-1} + \cdots + a_{n-1} s + a_n \tag{3}$$

If we want to express the operation of the *SF* by equivalent scheme using *TFR* forms, Fig. 1 can be used, where the feedback regulator $R_f = K_k$is obtained from the basic equation (complementary sensitivity function, *CSF*) of the closed-loop

$$T_{ry}(s) = \frac{k_r\, B(s)}{R(s)} = \frac{k_r\, B(s)}{A(s) + K(s)} = \frac{k_r\, P}{1 + K_k\, P} \tag{4}$$

where k_r is obtained by requiring that the static gain of T_{ry} should be equal to one. The calibrating factor k_r is necessary because the closed-loop using *SF* is not an integrating one. Equation (4) clearly shows, that the open-loop zeros remain unchanged and the closed-loop poles will be the required ones. The solution formally makes the characteristic polynomial of the closed-loop equal to the desired polynomial ("placed poles")

$$R(s) = s^n + r_1 s^{n-1} + \cdots + r_{n-1} s + r_n \tag{5}$$

Here it is obtained that

$$R_f = K_k(s) = \frac{K(s)}{B(s)} = \frac{R(s) - A(s)}{B(s)} \tag{6}$$

which corresponds to the state feedback vector in the classical *SVR*.

2 Observer-Based State-Feedback with Equivalent *TFR* Forms

The practical applicability of the *SF* theory was introduced by the development of the observers capable to calculate the unmeasured state variables. The most general *SF*/Observer (*SFO*) topology discussed above can also be given using equivalent *TFR* forms of *SF* and is shown in Fig. 2.

L. Keviczky (✉) • C. Bányász
Hungarian Academy of Sciences, Computer and Automation Research Institute and MTA-BME Control Engineering Research Group, H-1111 Budapest, Kende u 13-17, Hungary
e-mail: keviczky@sztaki.hu; banyasz@sztaki.hu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_6,

The usual classical design goal for the observer is to determine the observer feedback so that its feedback closed-loop system has the characteristic polynomial

$$Q(s) = s^n + q_1 s^{n-1} + \cdots + q_{n-1}s + q_n \tag{7}$$

The *TFR* $K_l(s) = L(s)/B(s)$ in Fig. 2 corresponds to the observer feedback vector in the classical *SVR*.

The pole-placement design goals for the *SF* and observer dynamics require

$$k(s) = R(s) - A(s) \quad and \quad L(s) = Q(s) - A(s) \tag{8}$$

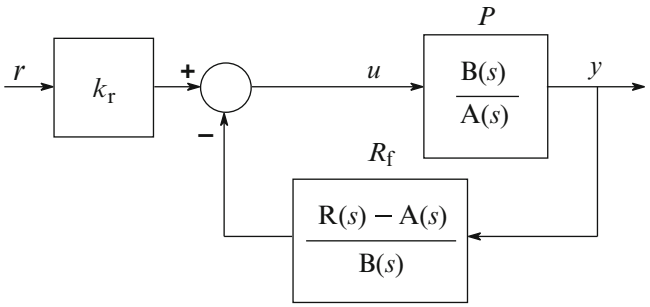


Fig. 1 Equivalent schemes of *SF* using *TFR* forms

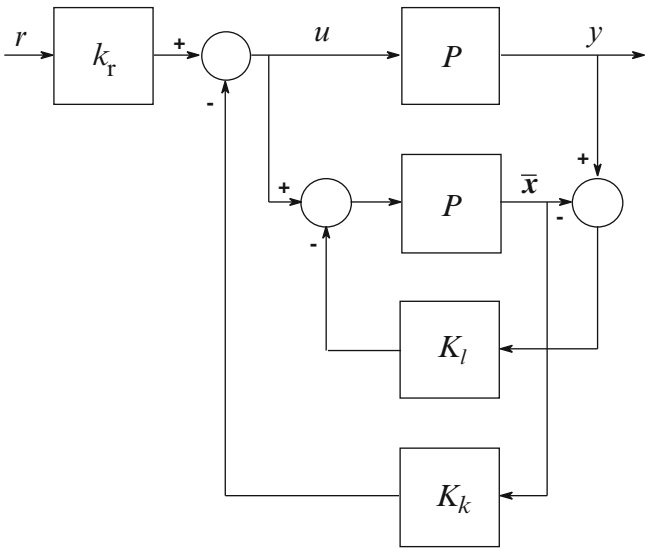


Fig. 2 Equivalent topology of the general basic *SFO* scheme using *TFR* forms

After some long, but straightforward block manipulations the equivalent *SFO* scheme can be transformed into another unity feedback closed-loop form given in Fig. 3.

It is interesting to observe that the transfer function of the closed-loop in Fig. 3 has a very special structure

$$\frac{P^2K_kK_l}{1 + P(K_k + K_l) + P^2K_kK_l} = \frac{PK_k}{1 + PK_k}\frac{PK_l}{1 + PK_l} = \frac{K}{R}\frac{L}{Q} \tag{9}$$

It is formally two simpler closed-loops cascaded, which dynamically completely corresponds to the characteristic equation: $R(s) = 0$ and $Q(s) = 0$. The overall transfer function of the *SFO* system is

$$T_{ry}(s) = k_r \frac{1 + PK_l}{PK_kK_l}\frac{PK_k}{1 + PK_k}\frac{PK_l}{1 + PK_l} = \frac{k_rP}{1 + PK_k} = \frac{k_rB}{R} \tag{10}$$

3 Model Error Properties

The above widely applied methodology has a common problem, that in all regulator and observer equations the true process P is used instead of the estimated model $\widehat{P}$ of the process. The equivalent *TFR* form of the *SF* using the model of the process is shown in Fig. 4.

The parallel scheme in Fig. 4 is used to compute the model error. Using (4) the $\widehat{T}_{ry}$ model-based version of T_{ry} is

$$\widehat{T}_{ry} = \frac{k_r\,P}{1 + K_k\,\widehat{P}} = \frac{k_rB}{R}\frac{\widehat{A}}{A} = T_{ry}\frac{\widehat{A}}{A}\widehat{A} \tag{11}$$

and its relative uncertainty

$$\ell_T = \frac{\widehat{T}_{ry} - T_{ry}}{\widehat{T}_{ry}} = \frac{\widehat{A} - A}{A} = \ell_A \tag{12}$$

which shows that $\ell_T = 0$ for $\ell_A = 0$. Introducing the additive $\Delta = P - \widehat{P}$ and relative plant model error

$$\ell = \frac{\Delta}{\widehat{P}} = \frac{P - \widehat{P}}{\widehat{P}} \tag{13}$$

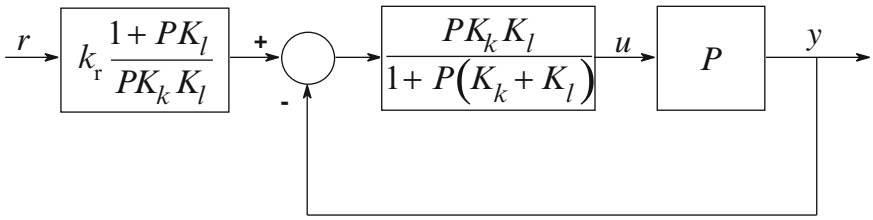


Fig. 3 Reduced equivalent topology of the general basic *SFO* scheme

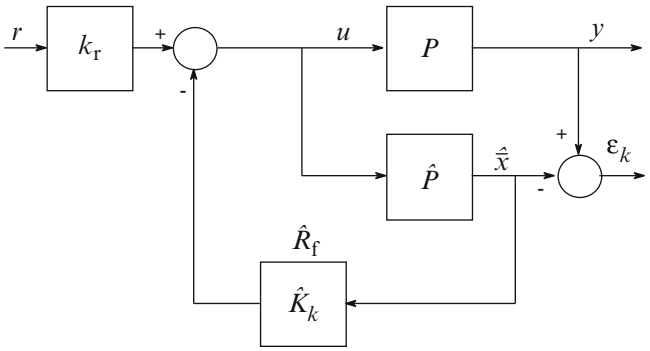


Fig. 4 The model based *SF* scheme and error

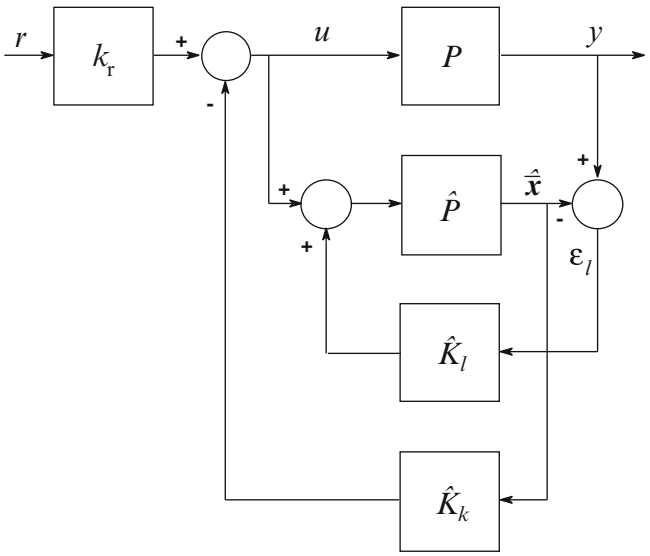


Fig. 5 Model based *SFO* scheme with *TFR* forms

the modeling error ε_k in Fig. 4 can be expressed as

$$\varepsilon_k = \frac{k_r \widehat{B}}{B}\,\ell\, r = T_{ry}\frac{\widehat{B}}{B}\,\ell\, r = \widehat{P}\,\ell\, u \tag{14}$$

The *SFO* scheme is widely applied in the practice with model-based *SVR*, so it is interesting how the model-based scheme in Fig. 5 influences the original modeling error ε_k.

After some long but straightforward computations

$$\varepsilon_l = \frac{\widehat{P}}{1 + K_l\widehat{P}}\,\ell\, u = \frac{\widehat{B}}{Q}\ell\, u = \frac{1}{1 + K_l\widehat{P}}\varepsilon_k \tag{15}$$

is obtained. Equation (15) clearly shows the influence of the *SFO* scheme, because it decreases the modeling error ε_k by $(1 + K_l\widehat{P})$. Selecting fast observer poles, one can reach quite small "virtual" modeling error ε_l in the major frequency domains of the tracking task.

Besides the radical model error attenuating behavior of the model-based *SFO* scheme, unfortunately it has a very important drawback, the nice cascade (9) structure changes to

$$\left.\frac{\widehat{P}^2K_kK_l(1+\ell)}{1+\widehat{P}(K_k+K_l)+\widehat{P}^2K_kK_l(1+\ell)}\right|_{\ell\to 0} = \frac{PK_k}{1+PK_k}\,\frac{PK_l}{1+PK_l} = \frac{K}{R}\,\frac{L}{Q} \tag{16}$$

which form is not factorable except for the exact model matching case, when $\ell \to 0$. On the basis of Fig. 5 and (16) it is easy to see that the poles of the observer feedback loop remain unchanged using the placement design equation forms model-based *SFO* (8), thus the only solution is to use the available model of the process, in this case $\widehat{A}$, i.e.,

$$K(s) = R(s) - \widehat{A}(s) \quad \text{and} \quad L(s) = Q(s) - \widehat{A}(s) \tag{17}$$

for the pole placing equations.

Because this design ensures the required poles only for small ℓ (see (16)), a serious robust stability investigation is required first. Next it is important to investigate where the actual pole is located for non zero ℓ, so how big the performance loss is coming from the model based *SFR*. These steps are usually neglected in most of the published papers, books and applications.

4 Introducing the Observer Based *YOULA*-Regulator

For open-loop stable processes the all realizable stabilizing (*ARS*) model based regulator $\widehat{C}$ is the *YOULA-parametrized* one:

$$\widehat{C}\left(\widehat{P}\right) = \left.\frac{Q}{1 - Q\widehat{P}}\right|_{\widehat{P}\to P} = \frac{Q}{1 - QP} = C(P) \tag{18}$$

where the "parameter" Q ranges over all proper ($Q(\omega = \infty)$ is finite), stable transfer functions [5], [6], see Fig. 6a.

It is important to know that the *Y-parametrized* closed-loop with the *ARS* regulator is equivalent to the well-known form of the so-called *Internal Model Control* (*IMC*) principle [6] based structure shown in Fig. 6b.

Q is anyway the transfer function from r to u and the *CSF* of the whole closed-loop for $\widehat{P} = P$, when $\ell \to 0$

$$\widehat{T}_{ry} = \frac{\widehat{C}P}{1 + \widehat{C}P} = QP\left.\frac{1+\ell}{1+(1-QP)\ell}\right|_{\ell\to 0} = QP = T_{ry} \tag{19}$$

is linear (and hence convex) in Q.

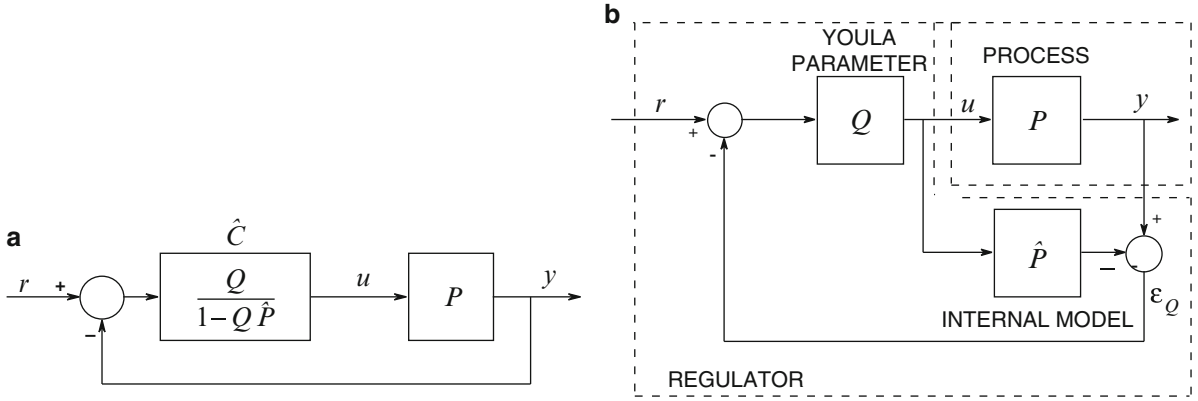


Fig. 6 The equivalent *IMC* structure of an *ARS* regulator

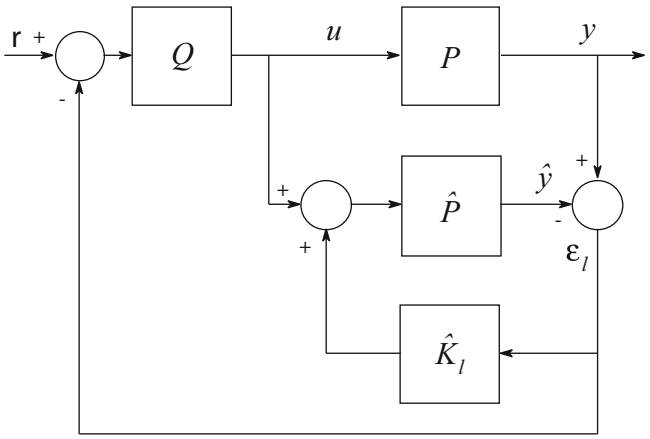


Fig. 7 The observer-based *IMC* structure

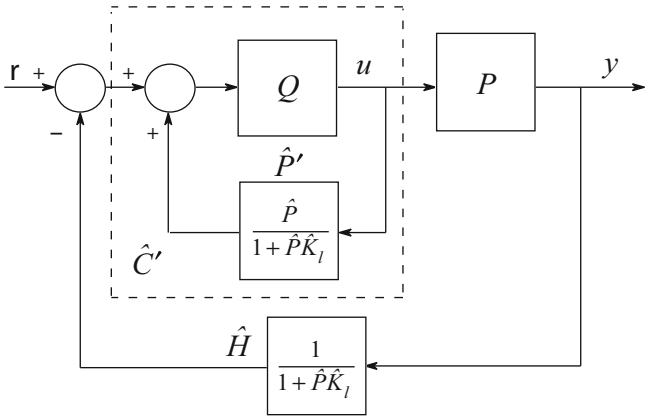


Fig. 8 Equivalent closed-loop for the observer-based *IMC* structure

It is interesting to compute the relative error ℓ_T of $\widehat{T}_{ry}$

$$\ell_T = \frac{T_{ry} - \widehat{T}_{ry}}{\widehat{T}_{ry}} = \frac{T_{ry}}{\widehat{T}_{ry}} - 1 = Q\left(P - \widehat{P}\right) = QP\frac{\ell}{1+\ell}$$
$$= T_{ry}\frac{\ell}{1+\ell} \tag{20}$$

The equivalent *IMC* structure performs the feedback from the model error ε_Q. Similarly to the *SFO* scheme it is possible to construct an internal closed-loop, which virtually reduces the model error to

$$\varepsilon_l = \frac{1}{1+\widehat{K}_l\widehat{P}}\left(y - \widehat{P}u\right) = \frac{1}{1+\widehat{K}_l\widehat{P}}\varepsilon_Q = \frac{1}{1+\widehat{L}_l}\varepsilon_Q$$
$$= \widehat{H}\ \varepsilon_Q \quad ; \quad \widehat{L}_l = \widehat{K}_l\widehat{P} \tag{21}$$

and performs the feedback from ε_l (see Fig. 7), where $\widehat{L}_l$ is the internal loop transfer function. In this case the resulting closed-loop will change to the scheme shown in Fig. 8.

This means that the introduction of the observer feedback changes the *Youla-parametrized* regulator to

$$\widehat{C}'\left(\widehat{P}'\right) = \frac{Q}{1 - Q\widehat{P}/\left(1+\widehat{K}_l\widehat{P}\right)} = \frac{Q\left(1+\widehat{K}_l\widehat{P}\right)}{1+\widehat{K}_l\widehat{P} - Q\widehat{P}} \tag{22}$$

The form of $\widehat{C}'$ shows that the regulator virtually controls a fictitious plant $\widehat{P}'$ which is also demonstrated in Fig. 8. Here the fictitious plant is

$$\widehat{P}' = \frac{\widehat{P}}{1+\widehat{K}_l\widehat{P}} = \frac{\widehat{P}}{1+\widehat{L}_l} \tag{23}$$

The closed-loop transfer function is now

$$\widehat{T}'_{ry} = \frac{\widehat{C}'P}{1+\widehat{C}'P} = \frac{QP\left(1+\widehat{K}_l\widehat{P}\right)}{1+\widehat{K}_l\widehat{P} - Q\widehat{P} + QP}$$
$$= QP\frac{1}{1+QP\frac{1}{1+\widehat{K}_l\widehat{P}}\frac{\ell}{1+\ell}}\Bigg|_{\ell\to 0} = QP = T_{ry} \tag{24}$$

The relative error ℓ'_{T} of $\widehat{T}'_{ry}$ becomes

$$\ell'_T = \frac{T_{ry} - \widehat{T}'_{ry}}{\widehat{T}'_{ry}} = \frac{T_{ry}}{\widehat{T}'_{ry}} - 1 = QP\frac{\ell}{1+\ell}\frac{1}{\left(1+\widehat{K}_l\widehat{P}\right)} = \ell_T\frac{1}{1+\widehat{L}_l} \tag{25}$$

which is smaller than ℓ_{T}. The reduction is by $\widehat{H} = 1/\left(1+\widehat{L}_l\right)$.

5 An Observer Based *PID*-Regulator

The ideal form of a YOULA-regulator based on reference model design [4], [5] is

$$C_{id} = \frac{(R_nP^{-1})}{1-(R_nP^{-1})P} = \frac{Q}{1-QP} = \frac{R_n}{1-R_n}P^{-1} \tag{26}$$

when the inverse of the process is realizable and stable. Here the operation of R_{n} can be considered a reference model (desired system dynamics). It is generally required that the reference model has to be strictly proper with unit static gain, i.e., $R_{\mathrm{n}}(\omega = 0) = 1$.

For a simple, but robust *PID* regulator design method assume that the process can be well approximated by its two major time constants, i.e.,

$$P \cong \frac{A}{A_2} \quad \text{where} \quad A_2 = (1+sT_1)(1+sT_2) \tag{27}$$

According to (26) the ideal YOULA-regulator is

$$C_{id} = \frac{R_nP^{-1}}{1-R_n} = \frac{R_n(1+sT_1)(1+sT_2)}{A\,(1-R_n)} \quad ; \quad T_1 > T_2 \tag{28}$$

Let the reference model R_{n} be of first order

$$R_n = \frac{1}{1+sT_n}$$

which means that the first term of the regulator is an integrator

$$\frac{R_n}{1-R_n} = \frac{1/(1+sT_n)}{1-1/(1+sT_n)} = \frac{1}{1+sT_n-1} = \frac{1}{sT_n} \tag{29}$$

whose integrating time is equal to the time constant of the reference model. Thus the resulting regulator corresponds to the design principle, i.e., it is an ideal *PID* regulator

$$C_{PID} = A_{PID}\frac{(1+sT_I)(1+sT_D)}{sT_I} = A_{PID}\frac{(1+sT_1)(1+sT_2)}{sT_1} \tag{30}$$

with

$$A_{PID} = T_1/AT_n \quad ; \quad T_I = T_1 \quad ; \quad T_D = T_2 \tag{32}$$

The YOULA-parameter Q in the ideal regulator is

$$Q = R_nP^{-1} = \frac{1}{A}\frac{(1+sT_1)(1+sT_2)}{1+sT_n} \tag{33}$$

It is not necessary, but desirable to ensure the realizability, i.e., to use

$$Q = R_nP^{-1} = \frac{1}{A}\frac{(1+sT_1)(1+sT_2)}{(1+sT_n)(1+sT)} \tag{34}$$

where T can be considered the time constant of the derivative action ($0.1\ T_{\mathrm{D}} \le T \le 0.5\ T_{\mathrm{D}}$). The regulator $\widehat{C}'$ and the feedback term $\widehat{H}$ must be always realizable. In the practice the *PID* regulator and the YOULA-parameter is always model-based, so

$$\widehat{C}_{PID}\left(\widehat{P}\right) = \widehat{A}_{PID}\frac{\left(1+s\widehat{T}_1\right)\left(1+s\widehat{T}_2\right)}{s\widehat{T}_1} \quad ; \quad \widehat{A}_{PID} = \frac{\widehat{T}_1}{\widehat{A}\ T_n} \tag{35}$$

$$\widehat{Q} = R_n\widehat{P}^{-1} = \frac{1}{\widehat{A}}\frac{\left(1+s\widehat{T}_1\right)\left(1+s\widehat{T}_2\right)}{1+sT_n} \tag{36}$$

The sheme of the observer based *PID* regulator is shown in Fig. 9, where a simple *PI* regulator

$$\widehat{K}_l = A_l\frac{1+sT_l}{sT_l} \tag{37}$$

is applied in the observer-loop. Here T_l must be in the range of T, i.e., considerably smaller than T_1 and T_2.

Note that the frequency characteristic of $\widehat{H}$ cannot be easily designed to reach a proper error suppression. For example, it is almost impossible to design a good realizable high cut filter in this architecture. The high frequency domain is always more interesting to speed up a control loop, so the target of the future research is how to select $\widehat{K}_l$ for the desired shape of $\widehat{H}$.

6 Simulation Examples

The simulation experiments were performed in using the observer based *PID* scheme shown in Fig. 9.

Example 1. The process parameters are: $T_1 = 20, T_2 = 10$ and $A = 1$. The model parameters are: $\widehat{T}_1 = 25$, $\widehat{T}_2 = 12$

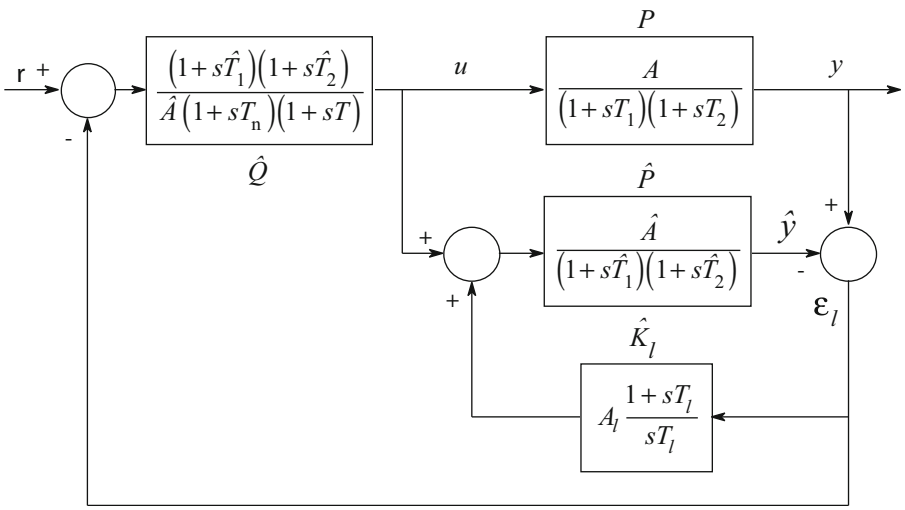


Fig. 9 An observer based *PID* regulator

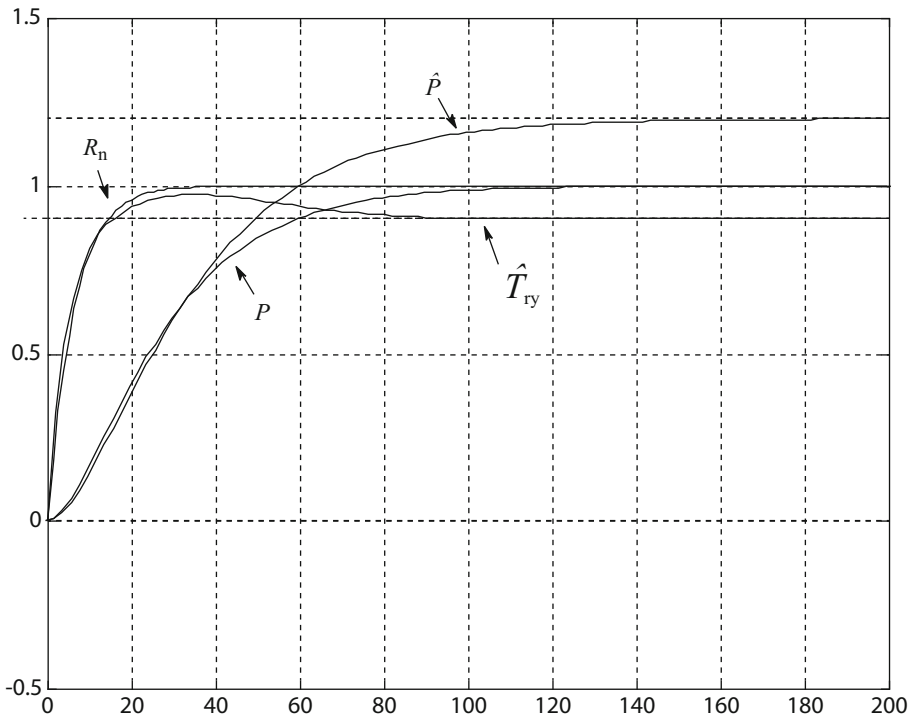


Fig. 10 Step responses using the observer based *PID* regulator

and $\widehat{A} = 1.2$. The purpose of the regulation is to speed up the basic step response by 4, i.e., $T_n = 5$ is selected in the first order R_n. In the observer loop a simple proportional regulator $\widehat{K}_l = 0.01$ is applied. The ideal form of Q (33) was used. Figure 10 shows some step responses in the operation of the observer based *PID* regulator.

It is easy to see that the $\widehat{T}'_{ry}$ very well approximates R_n in the high frequencies (for small time values) in spite of the very bad model $\widehat{P}$.

Example 2. The process parameters and the selected first order R_n are the same as in the previous example. The model parameters are: $\widehat{T}_1 = 30$, $\widehat{T}_2 = 20$ and $\widehat{A} = 0.5$. In the observer loop a *PI* regulator (37) is applied with $A_l = 0.001$ and $T_l = 2$. The ideal form of Q (33) was used. Figure 11 shows some step responses in the operation of the observer based *PID* regulator.

It is easy to see that the $\widehat{T}'_{ry}$ well approximates R_n in the high frequencies (for small time values) in spite of the very bad model.

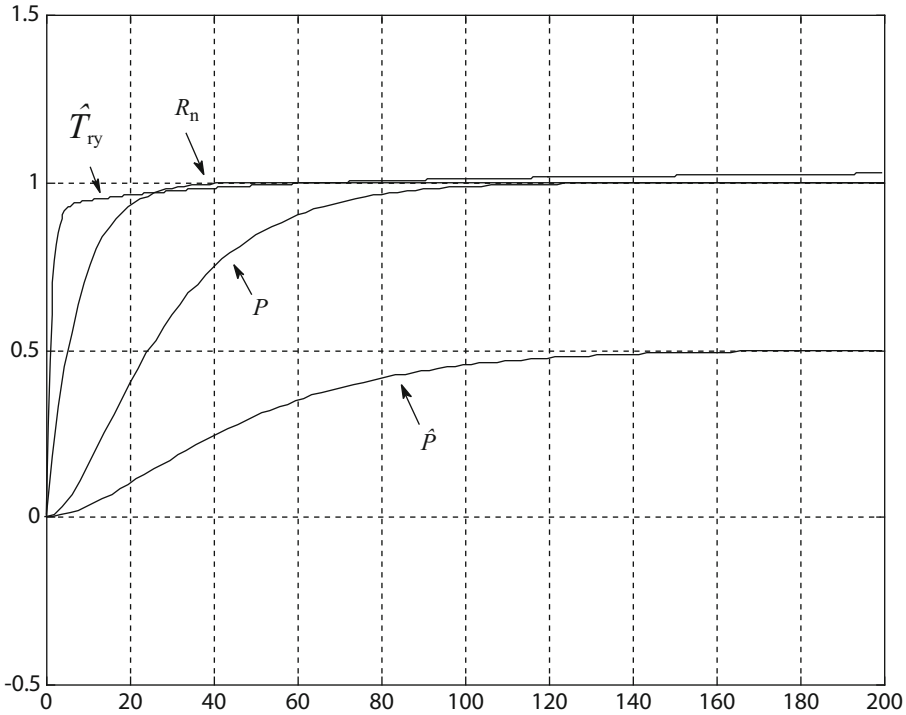


Fig. 11 Step responses using the observer based *PID* regulator

7 Conclusions

The *TFR* of the classical methods are introduced to get a simple and useful tool to analyze and explain further behaviors, which are difficult to obtain using *SVR*. Using *TFR* it was shown, if the *SVR* used in the *SFO* scheme is model-based then the original (without observer) model error decreases by the sensitivity function of the observer feedback loop. This model error reducing capability gives the theoretical background of the success of practical model-based *SFO* applications.

Finally the *SFO* method was applied for the classical *IMC* structure, opening a new class of methods for open-loop stable processes. This new method combines the classical YOULA-parametrization based regulators with the *SFO* scheme. Using this new approach an observer based *PID* regulator was also introduced. This regulator works well even in case of large model errors as some simulations showed.

References

1. Åström K.J. (2002). *Control System Design Lecture Notes*, U of California, Santa Barbara.
2. Bányász, Cs. and L. Keviczky (2004). State-feedback solutions via transfer function representations, J. Systems Science, 30, 2, pp. 21-34.
3. Kailath T. (1980). *Linear Systems*, Prentice Hall.
4. Keviczky L. (1995). Combined identification and control: another way. (Invited plenary paper.) *5th IFAC Symp. on Adaptive Control and Signal Processing*, *ACASP'95*, Budapest, H, pp. 13-30.
5. Keviczky L. and Cs. Bányász (2001). Iterative identification and control design using K-B parametrization, In: Control of Complex Systems, Eds: K.J. Åström, P. Albertos, M. Blanke, A. Isidori, W. Schaufelberger and R. Sanz, Springer, pp. 101-121.
6. Maciejowski J.M. (1989). *Multivariable Feedback Design*, Addison Wesley.

This work was supported in part by the MTA-BME Control Engineering Research Group of the HAS, at the Budapest University of Technology and Economics and by the project TAMOP 4.2.2.A-11/1/KONV-2012-2012, at the Széchenyi University of Győr.

The Compensation of N-th Order Bilinearity Applied with Model Based Controller

Lukasz Gadek, Leszek Koszalka, and Keith J. Burnham

1 Introduction

Bilinear structure, [1] and [2], is a transient between linear and non-linear models. As it can be found in the literature, bilinear models are considered as a class for which linear tools are applicable. It is more convenient to handle bilinear structure than extensive non-linear models. The approach to linearise the bilinear plant by a compensator has been used from 90s as in [3] or [4]. However, the existing compensator is adjusted for a first order bilinear term exclusively. Establishment of a compensator for n-th order is a novel introduced in this paper.

The paper is structured as following. Section 2 contains an overview on the bilinear structure in comparison with a standard linear ARX model. Section 3 presents a mathematical derivation of the n-th order compensator. In Section 4 a model based General Predictive Controller is introduced in two variations: linear and bilinear. Linear GPC without compensation and bilinear GPC are used as reference controllers in an experiment. Section 5 consists of a description of experiments in which performance of the compensator is analysed. First test are done to establish a robust linearisation (Section 5.1) then result of GPC control with and without the compensator is presented in Section 5.2. Conclusion on the compensator is in Section 6 while future work is underlined in Section 7.

2 Bilinear Model vs ARX

The bilinear model, in brief, is an extension of ARX with a bilinear part [5]. The extension itself is based on a simple non-linearity defined as a product of gain and varying systems states as in (1) where u and y are respectively input and output to the plant, z is time shift operator, A and B are vectors of coefficients.

$$y_k = -Ay_k(z^{-1}) + Bu_k(z^{-1}) + y_k(z^{-1})\eta u_k(z^{-1}) + e_k \quad (1)$$

Matrix η is bilinear part's gain. In this paper it is used exclusively in a diagonal form [6].

The form of equivalent transfer function in (2) is derived from (1) and is used to establish the compensator in Section 3.

$$\frac{Y}{U} = \frac{z^{-1}\left[\sum_{j=0}^{n_b-1} b_j z^{-j} + \sum_{q=1}^{n_\eta} \eta_q (z^{-q}Y) z^{-q+1}\right]}{1 + \sum_{i=1}^{n_a} a_i z^{-i}} \quad (2)$$

Bilinear structure is characterised by varying steady-state (SS) gain and dynamic properties with respect to the operating point, i.e. bilinear model is a special case of State Dependant Parameter model as in [7]. Such improvement over the ARX structure allows more flexibility in modelling a non-linear plant and hence higher convergence between estimated and actual output - an essential point of the model based control.

3 N-th Order Bilinear Compensation

Accordingly with Fig. 1 the compensation is aimed to linearise the identified bilinearity, i.e. neglect the bilinear term. To avoid ambiguity of signals following notation is used: u^c

L. Gadek (✉) • L. Koszalka
Wroclaw University of Technology, Wroclaw, Poland
e-mail: lukasz.gadek@pwr.edu.pl; leszek.koszalka@pwr.edu.pl

K.J. Burnham
Control Theory and Application Centre, Coventry, UK
e-mail: k.burnham@coventry.ac.uk

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_7,

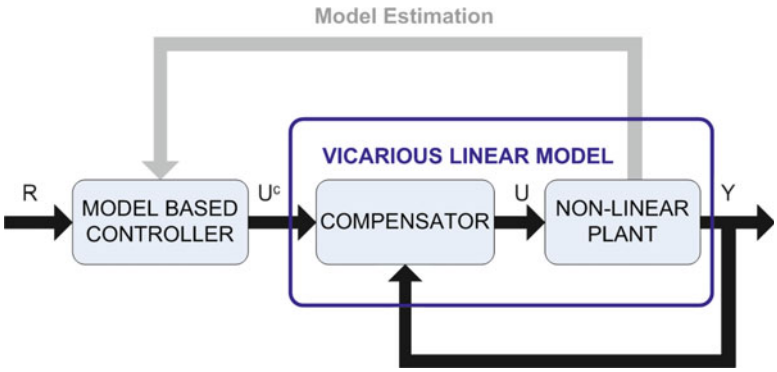


Fig. 1 Compensation and control scheme utilized in the paper

is output of the controller and input to compensator, u is input to a plant or output of the compensator.

If the plant could be approximately described with a transfer function $M(z)$ from (2) and the compensator is denoted as $C(z)$ then the aim is to obtain (3).

$$C(z)M(z) \quad = M^c(z) = \frac{z^{-1}\left[\sum_{j=0}^{n_b-1} b_j z^{-j}\right]}{1+\sum_{i=1}^{n_a} a_i z^{-i}} \tag{3}$$

where: $M(z) = \frac{Y}{U}$ (M is equivalent of TF), $C(z) = \frac{U}{U^c}$ and $M^c(z) = \frac{Y}{U^c}$.

When (2) is substituted into (3) then an explicit equation of the compensator is obtained and presented in (4).

$$\begin{aligned} C(z) &= \frac{\left[1+\sum_{i=1}^{n_a} a_i z^{-i}\right]\left[\sum_{j=0}^{n_b-1} b_j z^{-j}\right]}{\left[\sum_{j=0}^{n_b-1} b_j z^{-j} + \sum_{q=1}^{n_\eta} \eta_q (z^{-q}Y) z^{-q+1}\right]\left[1+\sum_{i=1}^{n_a} a_i z^{-i}\right]} \\ &= \frac{\sum_{j=0}^{n_b-1} b_j z^{-j}}{\sum_{j=0}^{n_b-1} b_j z^{-j} + \sum_{q=1}^{n_\eta} \eta_q (z^{-q}Y) z^{-q+1}} \end{aligned} \tag{4}$$

Now, the equivalent of compensator's transfer function in (4) is rewritten as a difference equation. Hence, a recursive and implementable form in (5) is obtained.

$$\begin{aligned} U\left[\sum_{j=0}^{n_b-1} b_j z^{-j} + \sum_{q=1}^{n_\eta} \eta_q (z^{-q}Y) z^{-q+1}\right] &= \left[\sum_{j=0}^{n_b-1} b_j z^{-j}\right] U^c \\ u_k &= \frac{-\sum_{j=2}^{n_\mu}\left[(b_{j-1} + \mu_j y_{k-j}) u_{k-j}\right] + \sum_{i=0}^{n_b} b_i u^c_{k-i}}{b_0 + \mu_1 y_{k-1}} \end{aligned} \tag{5}$$

Time dependencies of the compensation must be emphasized - calculation of the output is utilizing input for current time stamp. Although, zero delay system is not feasible in practical scenarios, most frequent it is delayed by one sample [9]. The compensator may be assumed a part of the plant (or controller) to avoid the delay, transfer functions may be merged to produce joint $M^c(z)$.

Analysing transfer function in (4), the compensator is highly sensitive to poles and zeros placement of the plant. If any of the zeros is placed outside the unity circle then the compensator poles are in unstable region. In the effect compensated plant (M^C) would consist of internally unstable signals which vein is difficult to predict as explained in [8]. Moreover, characteristic equation of the compensator is impacted by bilinear coefficients which may be a cause of exceeding the unity circle for high values of output (see Section 5.1).

4 Generalised Predictive Control

Two forms of the GPC are utilized basing on [9].The aim of GPC is to minimise a criterion which includes future states of the plant as in (6) where d is the delay of plant, λ is tuneable parameter, H_c and H_p are respectively control and output prediction horizons.

$$J_k = \sum_{i=d}^{H_p}\left[y_{k+i} - r_{k+i}\right]^2 + \sum_{j=1}^{H_c}\lambda\left[\Delta u_{k+j-1}\right]^2 \tag{6}$$

The prediction is based on a model, hence any discrepancies between the model and the plant can be observed easily as perturbation of the control.

Two variants of GPC are used, the difference is in derivation of a control law based on: linear ARX model as shown in [9], bilinear model transfer function equivalent with bilinear term in the denominator as in [3]. In this approach past state of input are affecting $\tilde{A}(z^{-1}) = A(z^{-1}) + \eta(z^{-1})u(z^{-1})$ coefficients. In practice the controller

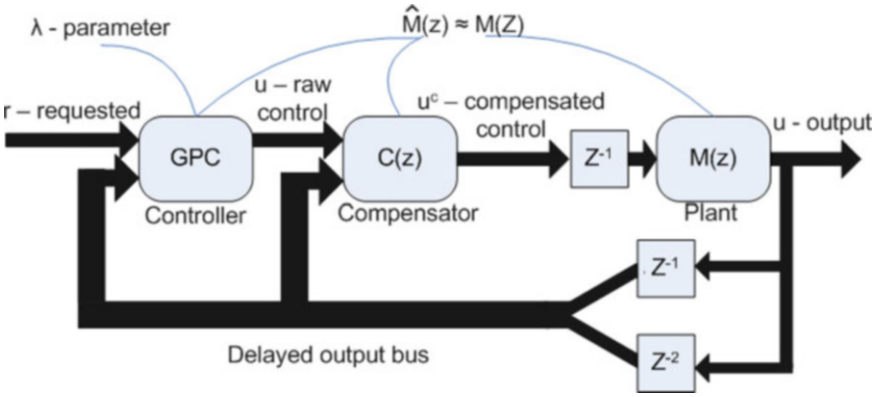


Fig. 2 A general scheme of the compensated GPC set-up

requires to recalculate $\tilde{F}$ and $\tilde{G}$ matrices [3] at every time sample, as $\tilde{a}_i(k)$ coefficients are considered as time varying.

The experiment set-up consists of three control scenarios:

- GPC basing on coefficients of linear model (referred as linear GPC),
- GPC utilizing bilinear model where the control law is updated in each iteration due to the point of operation (referred as bilinear GPC),
- GPC using linear terms of bilinear model followed by the compensator as in Fig. 2 (referred as compensated GPC).

The utilization of linear controller is aimed to outline the superior efficiency of bilinear model. The comparison of bilinear and compensated GPC is essential to evaluate the compensator. All controller use the same settings of prediction horizon $h_c = h_p = 5$, delay ($d = 1$) and the same set of tuneable parameter λ.

5 Investigations

In this Section implementation aspects and efficiency of the compensator are presented. The assessment is done with respect to linearisation capabilities and issues in Section 5.1 and with respect to the full control task in Section 5.2. The complete experimentation system is done in Matlab environment.

The quality of control in Section 5.2 is assessed due to two numerical criteria: Mean Square Error (7) and Control Cost (8) where N is length of the experiment.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - r_i)^2 \tag{7}$$

$$CC = \frac{1}{N}\sum_{i=1}^{N}u_i^2 \tag{8}$$

Two properties altered by the compensation are analysed: gain (or steady state gain) - K_{SS} and settling time t_{SS}, i.e. time after which the error $r_k - y_k$ is lower than 2 %.

Table 1 Coefficients of both plants used in Section 5.1

Plant	a_1	a_2	b_0	b_1	μ_1	μ_2
I	− 1. 384	0. 582	0. 092	− 0. 079	− 0. 016	− 0. 012
II	− 1. 174	0. 179	0. 092	− 0. 075	− 0. 016	0. 014

5.1 Vicarious Linear Plant

First experiment is performed on two ideal bilinear second order plants with minimum delay. The aim is to validate the nominal performance and issues with compensation for two cases:

- Stable internal signals for operating area when denominator of (2) has zeros in unity circle for all plausible y,
- Unstable internal signals in part of the operating range.

Sampling time T_s is assumed to be 1s for simplicity. Both systems are excited by stair-wise input from 0 to 8 with each step having an amplitude of 0. 5. Coefficients of both plants are described in Table 1.

First, assessment of compensators stability is done by calculation of poles placement with respect to the point of operation. Following assumptions are made. Input range u is 0 to 8. Output is simplified to be in steady-state, i.e. $Y = y_{k-1} = y_{k-2}$. Hence, compensator of plant I has poles described by (9).

$$z = (0.079 + 0.012Y)/(0.092 - 0.016Y)$$
$$|z| \leq 1 \quad \forall \quad Y \leq 0.4643 \tag{9}$$

Estimation of plant I gain for ($u = 8$) is resulting with $K_{ss} = 0.\ 0310$. By simple calculation, maximum expected output is $y = 0.\ 24$, hence stability condition of the compensator in (9) holds for the entire feasible range.

Characteristic equation of the plant II compensator and stability criterion are as in (10).

$$z = (0.075 - 0.014Y)/(0.092 - 0.016Y)$$
$$|z| \leq 1 \quad \forall \quad Y \leq 5.66 \tag{10}$$

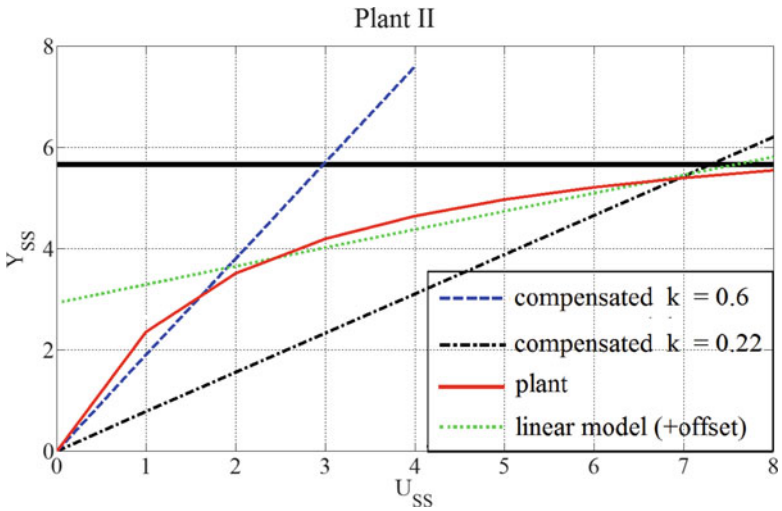


Fig. 3 Steady state gain of the plant II, linear approximation of model and compensated system with the most suitable gain K_c. Solid horizontal line is a maximum stable output

When excited with maximum input, plant II output in steady state is 5. 546. Steady state gain of the compensated vicarious plant is greater than the actual plant gain (please refer to Fig. 3), therefore it is rapidly exceeding stability criterion from (10). To sustain stability, the compensator (5) is modified with calibrateable gain K_c and saturation of output y_{max} as in (11).

$$u_k^c = \frac{-\sum_{j=2}^{n_\mu}\left[(b_{j-1}+\mu_j y_{k-j}^*)u_{k-j}\right] + K_c\sum_{i=0}^{n_b} b_i u_{k-i}^c}{b_0+\mu_1 y_{k-1}^*} \tag{11}$$

where $y^*$is minimum of y and y_{max}

The amendment helps sustaining stability and match K_{ss} of the actual and compensated plant. Following combination of K_c and y_{max} parameters are utilised in the experiment:

- Plant I compensated with $K_c = 0.6$,
- Plant II compensated with $K_c = 0.22$ and stability limit $y_{max} = 5.4$.

Fig. 4 presents the comparison of the plant properties with and without the modified compensator from (11). Compensated plants have constant properties which is characteristic for linear systems. The only exception occurs according to the stability criterion (10), i.e. when the saturation of compensation is active ($U > 7.5$ for plant II).

As a summary, it may be stated that the compensation is efficient in linearising the bilinear plant. However, when designing or considering the bilinear compensation, an analysis of compensator stability must be performed.

5.2 Control Task

The last experiment considers a full control task using three types of GPCs.Results for respective plants are compared with reference controllers in Table 2. In this experiment compensation gain was set to unity to avoid discrepancies. As it can be seen, with compensation MSE is improved significantly. This applies even for the case of saturated compensator (plant II). On the other hand, CC of u^c has risen with compensation, yet this is insignificant drop if compared to the increase of MSE.

6 Conclusion

The bilinear compensator (11) is validated to be an efficient tool to linearise bilinear plants in Section 5.1 and hence to improve controllability of such in Section 5.2. However, following constraints must be considered:

- In digital systems compensation must be included in controller's logic or plant's hardware, as no additional delay due to the compensator should be introduced (see Section 3)
- Zeros of the plant must be placed within the stable region as those become compensator poles,
- Stability consideration must include relocation of poles and zeros due to bilinearity (see Section 5.1) which can be limited by defining y_{max} - saturation of y in the compensator.

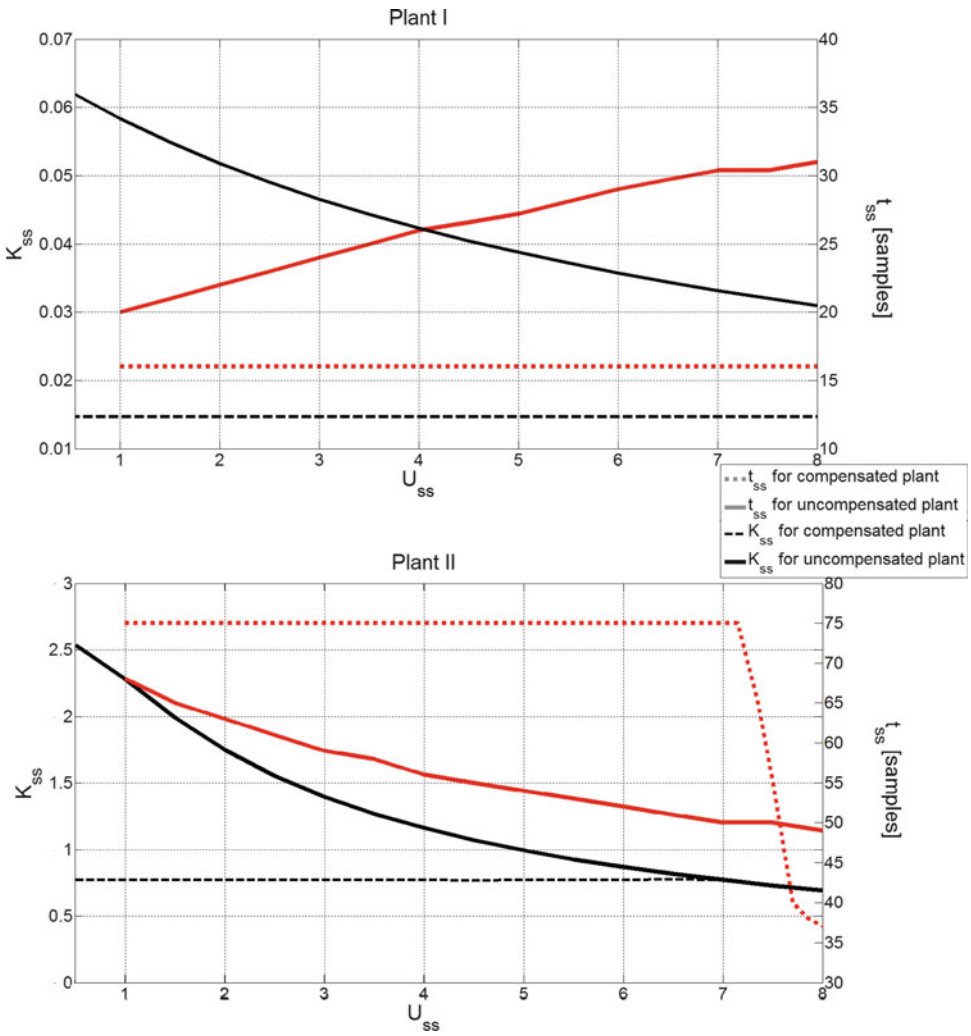


Fig. 4 Properties of compensated and uncompensated plant I (top) and plant II (bottom). Steady state gain and settling time are presented. Distortion of compensated plant II properties for $u > 7$ is due to the saturation of output

Table 2 Plant I and II control test result

	MSE		CC	
Name	value	% best	value	% best
Plant I				
linear GPC	$16 \cdot 10^{-5}$	1300%	0.2474	100%
bilinear GPC	$2.4 \cdot 10^{-5}$	200%	0.3588	144%
GPC+compensator	$1.2 \cdot 10^{-5}$	100%	0.4021	163%
Plant II				
linear GPC	$5.2 \cdot 10^{-3}$	433%	0.0036	100%
bilinear GPC	$2.1 \cdot 10^{-3}$	175%	0.0041	114%
GPC+compensator	$1.2 \cdot 10^{-3}$	100%	0.0044	122%

It has been observed that for the ideal case (bilinear plant, perfect model) a significant improvement over linear and bilinear GPC is obtained (Table 2). However, it can be easily deduced that the efficiency of compensator is dependant significantly on quality of the estimation. Hence, result is not likely to be resembling for a real plant.

7 Future Work

As in this paper the empirical work is limited to ideal bilinear plants, the next step is to implement the compensator in other non-linear plants. A good example of a practical non-linear plant similar to the bilinear in terms of properties is the four water-tank systems [10] or high-temperature industrial furnace [4]. The interesting point to be tested is application of the compensator to linearise the saturation, in e.g. electric motor, which may improve stability of MBC tuned to achieve dead-beat response.

8 Acknowledgement

This work was supported by the statutory funds of the Department of Systems and Computer Networks, Faculty of Electronics, Wroclaw University of Technology, Wroclaw, Poland.

References

1. Mohler R.R.: Natural Bilinear Control Processes. IEEE Transactions on Systems Science and Cybernetics, vol.6, 192-197 (1970)
2. Bruni C., DiPillo G., Koch G.: Bilinear systems: An appealing class of ”nearly linear” systems in theory and applications. Automatic Control, IEEE Transactions on, vol.19, 334-348 (1974)
3. Goodhart S.G., Burnham K.J., James D.J.G.: Bilinear self-tuning control of a high temperature heat treatment plant. Control Theory and Applications, IEE Proceedings, vol.141, no.1, 12-18 (1994)
4. Martineau S., Burnham K.J., Haas O.C.L., Andrews G., Heeley A.: Four-term bilinear PID controller applied to an industrial furnace. Control Engineering Practice, vol.12, 457-464 (2004)
5. Larkowski T., Burnham K.J.: System Identification, Parameter Estimation and Filtering. Published by WUT, Wroclaw (2011)
6. Pearson R. K.: Discrete-Time Dynamic Models. Oxford University Press, New York, USA (1991)
7. Taylor C.J., Chotai A., Burnham K.J.: Controllable forms for stabilising pole assignment design of generalised bilinear systems. Electronics Letters, Vol.47, Issue 7, 437-439 (2011)
8. Doyle J.C., Francis A.B., Tannenbaum A.R: Feedback Control Theory. Dover (2009)
9. Burnham K.J., Larkowski T.: Self-Tuning and Adaptive Control. Published by WUT, Wroclaw (2011)
10. Gatzke E.P., Meadows E.S., Wang C., Doyle F.J.: Model based control of a four-tank system. Computers and Chemical Engineering, vol.24, 1503-1509 (2000)

Estimation for Target Tracking Using a Control Theoretic Approach – Part 2

Stephen C. Stubberud, Arthur M. Teranishi, and Kathleen A. Kramer

1 Introduction

The Kalman filter, along with its many variants, is the standard estimator in many applications such as control systems [1], systems identification, and target tracking [2]. While the Kalman filter has been demonstrated to be effective on a number of applications (both real-world and textbook), there have been a number of its detractors who present problems where the Kalman filter fails. Often, these applications are on problems where the underlying assumptions of the Kalman filter have been violated egregiously. One assumption that often is violated is that of observability.

The Kalman filter, as with any state estimation technique, is a variation of a state observer [3]. For an observer to be effective, the mode must be excited and the state must be observable. The mode requirement is evident: if one does not excite the system with a desired frequency, the observer cannot "observe" how the system behaves to that frequency. More importantly, if the sensors that measure the system cannot provide measurements of certain dynamics, either directly or indirectly, then there exist unobservable dynamics. While this is an important issue in all applications of state estimation, it is of particular concern for the sensor fusion problem. In target tracking, if a sensor provides only an angle observation of the target, for example, then the target's position and velocity cannot be completely known. This is a result of an unobservability of the target state. This unobservability can be alleviated in the single platform case only when the sensor platform's dynamics are at least one derivative above the target's dynamics (e.g., when there is a moving platform with a stationary target).

To compensate for the unobservability, some techniques have been employed to modify the estimate based on some prior or perceived knowledge. Estimation techniques such as particle filters [4], interacting multiple models (IMMs) [5], and constrained estimators [6] have been used as well. However, all of these techniques are estimators, and therefore they rely on the observations and measurements to drive the state estimates to the correct values. If the measurements do not provide the information necessary for complete observability, then the estimates cannot be corrected properly.

In [7], the first of three papers was presented that examined the use of a control input in the Kalman filter to correct the state vectors. The benefit of the control law is that it can exist in many forms and is designed to use external information. It can provide a system input to the estimator. Since estimation techniques were originally designed to work in conjunction with control laws [8], the underlying theory of the estimation technique is not violated and avoids many of the ad hoc techniques used to modify the estimation techniques currently in practice.

In this second of the three papers on the use of the control law to aid in the estimation using a Kalman filter, the problem of the system with unobservability is researched as applied to the target tracking problem. In the next section, the Kalman filter with a control input is presented. Section 3 follows with the conceptual implementation of the control law. This is followed by the definition of the example and the definition of the control law. Section 5 shows the results and discusses their implications on the problem of estimation in the presence of observability issues.

S.C. Stubberud (✉)
Oakridge Technology, Del Mar, USA
e-mail: scstubberud@ieee.org

A.M. Teranishi
Asymmetric Associates, Long Beach, USA
e-mail: art.t@cox.net

K.A. Kramer
Department of Electrical Engineering, University of San Diego, San Diego, USA
e-mail: kramer@sandiego.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_8,

2 Control-Based Kalman Filter

The standard extended Kalman filter (EKF) for the tracking problem is given as the recursive equations

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}^T\left(\mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_{k|k-1}^T + \mathbf{R}_k\right)^{-1} \tag{1}$$

$$\mathbf{x}_{k|k} = \mathbf{x}_{k|k-1} - \mathbf{K}_k\left(\mathbf{z}_k - \mathbf{h}_k\left(\mathbf{x}_{k|k-1}\right)\right) \tag{2}$$

$$\mathbf{P}_{k|k} = \left(\mathbf{I} - \mathbf{K}_k\mathbf{H}_{k|k-1}\right)\mathbf{P}_{k|k} \tag{3}$$

$$\mathbf{x}_{k+1|k} = \mathbf{F}\mathbf{x}_{k|k} \tag{4}$$

$$\mathbf{P}_{k+1|k} = \mathbf{F}\mathbf{P}_{k|k}\mathbf{F}^T + \mathbf{Q}_k \tag{5}$$

where the state vector is given as the position velocity states

$$\mathbf{x}^T = [x \quad \dot{x} \quad y \quad \dot{y} \quad z \quad \dot{z}] \tag{6}$$

and the motion model $\mathbf{F}$ is the straight-line motion model as defined in [9]. When the observations are position measurements

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{7}$$

or range–bearing– elevation measurements

$$\mathbf{h}\left(\mathbf{x}_{k|k-1}\right) = \begin{bmatrix} \rho \\ \alpha \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \sqrt{\left(x_{tgt} - x_{platform}\right)^2 + \left(y_{tgt} - y_{platform}\right)^2 + \left(z_{tgt} - z_{platform}\right)^2} \\ \arctan\left(\dfrac{\left(x_{tgt} - x_{platform}\right)}{\left(y_{tgt} - y_{platform}\right)}\right) \\ \arctan\left(\dfrac{\left(z_{tgt} - z_{platform}\right)}{\sqrt{\left(x_{tgt} - x_{platform}\right)^2 + \left(y_{tgt} - y_{platform}\right)^2}}\right) \end{bmatrix} \tag{8}$$

the state vector of Eq. (2) is fully observable. However, such measurements are not always available. In some cases, due to sensor limitation, jamming, or stealth operations, the sensor on a platform may only provide an angle-only observation,

$$\mathbf{h}\left(\mathbf{x}_{k|k-1}\right) = \begin{bmatrix} \alpha \\ \varepsilon \end{bmatrix} \tag{9}$$

or when a synthetic aperture radar (SAR) is used, the observation is a range – range rate measurement

$$\mathbf{h}\left(\mathbf{x}_{k|k-1}\right) = \begin{bmatrix} \rho \\ \dot{\rho} \end{bmatrix} = \begin{bmatrix} \rho \\ \dfrac{x}{\rho}\dot{x} + \dfrac{y}{\rho}\dot{y} + \dfrac{z}{\rho}\dot{z} \end{bmatrix} \tag{10}$$

In these cases, the EKF can become overconfident in covariance, i.e., go to zero, while the true error can be growing without bound. A significant problem occurs in that not only is the position partially unobservable but the velocity vector is indirectly observed from the target position over time and, thus, is unobservable as well. While some may use ad hoc methods to adjust the Kalman filter's behavior to compensate for the unobservability, this violates the basic tenets of the Kalman filter. As presented in [10], to modify the estimation problem requires that the underlying principles be changed. However, the EKF was developed originally for feedback control. Instead of modifying the EKF, the proposed method for the target tracking problem is to only track the position and modify the velocities using a control law. So the new state vector becomes

$$\mathbf{x}^T = [x \quad y \quad z] \tag{11}$$

The EKF control application modifies the prediction equations, Eqs. (4) and (5) such that they become

$$\mathbf{x}_{k+1|k} = \mathbf{F}\mathbf{x}_{k|k} + \mathbf{g}(\mathbf{u}_k) \tag{12}$$

$$\mathbf{P}_{k+1|k} = \mathbf{F}\mathbf{P}_{k|k}\mathbf{F}^T + \mathbf{G}E\left[\mathbf{u}\mathbf{u}^T\right]\mathbf{G}^T + \mathbf{Q}_k \tag{13}$$

while the other equations remain unchanged. The motion model reduces to a 3x3 identity matrix as the velocity is and input vector. The input vector is modeled as a control law, which can be modified using a myriad of techniques including intelligent based methods.

3 Control Design

In [7], the control that was used was simply the velocity vector created from the range and bearing measurements over time. The noisy measurements were fully observable with respect to position. Therefore, the noisy positions could be calculated over two time points and a velocity estimate could be created as shown in Eq. (14).

$$\mathbf{u}^T = \left[\frac{x(t_k) - x(t_{k-1})}{t_k - t_{k-1}} \quad \frac{y(t_k) - y(t_{k-1})}{t_k - t_{k-1}} \quad \frac{z(t_k) - z(t_{k-1})}{t_k - t_{k-1}} \right] \tag{14}$$

In this effort, the control law is modified in two ways. The first way is to use the approach of Eq. (14) when the state or state component is fully observable. The second is to use the observability Gramian

$$\Gamma_i = \left\langle f_i, \frac{\partial \mathbf{h}}{\partial x_i} \right\rangle_{\mathbf{P},\mathbf{R}} \quad \forall i \in (1,2,3) \tag{15}$$

to provide a weighting to the velocity control and its associated uncertainty when the state is fully or partially unobservable.

The velocity control will be smoothed over time. This done by windowing a two-point point with a ten point exponential decay weighted average

$$u_i = \sum_{l=1}^{10} \alpha_l \Gamma_i^l \left(\frac{x_i(t_l) - x_i(t_{l-1})}{t_l - t_{l-1}} \right) \tag{16}$$

The weighting is $1/2^l$. In tracking applications, often a speed estimate of the target can be made based on the scenario. While this estimate has a large uncertainty, the resulting velocity is a good starting point for the filter. In this effort, the target is assumed to head in the negative y-axis direction with a constant velocity.

4 Scenarios

Four baseline scenarios were generated for this investigation. These four scenarios represent four separate intercept problems. In the first scenario, the platform starts at (0,0) and heads at 1000 ft/sec towards the target that is initially 600,000 ft (or 100 nmi) away. The target approaches the platform at 1000 ft/sec. This is shown in Figure 1. The platform is assumed to have a bearings-only tracking capability that provides updates at 1 second. We assume that the measurements have a 1.0 degree error that is normally distributed. The scenario ends when the target is 10 nmi away from the platform.

The second scenario changes the target initial location and heading. The target begins at approximately 60 nmi north of the target and approximately 30 nmi to the east. The target heads at 1000 ft/sec to the west. The other parameters are the same. The scenario is shown in Figure 2.

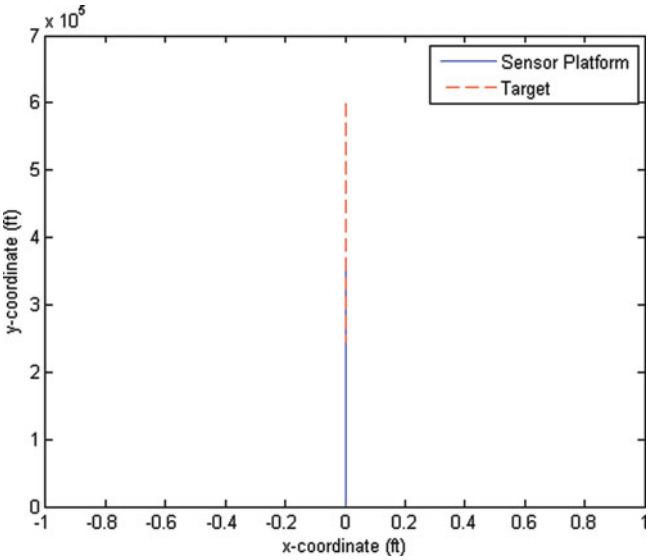


Fig. 1 Head on target is tracked by the platform

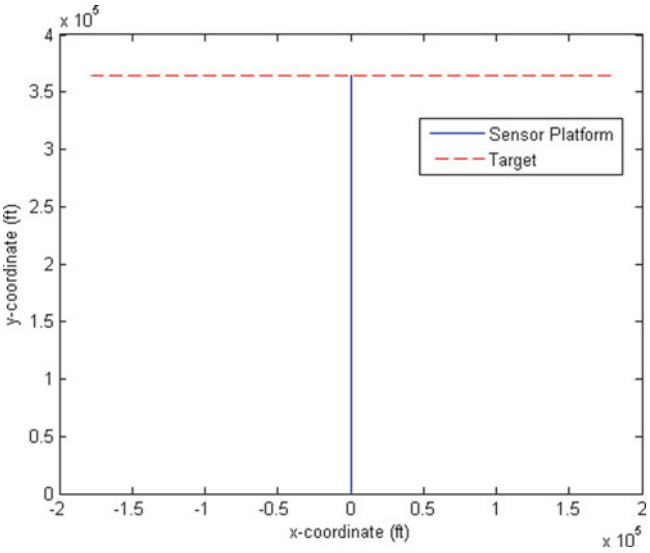


Fig. 2 The target begins from the east and heads across the sensor platform's trajectory headed west

The third scenario varies scenario two by the trajectory of the platform. The target platform has a velocity of 707 ft/sec to the west and 707 ft/sec to the south. The scenario ends when the target passes the noise of the sensor platform. The scenario trajectories are seen in Figure 3.

The final scenario again has the straight on target. The platform weaves giving the track an observability that would not exist without the platform maneuver. The platform speed is 1000 ft/sec with a 45 degree weave angle from north. The weaves have a 5 nmi between turns. The scenario is seen in Figure 4.

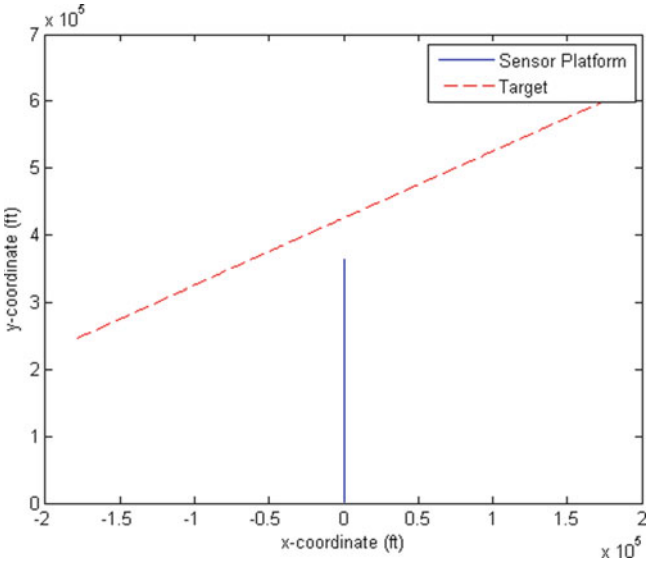


Fig. 3 The target crosses the platform trajectory at an angle which changes the aspect angle of the target track

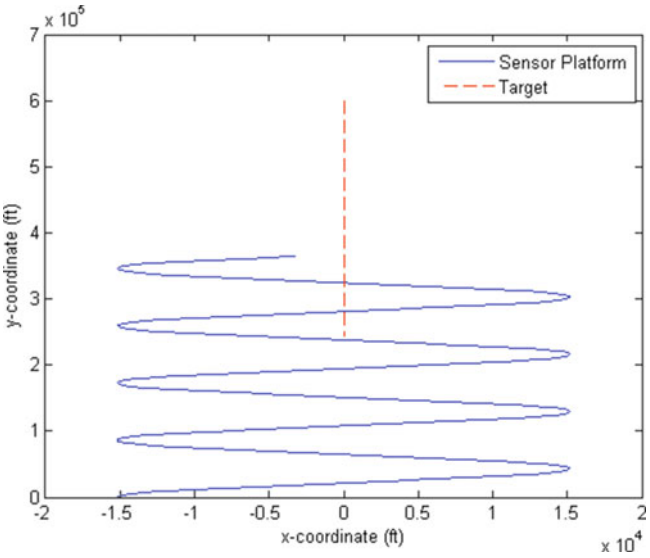


Fig. 4 Scenario 4 has the sensor platform weave which is used to track targets when angle-only measurements are solely used

In all four scenarios the straight-line motion model is used in the tracking algorithm. The initial velocity is assumed to be 900 ft/sec south. The initial target is assumed to be 100 nmi from the platform at the angle reported as this would be the maximum detection range. The initial uncertainty on position is assumed to be 20 nmi in the range direction. The uncertainty in velocity is assumed to be 800 ft/sec. The process noise $\mathbf{Q}$ is given as integrated white noise [9] with a scale factor of 1.7.

5 Results

The four scenarios were all implemented using the control implementation and a standard tracking method. The standard tracking method utilized a straight-line motion model with a state vector that includes the three-dimensional positions and their associated velocity components. The initial error covariance $\mathbf{P}$ was set to be

$$\mathbf{P}_0 = \begin{bmatrix} 2,500 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1,000,000 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2,500 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1,000,000 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{17}$$

for the standard method while the first, third, and fifth diagonal elements were used to set the initial $\mathbf{P}$ for the control-based tracking technique. For the first scenario, the results are seen in Figures 5 (control) and 6 (standard

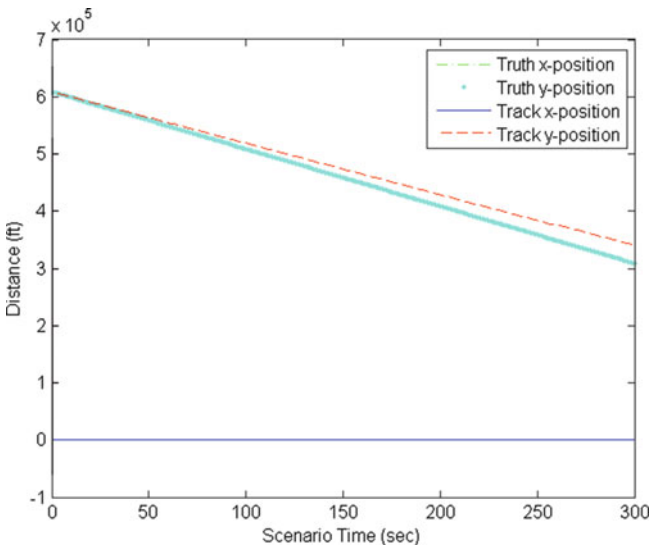


Fig. 5 The control approach allows for the unobservable state to be tracked based on problem specific knowledge

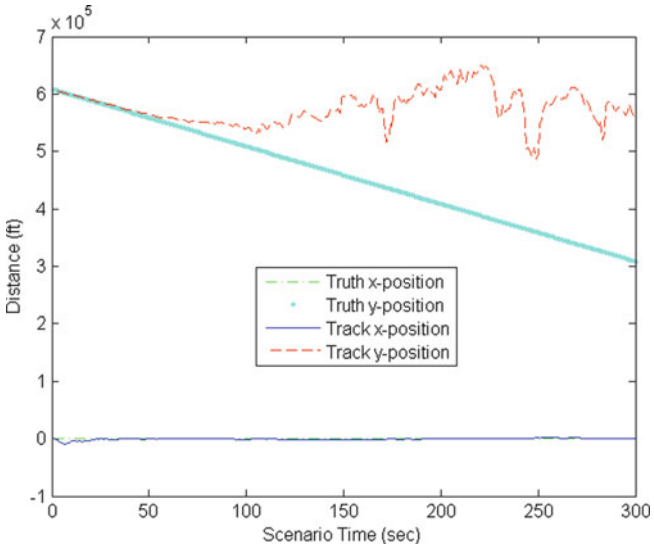


Fig. 6 The standard tracking approach is unable to handle the unobservable state for a significant period of time

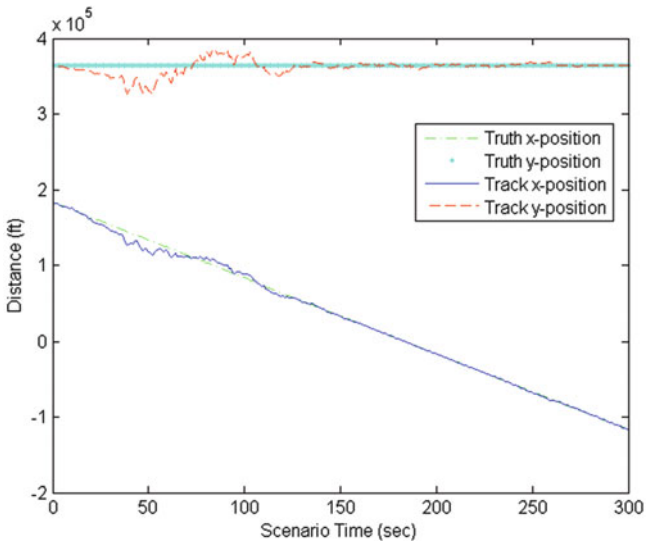


Fig. 8 The standard tracking approach has greater observability and tracks significantly better

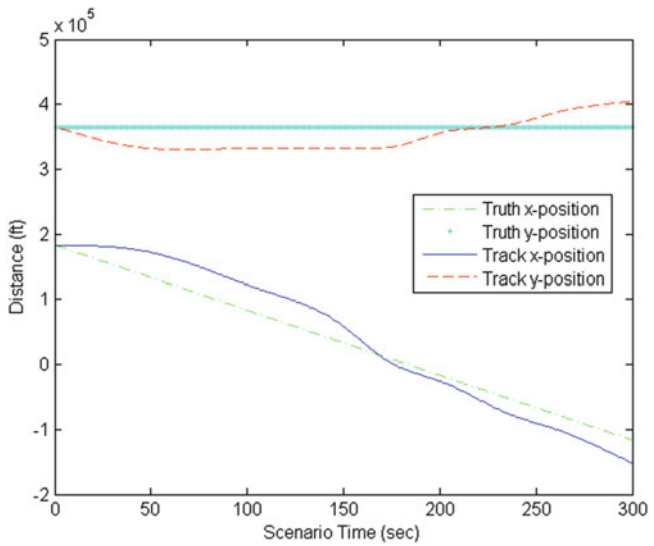


Fig. 7 For the crossing target, the track can be held well with the control technique

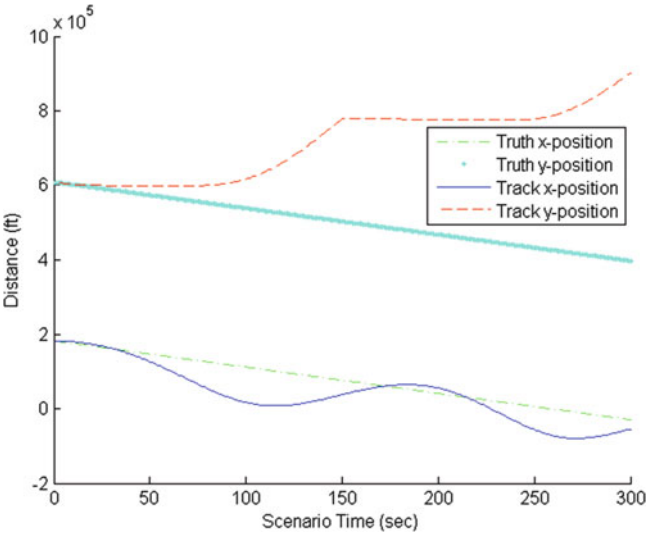


Fig. 9 The lightly observable state of y-axis position starts to track poorly

method). The results clearly indicate that by modeling the intercept tracking problem into a control signal, the target may be tracked.

The results in the second scenario are shown in Figures 7 and 8. The state estimation technique in Figure 8 performs noticeably better than does the control-based method. In this case the states have greater observability. As discussed in Part 1 of this research [7], the control approach does poorer when the states are completely observable. But the control-based approach's performance is still within acceptable limits based on the covariance.

When the target cuts an angle across the sensor platform, the observability of the states is greatly varying. In Figures 9 and 10, the results are shown. Without the control, a bias exists in the x-position that requires a closest point of approach (CPA) event to occur to correct for it. The CPA is a standard tracking phenomena that helps localize the target in at least one dimension. The control compensates for the x-coordinate error and does a better job of tracking the y-coordinate that without the modification. However, both techniques clearly have poor performance on this scenario.

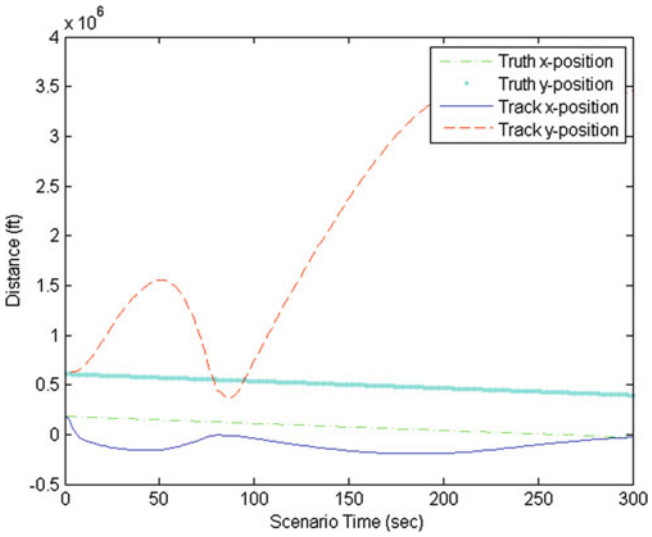


Fig. 10 The standard tracking approach cannot track either axis well

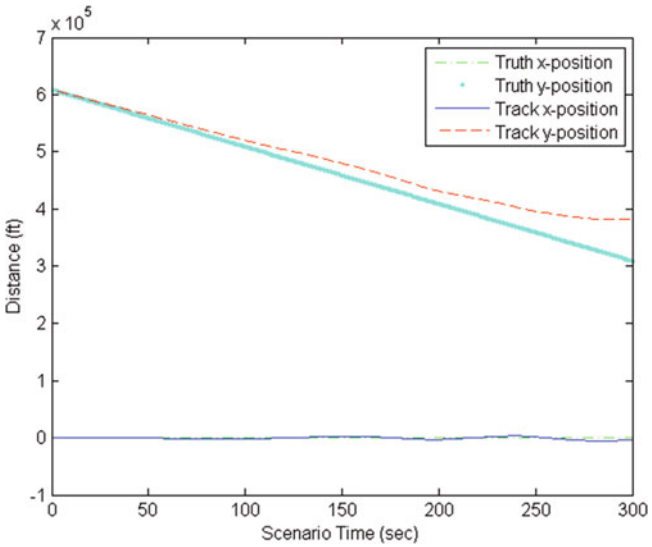


Fig. 11 The control approach lags in the y-axis tracking but still performs well within performance criteria

The final scenario results are seen in Figures 11 and 12. In this case, the states of the track are fully observable since the platform is maneuvering. The control method has the lag in the y-axis tracking but still tracks the target within the covariance error. The standard tracker is noticeably better as the observability benefits the estimator.

When the target track contains an unobservable state the control based approach performs better. When observability is not an issue, the control-based tracking approach perform inferiorly to that of the estimator but still meets covariance –based performance requirements.

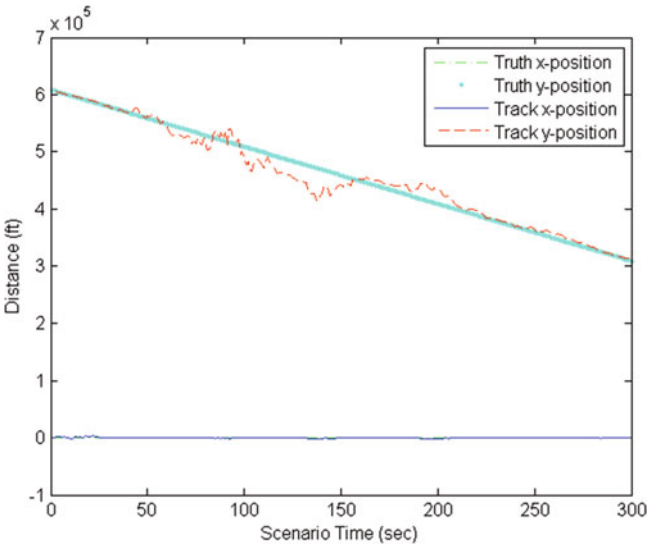


Fig. 12 The fully observable states give a performance edge to the standard tracking method

6 Conclusions

A fixed control technique was applied to the standard intercept tracking problem. The results continue to support the effectiveness of this method to target tracking. When observability issues arise, the technique has demonstrated its utility. The approach used a fixed control methodology that was designed for the head-on target intercept model. It was still effective in the other scenarios.

The fixed control needs to be modified to an intelligent system in the third part of this research. Once the intelligent design is incorporated, the technique can be analyzed in covariance behavior and incorporated into a interacting multiple model structure.

References

1. R. G. Brown, Introduction to Random Signal Analysis and Kalman Filtering. New York: Wiley, 1983.
2. S. Blackman and R. Popoli, Design and Analysis of Modern Tracking Systems. Boston: Artech House, 1999.
3. M. S. Santina, A. R. Stubberud, and G. H. Hostetter, Digital Control System Design. Fort Worth: Saunders College Pub., 1994.
4. M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking," IEEE Trans. on Signal Processing, Vol. 50, No. 2, February 2002, pp. 174-188.
5. H. A. P. Blom and Y. Bar-Shalom, "The Interacting Multiple-Model Algorithm for Systems with Markovian Switching Coefficients," IEEE Trans. on Automatic Control,Vol. 33, No. 8, August 1988, pp. 780-783.
6. A. R. Stubberud, "Constrained Estimate of the State of a Time-Variable System," Proc. of the 14th International Conference on Systems Engineering, Coventry, UK, August 2000, pp. 513-518

7. Stephen C. Stubberud, Arthur M. Teranishi, “Estimation for Target Tracking Using a Control Theoretic Approach - Part I,” Advances in Systems Science - Proceedings of the International Conference on Systems Science 2013, ICSS 2013, Wroclaw, Poland, September 2013, pp. 35-43.
8. D. L. Snyder and I. B. Rhodes, “Filtering and control performance bounds with implications on asymptotic separation,” Automatica, Vol. 8, No. 6, November 1972, Pages 747-753.
9. Bar-Shalom, Y. and X.-R. Li, Estimation and Tracking: Principles, Techniques, and Software, Artech House Inc., Norwood, MA, 1993.
10. Yang, C. and E. Blasch, “Fusion of Tracks with Road Constraints,” Journal of Advances of Information Fusion, Vol. 3, No. 1, pp. 14 –31, June, 2008.

Identification of Fractional Order Models: Application to 1D Solid Diffusion System Model of Lithium Ion Cell

Walid Allafi, Ivan Zajic, and Keith J. Burnham

1 Introduction

Mathematical modelling of Li-ion and Ni/MH batteries, including solid-phase processes, has recently gained a considerable attention in the literature [1, 2]. The charging and discharging process of the battery consists of de-intercalation of lithium from the positive and negative electrode, respectively. Subsequently, the diffusion of lithium ions through the electrolyte occurs, which is followed by the intercalation of lithium back into the negative and positive electrodes. Consequently, the battery model is commonly derived in terms of the concentration of solution using the porous electrode theory [3]. Expressing the lithium intercalation in the electrode as a solid diffusion and solving the resulting spherical diffusion equation is a commonly adopted modelling approach [4]. In this approach, the pseudo 2D model is constructed, where one of the dimensions is a spread between two collectors while the other dimension extends into the solid particle. In this regard, [5] derived an extended one-dimensional model of Li + battery, which is also adopted in this work. In this model, the one dimensional concentration as well as the distribution of the electric potential is governed by the boundary-value-problems.

The development of a mathematical model of the Li + battery consists of establishing the governing equations for the depended variables, such as the concentration of lithium ions. Additionally, the initial as well as the boundary conditions must be stated and suitable numerical solver technique must be selected. A detailed review of the governing equations, which could be applied for the porous electrode case, can be found in [6]. The overall energy balance equations for the battery systems can be found in [7].

1.1 Motivation for Fractional Order Modelling

There are several studies in electrochemistry for deducing the concentration of electro-active species in the electrode surface. It has been experimentally found in [8] that the characteristics of the surface of concentration of the active material can be described as $m(t) = D_t^{-0.5} i(t)$, where $i(t)$ denotes the electric current and the term $D_t^{-0.5} i(t)$ is a half integral of the current reactor [8]. The same can be noted about the Randle's model [9], which can also be considered for the battery modelling problem. This is due to a simplified resolution of the electrochemical diffusion equation [9]. The fractional order behaviour of Randle's model is due to the fractional impedance $W(s)$. This impedance is also known as Warburg cell and is a fractional order integrator of order $\alpha = 0.5$.

1.2 Fractional Order Model Identification

The refined instrumental variable method for parameter estimation of continuous-time linear transfer-function models (RIVC), proposed by P.C. Young et al. in [10], and is considered. The RIVC method is statistically optimal under the assumption of an autoregressive moving-average (ARMA) additive noise model and is suitable for identification of continous-time (CT) hybrid Box-Jenkins transfer function models [10]. This model is hybrid in the sense that the deterministic part of the model is estimated in the CT domain, while the noise model is estimated in discrete-time (DT) domain. Simplified version of RIVC method, abbreviated SRIVC, assumes output error noise scenario

W. Allafi (✉) • I. Zajic • K.J. Burnham
Control Theory and Applications Centre, Coventry University, Coventry CV1 5FB, UK
e-mail: allafiw@uni.coventry.ac.uk; i.zajic@coventry.ac.uk; ctac@coventry.ac.uk

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_9,

only while providing consistent, however, suboptimal, results under the ARMA noise scenario.

The extended SRIVC method for fractional order model estimation, abbreviated SRIVCF, has been introduced in [11], while the corresponding extension of RIVC method is provided in [12]. The advantage of these methods is the use of directly measured sampled input-output signals for the estimation of otherwise continuous-time models.

2 Li-ion Solid State Diffusion Mathematical Model

The current section presents a well-established solid-phase diffusion partial differential equation [1], occurring in spherical particles as illustrated in Fig. 1. The diffusion process follows the Fick's law and is presented in the spherical coordinates as follows:

$$\frac{\partial c}{\partial t} = D\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial c}{\partial r}\right) \tag{1}$$

where at $t = 0$, for $0 \leq r \leq R$, $c(t = 0) = c_0$ and

$$\left.\frac{\partial c}{\partial r}\right|_{r=0} = 0 \tag{2}$$

$$-D\left.\frac{\partial c}{\partial r}\right|_{r=R} = j(t) \tag{3}$$

The parameter c denotes the lithium concentration in the particles of solid, D denotes the diffusion coefficient, $j(t)$ refers to the boundary flux at time instance t and R represents the particle radius.

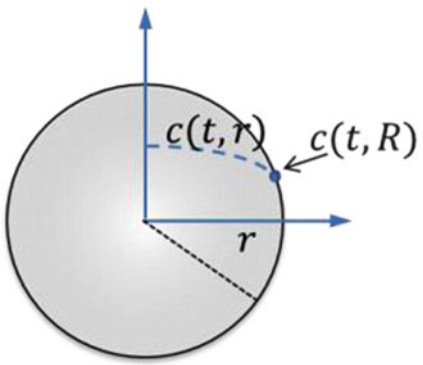


Fig. 1 Solid state diffusion particle for Li-ion cell

If the dimensionless form of the above model is considered, the diffusion coefficient as well as the radius R need to be assumed as constants. The model (1) is then represented as follows:

$$\frac{\partial \overline{c}}{\partial \tau} = D\frac{1}{\overline{r}^2}\frac{\partial}{\partial \overline{r}}\left(\overline{r}^2\frac{\partial \overline{c}}{\partial \overline{r}}\right) \tag{4}$$

where $\overline{c}(\tau = 0) = \overline{c}_0$ and

$$\frac{\partial \overline{c}}{\partial \overline{r}}|_{\overline{r}=0} = 0 \tag{5}$$

$$\frac{\partial \overline{c}}{\partial \overline{r}}|_{\overline{r}=R} = -\delta(\tau) \tag{6}$$

The following dimensionless variables are introduced $\overline{c} = c/c_{max}$ and $\tau = Dt/R^2$, where c_{max} [mol.m^{-3}] denotes the maximum concentration of the lithium in the particle.

3 Simplified Refined Instrumental Variable Method for Continuous-Time Fractional-Order Models

Since no measurement noise is introduced during the proposed model order reduction procedure, the SRIVCF method is adopted for the fractional-order model estimation. The minimised error function is given by

$$e(t_k) = y(t_k) - \frac{B(s^\alpha)}{A(s^\alpha)}u(t_k) \tag{7}$$

$$= A(s^\alpha)\frac{1}{A(s^\alpha)}y(t_k) - B(s^\alpha)\frac{1}{A(s^\alpha)}u(t_k)$$

where $B(s^\alpha)$ and $A(s^\alpha)$ represent input and output polynomials and differential operator s is defined such that $s = \frac{d}{dt}$. Introducing filtered forms of $y(t_k)$ and $u(t_k)$ as $y_{fA}(t_k)$ and $u_{fA}(t_k)$, respectively, leads to

$$e(t_k) = A(s^\alpha)y_{fA}(t_k) - B(s^\alpha)u_{fA}(t_k) \tag{8}$$

Subsequently, the expression (8) is rearranged into a pseudo-regression form, which can be solved as a least squares problem, i.e.

$$y_{f_A}^{\alpha n}(t_k) = \varphi_{fA}^T\theta + e(t_k) \tag{9}$$

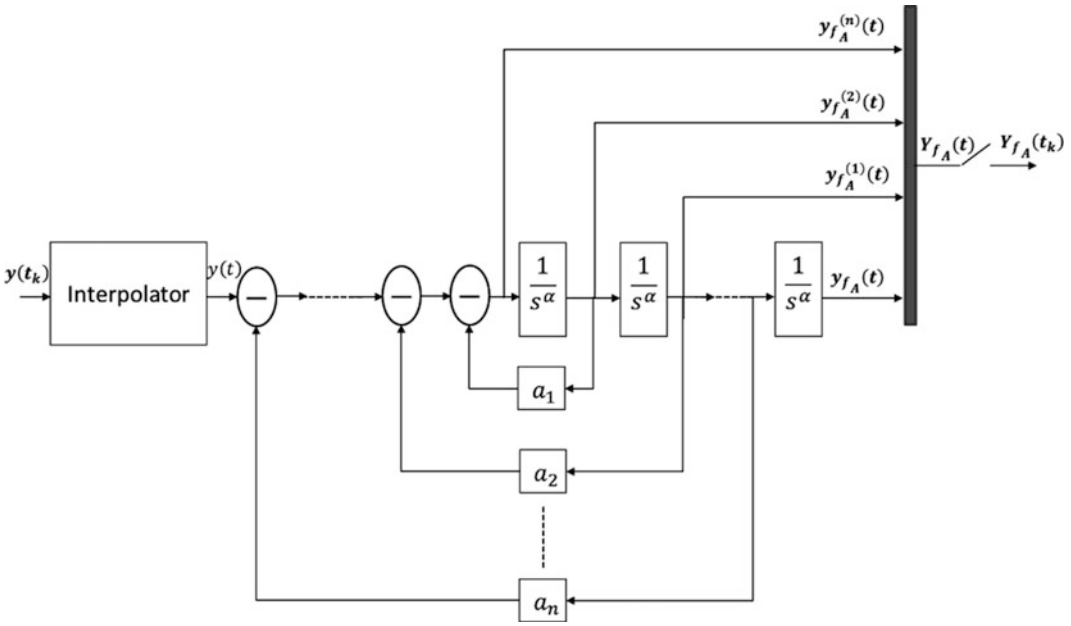


Fig. 2 State variable filter implementation. Filtering of output signal is considered

where $y_{fA}^{\alpha n}(t_k)$ denotes the highest output time derivative of order n and fraction α. The regression vector, parameter vector and filtered fractional order time derivatives of output and input signals are obtained, respectively, as follows

$$\begin{aligned}\varphi_{f_A(t_k)} = &\left[-y_{f_A}^{\alpha(n-1)}(t_k), -y_{f_A}^{\alpha(n-2)}(t_k), \ldots\right.\\ &\left.-y_{f_A}(t_k) \quad u_{f_A}^{\alpha(m)}(t_k) u_{f_A}^{\alpha(m-1)}(t_k), \ldots u_{f_A}(t_k)\right]^T\\ \theta = &[a_1, a_2, \ldots, a_n b_0, b_1, \ldots, b_m]^T\end{aligned} \tag{10}$$

The state variable filtering operation of output and input signals is illustrated in Fig. 2 for the output signal only. The filtering operation has been implemented as the following state-space equation, which has been discretised by adopting a zero-order-hold assumption on filtered input-output signals, hence

$$\begin{bmatrix} y_{f_A}^{\alpha(n)}(t) \\ y_{f_A}^{\alpha(n-1)}(t) \\ \vdots \\ y_{f_A}^{\alpha}(t) \end{bmatrix} = \begin{bmatrix} -a_1 & -a_2 & \cdots & -a_n \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{f_A}^{\alpha(n-1)}(t) \\ y_{f_A}^{\alpha(n-2)}(t) \\ \vdots \\ y_{f_A}(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} y(t) \tag{11}$$

The SRIVCF method is implemented as a two stage algorithm provided below:

Step 1: The initial model parameters need to be obtained for the purpose of deducing the filtered fractional order time derivatives of input-output signals, which are subsequently used in the stage number two. A several techniques exist, such as fractional-order (FO) least squares, FO frequency analysis [13] or FO state variable filtering can be applied. In the following numerical study the FO frequency analysis is adopted. The obtained initial parameters are used by the fractional-order commensurate stable filter, whose denominator has a same order as the $A(s^\alpha)$ polynomial, i.e.

$$f(s^\alpha) = \frac{1}{(s^\alpha + \lambda)^n} \tag{12}$$

Step 2: Iteratively repeat sub-steps (I) to (III) until the obtained sum of the squares of the differences between $\widehat{\theta}_i$ and $\widehat{\theta}_{i-1}$ is satisfactorily small.

I. Pre-filter $y(t_k)$ and $u(t_k)$ to generate the filtered time derivatives of the output $y(t_k)$ and input $u(t_k)$ signals as in (11), respectively:
II. Generate a vector $y^{IV}(t_k)$ of instrumental variables using the estimated model from the previous iteration step.
III. The parameter estimates are obtained by:

$$\widehat{\theta}_i = \left(\frac{1}{N}\sum_{k=1}^{N} \varphi_{f_A}^{IV}(t_k)\varphi_{f_A}(t_k)\right)^{-1} \frac{1}{N}\sum_{k=1}^{N} \varphi_{f_A}^{IV}(t_k) y_f^{\alpha(n)}(t_k)$$

where the regression vector obtained by (10) and regression vector of instrumental variables is defined as follows:

$$\begin{aligned}\varphi_{fA}^{IV}(t_k) = &\left[-y_{fA}^{\alpha(n-1)}(t_k)^{IV}, -y_{fA}^{\alpha(n-2)}(t_k)^{IV} \ldots - y_{fA}(t_k)^{IV}\right.\\ &\left. u_{fA}^{\alpha(m)}(t_k), u_{fA}^{\alpha(m-1)}(t_k), u_{fA}^{\alpha(m-2)}(t_k), \ldots u_{fA}(t_k)\right]\end{aligned}$$

The parameters of the state variable filter (12) are being iteratively updated using the latest parameter vector estimates $\widehat{\theta}_i$, hence

$$f(s^{\alpha}) = \frac{1}{\tilde{A}(s^{\alpha})}$$

4 Frequency Domain Comparison of Fractional Order Modelling Approach With Finite Volume Method

The analytical solution to the 1D solid state diffusion equation, provided in (4) in form of a partial differential equation with $r = R = 1$ and $D = 1$, is given by Jascobsen and West model [14], i.e.

$$G(r,s) = \frac{\overline{c}(1,s)}{\delta(s)} = \frac{-\tan h(\sqrt{s})}{\tan h(\sqrt{s}) - \sqrt{s}} \tag{13}$$

The bode diagram of the analytical solution (13) is shown in Fig. 3 as a solid black line.

One of the common approaches to numerically solve the equation (4) is to adopt the Finite Volume Method (FVM). In this method the particle is divided into a number of control volumes (CVs). The Bode diagram of dimensional concentration on the surface of the particle using FVM for 10, 50, and 100 CVs is shown in Fig. 3 as plus, star and square symbols, respectively. It can be seen that the higher the number of control volumes the better model fit to analytic solution is achieved in high frequency.

For comparison, the following FO model of the analytical solution (13) to the diffusion model (4) has been identified

$$\widehat{G}(s) = \frac{\overline{c}(s)}{-\delta(s)} = \frac{\left(s^{0.5} + 3\right)}{s} \tag{14}$$

where the FO derivative term $s^{0.5}$ has been approximated by Oustaloup method [15] to be 10^{th} order linear transfer function. In other words, the FO model (14) has the same order as FVM with 10 CVs. The analytical solution and approximated Bode diagram of FO model (14) are presented in Fig. 3 as dots and triangles, respectively. Note, that despite the fact the approximated FO model is simulated by 10^{th} order transfer function model it outperforms the FVM with 100 CVs (for high frequencies).

From the above it follows, that the FO modelling approach can be considered as a model order reduction technique similarly to truncation based, projection based and data based methods [16]. In this case, the high order model is approximated with the FO model using presented SRIVCF algorithm.

5 Numerical Example: SRIVCF Estimation

The partial differential equations (4-6) are numerically solved by applying FVM [17]. The spatial domain is divided into 1000 CVs with sampling interval 10^{-6}. The FO system is simulated using the Oustaloup approximation method [15].

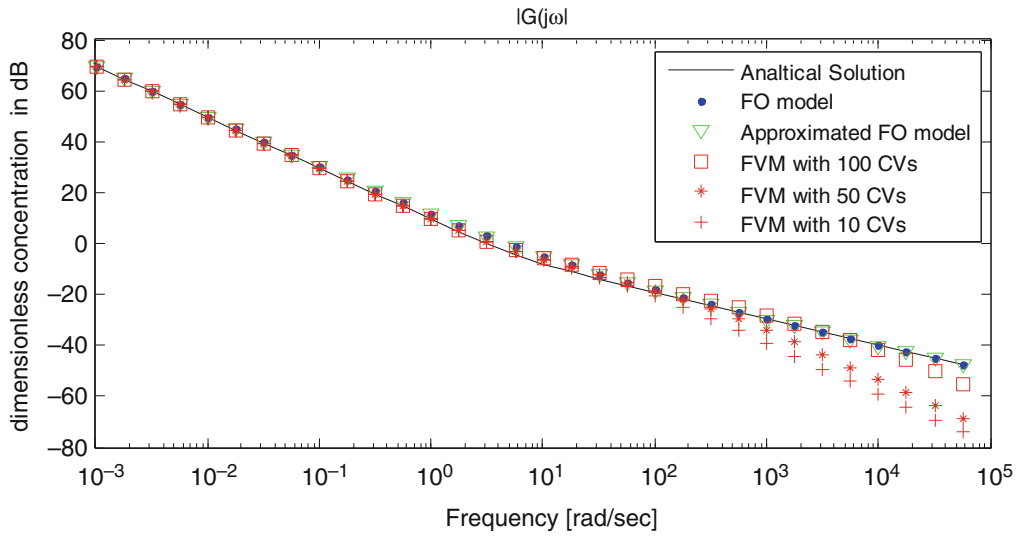


Fig. 3 Demonstrates the bode diagram of analytic system, FO model and it approximation, FVM with 100, 50, and 10CV

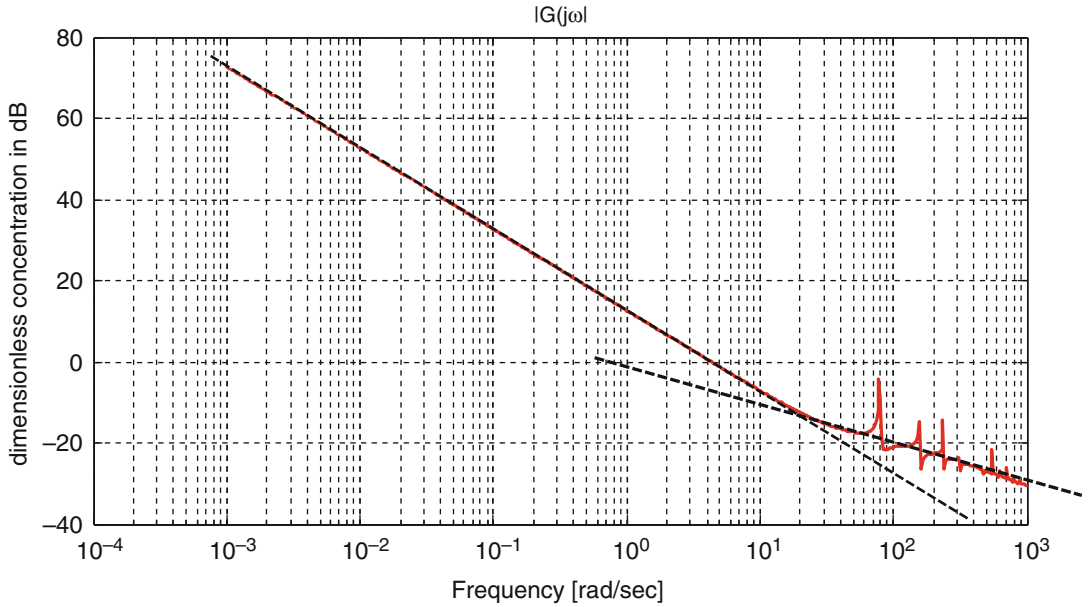


Fig. 4 Estimated Bode diagram of the system. Solid line represents the SPA estimated Bode plot, while the dashed lines represent fitted asymptotic lines

5.1 Fractional Order Selection and Initial Model Parameter Estimation

Fractional model order can be identified based on estimated Bode diagram of the system, where empirical transfer function estimation (ETFE) or spectral analysis (SPA) is a commonly adopted method. Subsequently, asymptotic lines are fitted to the magnitude curve of the estimated Bode diagram [13]. This allows for approximate estimation of the fractional order from slopes of the asymptotic lines and pole location is determined from the intersection of these lines.

The SPA has been applied to estimate the magnitude Bode diagram of the system based on the simulated (measured) data. Fig. 4 shows the SPA estimated Bode plot (solid red line) together with the fitted asymptotic lines (dashed lines). It can be noted, from the Fig. 4, that the first asymptotic line has a slope of -20 dB/decade from which it can be deduced that the system has a single pole in the origin. Afterwards, the slope of asymptotic line increases to -10 dB/decade at frequency of 20 rad/s. Therefore, the system zero is located at approximately -20. Moreover, the increase of slope by only 10 dB/decade indicates a fractional order of 0.5. The initial transfer function becomes

$$\widehat{G}(s) = \frac{\overline{c}(s)}{-\delta(s)} = \frac{\left(s^{0.5} + 20\right)}{s}$$

Table 1 The calculated IAE performance measure together with corresponding frequency ranges for different integer model orders of approximated FO models

Approximated integer order of fractional $s^{0.5}$	IAE	Frequency range in rad/s
n = 10	0.0301	$[10^{-5},10^{7}]$
n = 5	0.0924	$[0.1,14*10^{5}]$
n = 3	0.2149	$[1,15*10^{5}]$

5.2 SRIVCF Estimation

The parameters of the FO model with $s^{0.5}$, initially identified in Sub-Section 5.1, are estimated (refined) by the SRIVCF method. The model is stated as follows

$$\widehat{G}(s) = \frac{\overline{c}(s)}{-\delta(s)} = \frac{\left(b_0 s^{0.5} + b_1\right)}{s}$$

where $\overline{c}(t_k)$ and $-\delta(t_k)$ are the output and input signals, respectively. The input is designed to be a pseudo-random binary sequence with magnitude ranging in interval $(-20, 20)$. The complete set of input and output data contains $(8*10^{6})$ samples.

The estimated model parameters are $b_0 = 0.9637$ and $b_1 = 1.3932$. Table 1 illustrates a mean integrated absolute error between FVM simulated and SRIVCF estimated $\overline{c}(t_k)$, abbreviated IAE, for different order of approximated FO models. Table 1 shows that the smaller the order the less

accurate results are obtained. However, there is no further requirement to vary (or re-estimate) the model parameters. In overall, this leads to relatively simple model order reduction technique.

6 Conclusion

It has been shown that the identified fractional order continuous-time transfer function model provides an accurate approximation of the diffusion process which occurs within the lithium ion batteries. The parameters of the identified fractional order model have been estimated using the extended version of the simplified refined instrumental variable method, which uses sampled (measured) input-output signals. The identified fractional order model can be subsequently used in a model based control designs and control oriented system analysis, where notions such as transfer-function form and phase lag are standard notions and tools of control engineering. Finally, it has been possible to obtain a reduced order proper integer order linear model directly from the fractional order operator $s^{-\alpha n}$ based on constraints imposed on the system, e.g. the demanded frequency range of operation.

For further work, the nonlinear phenomena of diffusion equations can be more accurately and perhaps more conveniently approximated by a special class of nonlinear fractional-order models. The extension of the simplified refined instrumental variable method to accommodate for such a classes of nonlinear fractional-order models is a part of the current and on-going research.

References

1. Thomas, K., Newman, J., Darling, R.: Advances in lithium-ion batteries, 345 (2002)
2. Doyle, M., Fuller, T.F.J., Newman, J.: Modeling of Galvanostatic Charge and Discharge of the Lithium/Polymer/Insertion Cell, J. of The Electrochemical Society, 140(6), 1526–1533 (1993)
3. Newman, J.: Electrochemical Systems, 2nd ed., Prentice-Hall, Englewood Cliffs, New Jersey, (1991)
4. Nagarajan, G.S., Van Zee, J.W., Spotnitz, R.M.: A Mathematical Model for Intercalation Electrode Behavior I. Effect of Particle-Size Distribution on Discharge Capacity, J. Electrochem. Soc. 145 (3), 771–779 (1998)
5. Doyle, M.: Design and simulation of Lithium Rechargeable Batteries. Ph.D. Thesis, University of California, Berkeley (1995)
6. Botte, G.G., Subramanian, V.R., White, R.E.: Mathematical modeling of secondary lithium batteries. 45 (15–16) 2595–2609 (2000)
7. Bernardi, D., Pawlikowski, E., Newman, J.: A General Energy Balance for Battery Systems, Electrochem. Soc. 32(1), 5-12 (1985)
8. Das, S.: Functional Fractional Calculus for System Identification and Controls. Springer-Verlag, Springer-Verlag, Heidelberg (2009)
9. Randles, J.E.B.: Kinetics of rapid electrode reactions, Disc. Faraday Soc, 11–19 (1947)
10. Young, P.C., Jakeman, A.J.: Refined instrumental variable methods of time-series analysis: Part III, extensions. Int. J. of Control, 31, 741–764 (1980)
11. Malti, R., Victor, S., Oustaloup, A., Garnier , H.: An optimal instrumental variable method for continuous time fractional model identification. Proc. of the 17th IFAC World Congress, pp. 14379-14384 (2008)
12. Allafi, W., Burnham, K.J: Identification of Fractional-Order Continuous-Time Hybrid Box-Jenkins Models Using Refined Instrumental Variable Continuous-Time Fractional-Order Method. Advances in Intelligent Systems and Computing 240, 785–794 (2014)
13. Ghanbari, M., Haeri, M.: Order and pole locator estimation in fractional order systems using bode diagram. Signal Process. 91 (2), 191–202 (2011)
14. Jacobsen, T., West, K., Diffusion impedance in planar, cylindrical and spherical symmetry, Electrochem. Acta., 40(2), 255–262 (1995)
15. Oustaloup, A., Levron, A., Mathieu, N.M.: Frequency-band complex noninteger differentiator: characterization and synthesis, IEEE Trans., Circuits Syst., I Fundam. Theory, 25–39, 47(1) (2000)
16. Aizad, T., Sumislawska, M., Maganga, O., Agbaje, O., Phillip, N., Burnham, K.J.: Investigation of Model Order Reduction Techniques: A Supercapacitor Case Study, Advances in Intelligent Systems and Computing 240, 795–804 (2014)
17. Gu, W.B., Wang, C.Y., Weidner, J.W, Jungstc, R.G., Nagasubramanianc, G.: Computational Fluid Dynamics Modeling of a Lithium/Thionyl Chloride Battery with Electrolyte Flow, J. Electrochem. Soc., 147(2), 427–434 (2000)

Robust Adaptive Control of the Dynamic Multilinked Object: Control of Robot Manipulator

Yuliya Lezhnina, Galina Ternovaya, and Viktoriya Zaripova

1 Introduction

Currently, there is one of the most intensive processes of development resources and methods for construction of the automated system for controlling workflows and productions. Almost all aspects of control systems have been revised. These are the structure and organization of hardware, the distribution of functions between different hardware, the algorithms of the realization of the separate functions, the role of the mathematical models in the management process, the forms and content of the interaction of men and technology. The large number of the interconnected subsystems influencing the each other complicates the traditional control problems, and the requirement of the decentralization comes to the fore [1]. The utilization of the decentralized algorithms corresponds to the nature of large interconnected systems, because it expects the distribution of the system components in space. And in addition to this, the decentralized management structure allows getting more qualitative and trusted control system, because it brings the governing body to the object and simplifies the system structure considerably [2, 3].

This paper focuses on the problem to stabilize and control the reference model for the dynamic multilinked object. The management bases on the measuring of estimates of the derivatives of the output signal with the help of the observers described in [4–6]. Research is also being carried out in the aim of robust adaptive control of the dynamic multilinked object without measuring of the derivatives of regulated variable. This scheme in contrast to different works doesn't require the realization of the filter condition for formation of regression vector, which considerably reduces the order of the closed system.

The most striking example of multivariable systems which has a strong enough influence of interconnections is manipulators of industrial robots [7]. We must note that mathematical formulation of the dynamics of manipulators is highly nonlinear, and it cannot be used without linearization for many algorithms. Another problem with the complexity of robot controlling is that when we add to the consideration of the equation of the engine, the relative degree of the closed system will have the result greater than one.

In this paper we propose to use the algorithms of robust and robust adaptive control for controlling an anthropomorphic robot as to ensure the movement of the end of manipulator on a predetermined path with a prescribed accuracy. The control signal must be formed for each axis without to use the value of speed and acceleration of its relative coordinates or measurable output variables of other axis. Decentralized control thus is realized by the manipulator when this process will exit.

2 The Robust Adaptive Control the Output of the Nonlinear Multilinked Object

There is an interconnected system in which the dynamic processes in local subsystems are described by equations

$$Q_i(P)y_i(t) = R_i(P)\left(u_i(t) + \psi_i(y_i) + f_i(t) + \sum_{i=1,\, i\neq j}^{k} y_{sij}(t)\right), \quad i = \overline{1,k}, \tag{1}$$

$$Q_{sij}(P)y_{sij} = R_{sij}(P)y_j(t), \qquad i \neq j, \quad i = \overline{1,k}. \tag{2}$$

Y. Lezhnina (✉) • V. Zaripova
Department of CAD systems, Astrakhan Civil Engineering Institute, Chita, Russia
e-mail: lejninau@mail.ru; vtempus2@gmail.com

G. Ternovaya
Department of Mathematic, Astrakhan State Technical University, Chita, Russia
e-mail: ternovaja@mail.ru

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_10,

where $Q_i(P)$, $R_i(P)$, $Q_{sij}(P)$, $R_{sij}(P)$ – is the linear differentiation operator whose elements depend on the vector of unknown parameters $\xi \in \Xi$, Ξ – is a known quantity of like value of vector ξ; deg $Q_i = n_i$; deg $R_i = m_i$; deg $Q_{sij} = n_{ij}$; deg $R_{sij} = m_{ij}$; $Q_i(P)$, $R_i(P)$ – is the linear differentiation operator whose elements depend on the vector of unknown parameters $\xi \in \Xi$, Ξ – is a known quantity of like value of vector ξ; $f_i(t)$ – is an unknown bounded perturbation action; $\psi_i(y_i)$ – is an unknown curvilinear function; $u_i(t)$ – is a scalar control action; $y_i(t)$ – is a scalar controlled variable of i -th subsystem which is accessible for measurement.

There the equations (1) describe the dynamic processes in the local subsystems, and (2) in the cross coupling. We had to design a control system that ensures the execution of the target condition without measuring the derivatives of the output signal $y_i(t)$ in the form of $\overline{\lim_{t\to\infty}} |e_i(t)| = \overline{\lim_{t\to\infty}} |y_i(t) - y_{mi}(t)| < \delta$. Demanded quality of transients in subsystems is set by the equations of local reference models in the form of $Q_{mi}(P)y_{mi}(t) = R_{mi}(P)r_i(t), \quad i = \overline{1,k}$. Here, $Q_{mi}(P)$, $R_{mi}(P)$ – are linear differential operators of orders; n_{mi}, m_{mi}, $n_{mi} \le n_i$, $m_{mi} \le m_i$; $r_i(t)$ – are scalar bounded preset signals. Also the use of the measured values of subsystems is not allowed in the other local control subsystems.

Hypothesis. Set Ξ is specified; polynomials $R_i(\lambda)$ are Hurwitz polynomials; orders deg $Q_i = n_i$; deg $R_i = m_i$ are known; relative degree $n_i - m_i > 1$ is known; influences $f_i(y_i, t)$ are bounded $|f_i(y_i, t)| \le C_i, \quad C_i > 0$; matrixes A_{sij}, B_{sij} are Hurwitz matrixes; $n_i - m_i$ derivative of reference models output are bounded, $|r_i(t)| \le const$ $\left|y_{mi}^{l_1}(t)\right| \le const$, $l_1 = \overline{1, n_i - m_i}$; $|\psi_i(y_i)| \le \varphi_i(y_i), \ \varphi_i(y_i) > 0$. Let us make the error equation $e_i(t) = y_i(t) - y_{mi}(t)$, subtracting equations of local reference models from (1)

$$Q_i(P)e_i(t) = R_i(P)\left(u_i(t) + \psi_i(y_i) + f_i(t) + \frac{Q_{mi}(P) - Q_i(P)}{R_i(P)}y_{mi}(t) + \right.$$
$$\left. - \frac{R_{mi}(P)}{R_i(P)}r_i(t) + \sum_{j=1, i\neq j}^{k} Q_{sij}^{-1}(P)R_{sij}(P)\left(e_j(t) + y_{mj}(t)\right)\right). \tag{3}$$

In view of the assumptions a value

$$\vartheta_i(t) = f_i(t) - \frac{R_{mi}(P)}{R_i(P)}r_i(t) + \frac{Q_{mi}(P) - Q_i(P)}{R_i(P)}y_{mi}(t)$$
$$+ \sum_{j=1}^{k} Q_{sij}^{-1}R_{sij}(P)y_{mj}(t)$$

is bounded.

In accordance with the approach presented in [8], let us define local control law in the form of

$$u_i(t) = -\theta_i T_i(P)\overline{e}_i(t) + \mu_i\varphi_i(y_i)sign(e_i),$$
$$\dot{\mu}_i = -\pi_i|e_i| - \gamma_i\mu_i, \quad \mu_i(0) = 0, \tag{4}$$

where μ_i **–** is a adjustable parameter; a number $\theta_i > 0$ and the linear differentiation operator $T_i(P)$ of order $n_i - m_i - 1$ are chosen for reasons of Hurwitz polynomial $Q_{0i}(\lambda) = Q_i(\lambda) + \theta_i R_i(\lambda)T_i(\lambda)$, and the function $\overline{e}_i(t)$ is a estimated error $e_i(t)$. Let us introduce the notation $y_{2si}(t) = Q_{sij}^{-1}R_{sij}(P)$ $e_j(t)$. Then the equation (2) will look like

$$Q_{0i}(P)e_i(t) = R_i(P)\left(\theta_i T_i(P)(e_i - \overline{e}_i) + \vartheta_i(t) + \psi_i(y_i) + \mu_i\varphi_i(y_i)sign(e_i) + \sum_{j=1, i\neq j}^{k} y_{2sij}(t)\right)$$

Implementing the control law (4) requires getting the estimate $\overline{e}_i(t)$ and its $n_i - m_i - 2$ derivatives, for which we will use the observer [9]

$$\dot{\overline{x}}_i = F_{0i}\overline{x}_i + H_i(e_i - \overline{e}_i), \quad \overline{e}_i = L_{0i}\overline{x}_i. \tag{5}$$

Here $\overline{x}_i \in \Re^{n_i - m_i}$; $L_{0i} = [1, 0, \ldots, 0]$; $H_i^T = \left[\frac{h_{1i}}{\chi_i}, \ldots, \frac{h_{n_i-m_i-1}}{\chi_i^{n_i-m_i-1}}\right]$; $F_{0i} = \begin{bmatrix} 0 & I_{n_i-m_i-1} \\ 0 & 0 \end{bmatrix}$; $\chi_i > 0$ – is small number; vector H_i gets out so that the matrix $F_i = F_{0i} + \overline{H}_i L_{0i}$ was Hurwitz, where $\overline{H}_i^T = [-h_{1i}, \ldots, -h_{n_i-m_i-1}]$; $I_{n_i-m_i-1}$ – identity matrix.

It is obvious, that the control law can be technically realizable, as it contains the known or measurable values. At the time, the solution of the problem gives the following statement.

The Statement If the hypothesis A1 is implemented, then the control law (3) with the observer (4) implements the limited nature of system trajectory (1),(2).

But it should be noted, that choosing the number θ_i of greater value, and the value χ of smaller, we can achieve the target condition $\overline{\lim_{t\to\infty}} |e_i(t)| = \overline{\lim_{t\to\infty}} |y_i(t) - y_{mi}(t)| < \delta$. In Fig. 1, there is the block diagram of robust stabilization system.

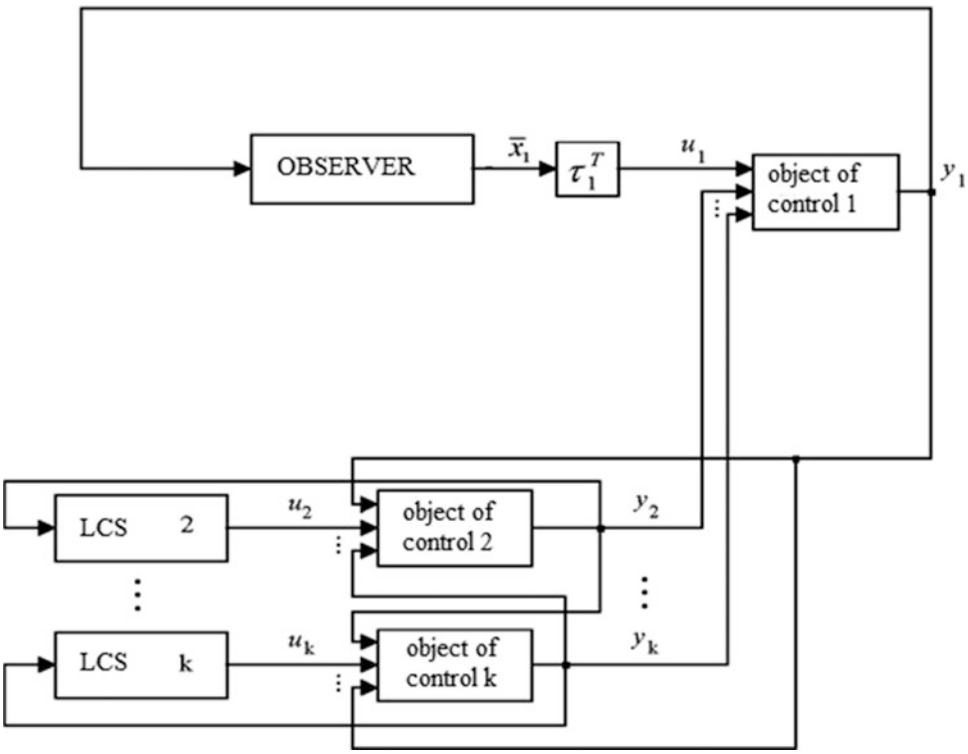


Fig. 1 The Block Diagram of robust Stabilization System

3 Mathematical Model of the Robot

To illustrate the operability of the algorithm, let's consider two degrees of freedom (fig. 2) connected to the rotary joint. Then the structure corresponds with a two-tier flat manipulator, which will allow to take into account the interactions, existing between the selected kinematic pairs. The manipulator system consists of a mechanical part of the system and drive, which provide the work of certain degrees of mobility of operation of a mechanism. Each of the degrees of mobility of manipulator is provided with separate actuator.

Using Lagrange's equations of type II, we represent dynamic equations of motion in kinematic pairs of robot system by means of nonlinear differential equations

$$\tau(t) = D(\theta)\ddot{\theta}(t) + h\left(\theta, \dot{\theta}\right) + c(\theta) \tag{6}$$

where $\tau(t)$ –n-dimensional vector of equivalent torque, produced by drives; $\theta(t) = col(q_1, \ldots, q_n)$ –n-dimensional vector of joint variables of the manipulator; $\dot{\theta}(t) = col(\dot{q}_1, \ldots, \dot{q}_n)$ –n-dimensional vector of manipulator rates; $\ddot{\theta}(t) = col(\ddot{q}_1, \ldots, \ddot{q}_n)$ –n-dimensional vector of manipulator acceleration; $D(\theta)$ –symmetric inertia matrix;

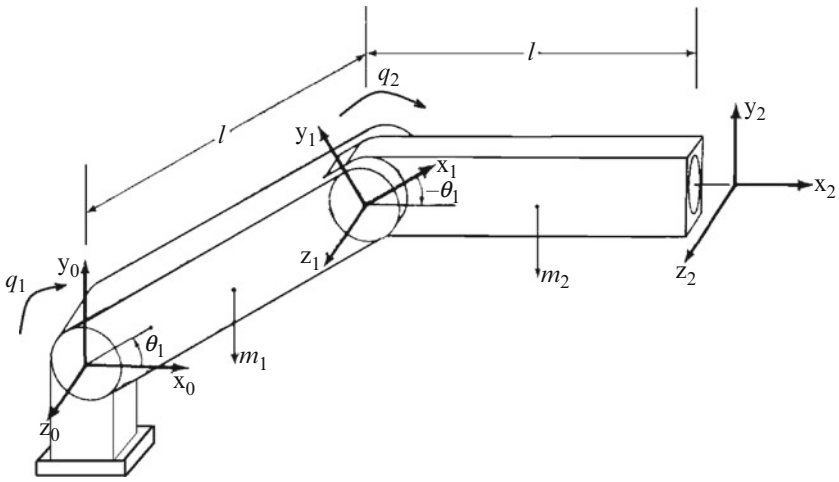


Fig. 2 Two-tier flat manipulator with rotary joints

$h(\theta, \dot{\theta})$ –vector of Coriolis force and centrifugal forces; $c(\theta)$ – vector of gravity forces.

To describe the rotational relationships between adjacent links representation of Denavit-Hartenberg is used [10]. For the considered particular case of a robot the following numeric parameters of tiers are chosen: $m_1 = 15, 91$ kg, $m_2 = 11, 36$ kg, $l = l_1 = l_2 = 0, 432$ m. Such a choice has defined their correspondence with the tiers 2 and 3 of the manipulator of robot Puma 560 of brand Unimation. The rim variables are angles q_1, q_2. Tier parameters possess the values $\alpha_1 = \alpha_2 = 0\ d_1 = d_2 = 0\ a_1 = a_2 = l$

To close the system of equations (6) it is necessary to obtain the expression for the generalized moments in the joints. Drives with DC motors are considered, whose mathematical model which is set as a system of differential equations with constant coefficients [11, 12]

$$\begin{cases} -C_{Mi}I_i + J_{ri}\ddot{q}_i = -\tau_i \\ L_{ri}\dot{I}_i + R_{ri}I_i + C_{Ei}\dot{q}_i = u_i, \end{cases}$$

where I_i – motor rotor current (A); u_i – voltage motor armature (V); B_{ci} – internal friction (H · м · sec/rad); L_{ri} – rotor inductance (henry); J_{ri} – rotor inertia (kg · m^2) ; R_{ri} – resistance of rotor winding (Ohm); C_{Mi} – coefficient of moment proportion (H · m/A); C_{Ei} – coefficient of proportion of electromotive force (B · sec/rad). Here u_i – control actions, which present the voltage of the DC motor, on whose value clipping is laid. Mechanical parameters of the robot are assumed to be variable and unknown, and their changes are considered to be relatively fast. Options of the drives are slow to change, and this change can be neglected. Thus, it is possible to synthesize control, suggesting that the models of drives are predetermined and unchangeable [13]. Then vector equations are valid for the manipulator [14]

$$\begin{cases} D(\theta)\ddot{\theta}(t) + h(\theta, \dot{\theta}) + c(\theta) = E_M I \\ E_{rI}\dot{I} + E_\omega \dot{\theta} + E_I I = u, \end{cases}$$

where E_M, E_{rI}, E_ω, E_I – diagonal $(n \times n)$ -matrixes with elements C_{Mi}, L_{ri}, C_{Ei} and R_{ri} correspondingly; $I = col(I_1, \ldots, I_n)$.

$$D(\theta) = \begin{bmatrix} \frac{1}{3}m_1l^2 + \frac{4}{3}m_2l^2 + m_2l^2C_2 + J_{r1} & \frac{1}{3}m_2l^2 + \frac{1}{2}m_2l^2C_2 \\ \frac{1}{3}m_2l^2 + \frac{1}{2}m_2l^2C_2 & \frac{1}{3}m_2l^2 + J_{r2} \end{bmatrix},$$

$$h(\theta, \dot{\theta}) = \begin{bmatrix} -\left(\frac{1}{2}m_2l^2S_2\dot{q}_2^2 - m_2l^2S_2\dot{q}_1\dot{q}_2\right) \\ \frac{1}{2}m_2l^2S_2\dot{q}_1^2 \end{bmatrix},$$

$$c(\theta) = \begin{bmatrix} \frac{1}{2}m_1glC_1 + \frac{1}{2}m_2glC_{12} + m_2glC_1 \\ \frac{1}{2}m_2glC_{12} \end{bmatrix}$$

where m_i and $l_i = l$ – mass i-th unit, $S_i = \sin(q_i)$, $C_i = \cos(q_i)$, $C_{ij} = \cos(q_i + q_j)$, $i = 1, 2$. Taking into account the equations of the engine, we will obtain a model of a closed system in the space of states, which will be used to design the control system.

$$\dot{x}_i = A_ix_i + B_iu_i + G_ih_i + G_ic_i,$$
$$y_i = L_ix_i,\ \ i = \overline{1,n},$$

where state vector $x_i^T = [q_i, \dot{q}_i, I_i]^T$, and matrixes A_i, B_i, G_i, L_i are presented as

$$A_i = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & C_{Mi}/d_{ii} \\ 0 & C_{Ei}/L_{ri} & -R_{ri}/L_{ri} \end{bmatrix} B_i = \begin{bmatrix} 0 \\ 0 \\ 1/L_{ri} \end{bmatrix},$$
$$G_i = \begin{bmatrix} 0 \\ -1/d_{ii} \\ 0 \end{bmatrix} L_i = [1\ \ 0\ \ 0]$$

3.1 Control System of the Robot Manopulator at Movement Along any Trajectory

We will consider the following problem – it is necessary to take across the manipulator from the initial point to the final when tracking by angles of rotation in kinematic couples of set trajectories. Change of homogeneous coordinates on some interval of a trajectory it is possible to describe look function "step + exponent"

$$\theta_{mi}(t) = d_{0i} + \left(1 - \sum_{j=1}^{k_{ji}} \exp\left(\frac{-t}{\tilde{t}_{ji}}\right)\right) d_{1i} \qquad (7)$$

where d_0, $(d_0 + d_1)$ – initial and final position; k_j – number of exponents, $\tilde{t}_j$ – time constant of i-th local subsystem. It is necessary to design the robust adaptive control to move the manipulator lengthways of set trajectories. Thus for control

law it isn't allowed to use of derivative of the control signals and output variable other local subsystems. To solve this problem we use the decentralized law of robust adaptive control, which was proposed in this paper. Since the relative degree of the selected object is $\gamma_i = 2$, then choose a $Q_{mi}(P) = P + 1$ is Hurwitz polynomial for $i = 1, 2, \pi = \gamma = 0.1$ is coefficients. When tracing control of select configuration of the robot arm using coefficients $\mu = 0, 01, \theta = 20$ and observer (5) with $\chi = 0, 01$.

Synthesized control law provides the manipulator movement from the initial configuration $\{q_{01} = -90^\circ,\ q_{02} = 0^\circ, \dot{q}_{01} = 0, \dot{q}_{02} = 0\}$ to the final position $\{q_{k1} = 0^\circ, q_{k2} = 90^\circ, \dot{q}_{k1} = 0, \dot{q}_{k2} = 0\}$. The rotation angles in the kinematic pairs to move lengthways the set of trajectories

$$q_{m1} = \frac{\pi}{2}\left(3e^{-t/0.3} - 4e^{-t/0.4}\right),$$
$$q_{m2} = \frac{\pi}{2}\left(1 + 3e^{-t/0.3} - 4e^{-t/0.4}\right)$$

These exponential laws provide continuity of change of speed and limited acceleration in fast movements of the robot.

Figure 3 shows the trajectory of homogeneous coordinates of units and the reference trajectory of the joints.

Fig. 4 shows errors trajectories. We have smooth transient process. Satisfactory indicators of quality of control process are received. Regulation time of makes $t_p = 2, 2$ if the admissible error makes $\varepsilon \leq 0, 01$ mm. If the error $\varepsilon \leq 0, 02$ mm in admissible, the control law copes with a problem of tracking the chosen trajectory at once.

Apparently, from results of the modeling, the offered control law provides high quality of transients in system. We will note that when modeling amplitude restriction $|u| \leq 85$ was imposed on a control signal of management as at technical realization signal level always is defined by a regulator design. However, even under the imposed restriction, an output signal rather precisely traces reference value.

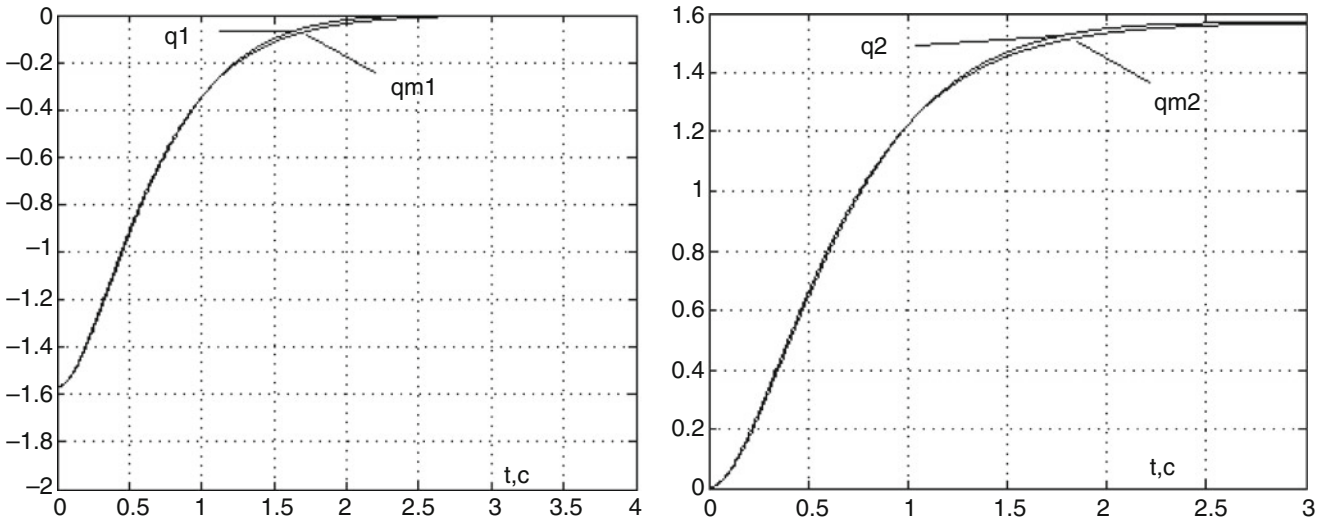


Fig. 3 Output trajectory

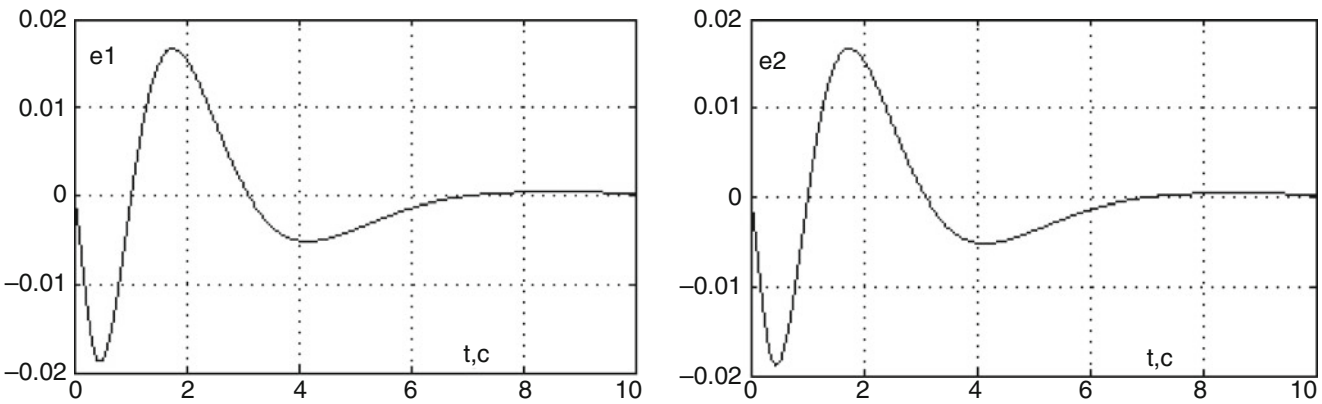


Fig. 4 Error trajectory

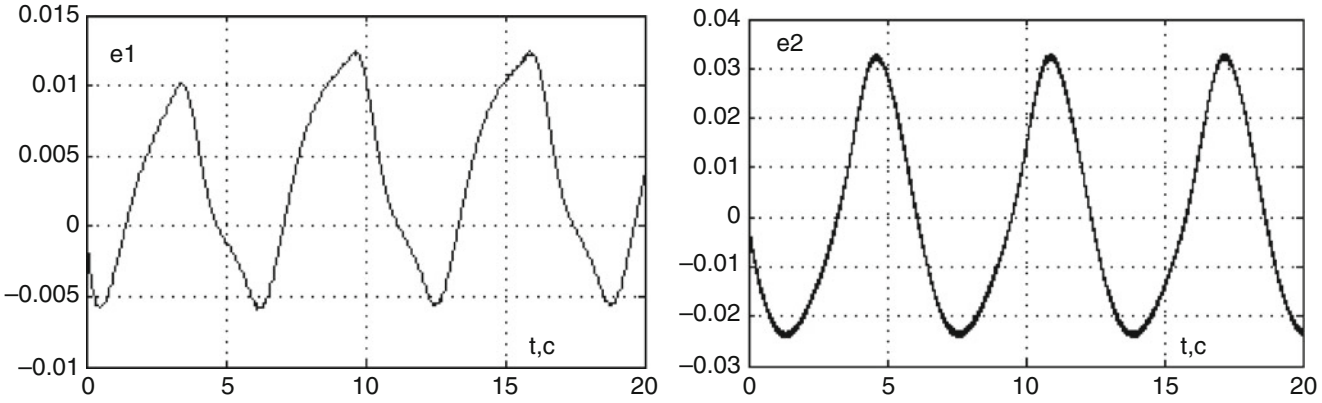


Fig. 5. Error trajectory in movement by circle

3.2 Control System of Robotic Manipulators Motion Along a Closed Path

Consider the solution of more complex problem [15, 16]. It is necessary to synthesize a control algorithm that carries out the movement of the end of the manipulator along a closed path. For this purpose at the first stage of synthesis the problem of the formation of program path is solved $\theta_m(t) = [q_{1np},\ q_{2np}]^T$, providing the movement of the end point of the manipulator along a circle. In this case, there is a clear link between the two angular degrees of freedom of the manipulator and the coordinates of the end point of the manipulator in the Cartesian coordinate system. Then, according to prescribed trajectory uniquely, we define the variation of angles and displacements of the manipulator. If the initial values of the homogeneous coordinates assumed equal $q_1 = 90^\circ,\ \ q_2 = 0^\circ$, then $x_0 = l - a,\ \ y_0 = l$. Obtain expressions for the signals α and β [17].

$$\alpha = arctg\frac{l + a\sin t}{l - a + a\cos t},\ \beta = \arccos\frac{l - a + a\cos t}{2l\cos t}$$

After forming the signals α and β, for program paths we will obtain the equations

$$q_{1m} = \alpha + \beta,\quad q_{2m} = \alpha - \beta$$

Control law is given in the form (4) with the observer (5). Fig. 5 shows the errors trajectories resulting from the simulation with the following parameters of the control law

$$Q_{mi}(P) = P + 1,\ i = 1,2,\ \pi = \gamma = 0.1,\ \mu = 0.01,$$
$$\theta = 20, |u2_1(t)| \le 2, |u2_2(t)| \le 8$$

$$\dot{\zeta}_1 = \zeta_1 + \frac{1}{\chi}(e_1(t) - \overline{e}_1), \dot{\zeta}_2 = \zeta_2 + \frac{6}{\chi}(e_2(t) - \overline{e}_2),$$
$$\chi = 0,01\ \overline{e}_i = \zeta_i,\quad i = 1,2.$$

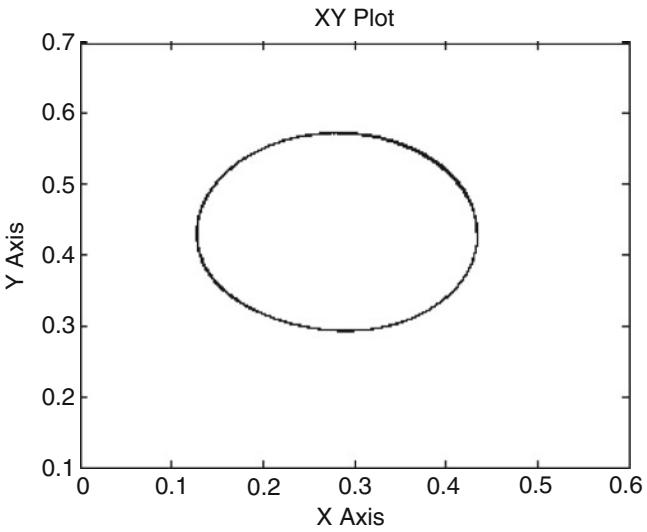


Fig. 6 Trajectory of manipulator end in movement by circle

These errors values are yielded the trajectory of the end of the manipulator shown in Fig. 6

4 Resume

It is proposed to use the observer with a high degree of stability for the synthesis of robust decentralized systems of stabilization by the technological multilinked object with a scalar input-output.

We've obtained the relations of robust decentralized control with the reference model describing the guaranteed estimate of the small parameter and the transient performance.

We've considered the problem of robust adaptive control for the nonlinear multilinked object with a scalar input-output. We've obtained and substantiated structure of adaptive control, ensuring the realization of control goal with an uncertain reference model.

We've confirmed and investigated the efficiency of synthesized algorithms of robust adaptive control by computational modeling in Simulink MatLab.

References

1. E. Freund, Path Control for a Redundant Type of Industrial Robot, Proc. Of VII International Sump. On Industrial Robots. (1977) 234-241, 107-114.
2. A. M. Tsykunov, Robust control algorithms with compensation of bounded perturbations, Automation and Remote Control. 7 (2007), 103–115.
3. V.O. Nikiforov, Robust high-order tuner of simplified structure, Automatica. 8(35) (1999) 1409–1415.
4. Tsykunov, A. M. Robust output control of linear dynamic objects, Mekhatronika, avtomatizatsiya, upravlenie, N.8, 2008, pp. 7–12.
5. Parsheva, E. A., Lezhnina, Yu. A. Robust decentralized control with disturbance compensation with nonlinear multivariable objects, Mekhatronica, avtomatizatsiya, upravlenie. N.6, 2011 pp. 2–7.
6. Brusin, V. A. A class of singular distributed adaptive systems, Autom. Telemekh., N.4, 1995, pp. 119–127.
7. R. P. Paul, Robot Manipulators–Mathematical programming and control, M.I.T. Press, Cambridge,1983.
8. Tsykunov, A. M. Robust control of linear singularly perturbed plants, Automation and Remote Control. V.8, N.72, 2011, pp. 1776–1789.
9. Atassi, A. N., Khalil, H. H. A separation principle for the stabilization of a class of nonlinear systems, IEEE Trans. on Automatic Control, V.9, N.44, 1999, pp. 1672–1687.
10. Mohsen Shahinpoor, A Robot Engineering Text Book, Harper and Row, 1987.
11. Belousov, I. R. Formation of dynamic equations of robot manipulators, Preprint of cervical cancer named after, M.V. Keldysh RAS, N.45, 2002.
12. Bobtsov, A. A. Robust control algorithm of uncertain object without measuring derivatives of adjusted variable, Autom. Telemekh. N.8, 2003, pp. 82–96.
13. Furtat, I. B., Fradkov, A. L., Tsykunov, A. M. Robust Synchronization of Linear Networks with Compensation of Disturbances, Proceedings of 18th IFAC World Congress, V.18, N.1, 2011, pp. 1255–1260.
14. Tao, G., Ioannou, P. A. Model reference adaptive control for plants with unknown relative degree, IEEE Trans. Automat. Control. V.6, N.38, 1993, pp. 976–982.
15. Parsheva, E. A. Robust Decentralized Control with Scalar Output of Multivariable Structurally Uncertain Plants with State Delay, Proceedings of the 18th IFAC World Congress, V.18, N.1, 2011, pp. 11489–11494.
16. Parsheva, E. A. Robust Decentralized Control with Scalar Output of Multivariable Plants with Uncertain Structures, 6th IFAC Symposium on Robust Control Design, V.6, N.1, 2009, pp. 54–59.
17. Lezhnina Yu.A., Ternovaya G.N., Galyauv E.R., Kvyatkovskaya I. Yu.: Decentralized robust control over robot manipulator. Applied Mechanics and Materials. 437, pp. 605–609 (2013)

Modeling and Identification of a Fractional-Order Discrete-Time Laguerre-Hammerstein System

Rafal Stanisławski, Marcin Gałek, Krzysztof J. Latawiec, and Marian Łukaniszyn

1 Introduction

Nonlinear block-oriented systems, including the Hammerstein, Wiener and feedback-nonlinear ones have attracted considerable research interest both from the industrial and academic environments [1, 2, 3, 4]. On the other hand, it is well known that orthonormal basis functions (OBF) have proved to be useful in identification and control of dynamical systems, including nonlinear block-oriented systems [5, 6, 7, 8]. In particular, an inverse OBF (IOBF) modeling approach has been effective in identification of a linear dynamic part of the Hammerstein system [5]. The approach provides the so-called separability in estimation of linear and nonlinear submodels [6], thus eliminating the bilinearity issue detrimentally affecting e.g. the ARX-based modeling schemes. The IOBF modeling approach is continued to be efficiently used here to model a linear fractional-order dynamic part of the Hammerstein system.

Recently, fractional-order dynamics have been given a huge research interest, mostly for linear systems [9, 10, 11, 12, 13, 14, 15, 16, 17].

Discrete-time fractional-order OBF-based modeling is a new research area and there is a few papers on the topic that has up to date been available [18, 19, 20, 21, 22]. Those papers illustrate that fractional-order discrete Laguerre filters can be very effective in modeling of dynamical systems.

A fractional-order Hammerstein system has been elegantly analyzed and identified in Ref. [23]. However, a computational burden of the approach is very high, in fact prohibitively high in adaptive estimation and control.

This paper presents a new, simple strategy for Hammerstein system identification, which is a combination of the inverse-OBF modeling concept and fractional-order generalization of discrete Laguerre filters. The effective combination gives rise to the introduction of a powerful method for identification of the fractional-order Hammerstein system.

2 Fractional-Order Discrete-Time Difference

A simple generalization of the familiar Grünwald-Letnikov difference [11] is the fractional difference (FD) in discrete time t, described by equation [9, 13, 14, 24]

$$\Delta^{\alpha}x(t) = \sum_{j=0}^{t} P_j(\alpha)x(t)q^{-j} = x(t) + \sum_{j=1}^{t} P_j(\alpha)x(t)q^{-j} \quad t = 0, 1, \ldots \tag{1}$$

where $\alpha \in (0, 2)$ is the fractional order, q^{-1} is the backward shift operator and

$$P_j(\alpha) = (-1)^j \gamma_j(\alpha) \tag{2}$$

with

$$\gamma_j(\alpha) = \binom{\alpha}{j} = \begin{cases} 1 & j = 0 \\ \dfrac{\alpha(\alpha-1)\ldots(\alpha-j+1)}{j!} & j > 0 \end{cases} \tag{3}$$

Note that each element in Eqn. (1) from time t back to 0 is nonzero so that each incoming sample of the signal $x(t)$ increases the complication of the model equation. In the limit, with $t \to +\infty$, we end up with computational explosion. Therefore in [25], truncated or finite fractional difference (FFD) has been considered for practical,

R. Stanisławski (✉) • M. Gałek • K.J. Latawiec • M. Łukaniszyn
The authors are with the Department fo Electrical, Control and Computer Engineering, Opole University of Technology, ul. Sosnkowskiego 31, 45-272 Opole, Poland
e-mail: r.stanislawski@po.opole.pl; k.latawiec@po.opole.pl; m.lukaniszyn@po.opole.pl

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_11,

feasibility reasons. Finite fractional difference (FFD) is defined as

$$\Delta^{\alpha}x(t,J) = x(t) + \sum_{j=1}^{J} P_j(\alpha)x(t)q^{-j} \tag{4}$$

where $J = \min(t,\overline{J})$ and $\overline{J}$ is the upper bound for j when $t > \overline{J}$.

In this paper, we assume that α is known.

Remark 1. Possible accounting for the sampling period T when transferring from the Grünwald-Letnikov continuous-time derivative to the Grünwald-Letnikov discrete-time difference results in dividing the right-hand side of Eqns. (1) and (4) by T^{α}. Operating without T^{α} as in the sequel corresponds to putting $T = 1$ or to the substitution of $P_j(\alpha)$ for $\frac{P_j(\alpha)}{T^{\alpha}}$, $j = 0, \ldots, t$.

3 Fractional-Order Discrete-Time Laguerre Filters

A classical (or integer-order, or "regular") OBF model of a dynamical system, or shortly, OBF system, can be presented in form

$$y(t) = \sum_{i=1}^{K} C_i L_i(q)u(t) + e(t) \tag{5}$$

where $u(t)$ and $y(t)$ are the system input and output, respectively, $L_i(z)$ and C_i, $i = 1, \ldots, K$, are orthonormal transfer functions and weighting parameters, respectively and $e(t)$ is the output error. In case of use of discrete Laguerre filters we have

$$L_i(z) = \frac{k}{z-P}\left(\frac{-Pz+1}{z-P}\right)^{i-1} \quad i = 1, ..., K \tag{6}$$

where $k = \sqrt{1-P^2}$ and P is a dominant pole. In the sequel, we limit our interest to the practically justified case of $P > 0$. The unknown parameters C_i, $i = 1, \ldots, K$, can be easily estimated using e.g. Recursive Least Squares (RLS) or Least Mean Squares (LMS) algorithms formalized in a linear regression fashion [26]. In our examples, RLS estimation is used. Pursuing an optimal Laguerre pole P_{opt} has been well established [7, 8, 27, 28, 29].

The Laguerre filters presented in Eqn. (6), can be factorized to the form [25, 20, 21]

$$L_i(q^{-1}) = G_L(q^{-1})\big(G_R(q^{-1}) - P\big)^{i-1} \quad i = 1, ..., K \tag{7}$$

with

$$G_L(q^{-1}) = \frac{kq^{-1}}{1-Pq^{-1}} \tag{8}$$

$$G_R(q^{-1}) = \frac{k^2q^{-1}}{1-Pq^{-1}} = kG_L(q^{-1}) \tag{9}$$

and the consecutive filter outputs being $y_L(t) = G_L(q^{-1})u(t)$ and $y_R^i(t) = G_R(q^{-1})U_i(t)$, $i = 1, ..., K-1$, with

$$U_i(t) = \begin{cases} y_L(t) & i = 1 \\ y_R^{i-1}(t) - PU_{i-1}(t) & i = 2, ..., K \end{cases} \tag{10}$$

The two filters can also be described as

$$\begin{aligned} &G_L^f: \\ &\Delta y_L(t) = (P-1)y_L(t)q^{-1} + ku(t)q^{-1} \end{aligned} \tag{11}$$

$$\begin{aligned} &G_R^f: \\ &\Delta y_R^i(t) = (P-1)y_R^i(t)q^{-1} + k^2U_i(t)q^{-1} \end{aligned} \tag{12}$$

where $\Delta y_L(t) = y_L(t) - y_L(t-1)$ and similar is $\Delta\ y_R{}^i(t)$, $i = 1, \ldots, K$.

The outstanding value of the factorization (7) of the expression (6) is that $G_L(q^{-1})$ and $G_R(q^{-1})$ are the first-order filters that can be easily adopted to the fractional-order form. The fraction-formalized filters $G_L^f(q^{-1})$ and $G_R^f(q^{-1})$ can now be described as

$$\begin{aligned} &G_L^f: \\ &\Delta^{\alpha} y_L(t) = (P-1)y_L(t)q^{-1} + ku(t)q^{-1} \end{aligned} \tag{13}$$

$$\begin{aligned} &G_R^f: \\ &\Delta^{\alpha} y_R^i(t) = (P-1)y_R^i(t)q^{-1} + k^2U_i(t)q^{-1} \end{aligned} \tag{14}$$

where $U_i(t)$ is as in Eqn. (10). Finally, the outputs from the FD versions of the $G_L(q^{-1})$ and $G_R(q^{-1})$ filters can be obtained as

$$\begin{aligned} &G_L^f: \\ &y_L(t) = (P-1)y_L(t)q^{-1} + ku(t)q^{-1} - \sum_{j=1}^{t} P_j(\alpha)y_L(t)q^{-j} \end{aligned} \tag{15}$$

$$\begin{aligned} &G_R^f: \\ &y_R^i(t) = (P-1)y_R^i(t)q^{-1} + k^2U_i(t)q^{-1} - \sum_{j=1}^{t} P_j(\alpha)y_R^i(t)q^{-j} \end{aligned} \tag{16}$$

The outputs for FFD versions of the $G_L(q^{-1})$ and $G_R(q^{-1})$ filters can be calculated as

$$G_L^f:$$
$$y_L(t) = (P-1)y_L(t)q^{-1} + ku(t)q^{-1} - \sum_{j=1}^{J} P_j(\alpha)y_L(t)q^{-j} \tag{17}$$

$$G_R^f:$$
$$y_R^i(t) = (P-1)y_R^i(t)q^{-1} + k^2U_i(t)q^{-1} - \sum_{j=1}^{J} P_j(\alpha)y_R^i(t)q^{-j} \tag{18}$$

Remark 2. Possible accounting for the sampling period T when transferring from the Grünwald-Letnikov continuous-time derivative to the Grünwald-Letnikov discrete-time difference results in multiplication of the two first components at the right-hand sides of Eqns. (17) and (18) by T^{α}.

Finally, the output (5) from the fractional-order Laguerre system is computed as

$$y(t) = \sum_{i=1}^{K} C_i U_i(t) \tag{19}$$

with $U_i(t)$ calculated in Eqn. (10).

4 System description

4.1 Non-fractional case [6, 8, 30]

The Hammerstein system (Fig. 1) consists of two cascaded elements, where the first one is a nonlinear memoryless gain and the second is a linear dynamic model. The whole Hammerstein system can be described by the equation

$$y(t) = G(q)[f(u(t)) + e_H(t)] = G(q)[v(t) + e_H(t)] \tag{20}$$

where $G(q)$ models a dynamic linear part, $f(.)$ describes a nonlinear function, $v(t)$ is the unmeasured output of the nonlinear part and $e_H(t)$ is the error/distur-bance term. An alternative output error/disturbance formulation is also possible. In case of use of the inverse OBF (IOBF) concept to model a linear dynamic part, the Hammerstein model equation can be presented in inverse form [6, 30]

$$\hat{\mathrm{G}}^{-1}(q)\hat{\mathrm{y}}(t) = v(t) \tag{21}$$

or

$$R(q)\hat{\mathrm{y}}(t) = v(t) \tag{22}$$

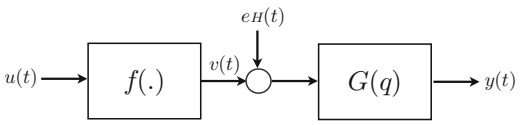


Fig. 1 Hammerstein system

where $R(q)$ is the inverse of the system model $\hat{\mathrm{G}}(q)$. In the IOBF concept, the inverse $R(q)$ of the system is modeled using OBF. An OBF modeling approach can now be applied to equation (22) instead of (21) and finally we can present equation (20) in the following form [6, 30]

$$y(t) + \sum_{i=1}^{M} C_i L_i(q)y(t) = \frac{1}{r_0}v(t-d) + e_1(t) \tag{23}$$

where $e_1(t)$ is the equation error, d is the time delay of the system, and r_0 is the leading coefficient of $R(q)$.

The nonlinear part of the Hammerstein system $f(.)$ can be approximated e.g. with the polynomial expansion

$$v(t) = f(u(t)) = a_1u(t) + a_2u^2(t) + \ldots + a_mu^m(t) \tag{24}$$

with the coefficient a_1 put to 1 without loss of generality [30].

Combining equations (23) and (24) we arrive at the equation describing the model output $\hat{y}(t)$ of the whole Hammerstein system

$$\hat{\mathrm{y}}(t) = -\sum_{i=1}^{M} C_i L_i(q)y(t) + \frac{1}{r_0}\sum_{i=1}^{m} a_i u^i(t-d) \tag{25}$$

with linear and nonlinear submodels separated from each other. Now that the bilinearity effect has been avoided thanks to the separation of the submodels, Eqn. (25) can be easily presented in the linear regression form.

4.2 Fractional-order case

We assume now that a linear dynamics is of fractional order. In order to embed the fractional-order Laguerre filters of Section 3 in the IOBF framework the following important remark is due.

Remark 3. It is essential that the Laguerre filters are, in the IOBF framework, driven by $y(t)$. This means that in order to calculate the fractional-order output equation (20) in the IOBF fashion we have to substitute $y(t)$ for $u(t)$ in Eqns. (13), (15) and (17).

Equation (25) can now be rewritten in form

$$\hat{y}(t) = -\sum_{i=1}^{M} C_i U_i(t) + \frac{1}{r_0}\sum_{i=1}^{m} a_i u^i(t-d) \tag{26}$$

which can be presented in a linear regression form

$$\hat{y}(t) = \varphi^T(t)\Theta \tag{27}$$

where $\Theta^T = [C_1 \quad \ldots \quad C_M \quad \beta_1 \quad \ldots \quad \beta_m]$ and $\varphi^T(t) = [-U_1(t) \quad \ldots \quad -U_M(t) \quad u(t-d) \quad \ldots \quad u^m(t-d)]$ with $\beta_i = a_i/r_0$ and $U_i(t)$, $i = 1, \ldots, M$ driven by $y(t)$ as in Remark 3. Now, the parameters Θ can be easily estimated using e.g. the RLS algorithm (or its adaptive version ALS).

5 Simulation Experiments

Example 1 Consider a discrete fractional-order Hammerstein system, with a static nonlinearity $f(u(t)) = u^3(t)$ and a fractional-order dynamic part described in state-space

$$\Delta^{\alpha} x(t+1) = A_f x(t) + Bu(t), \tag{28}$$

$$y(t) = Cx(t) + Du(t) \tag{29}$$

with

$$A_f = \begin{bmatrix} -0.4 & -0.03 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$
$$C = [0 \quad 0.23], \; D = [0],$$
$$\alpha = 0.5$$

The dynamic part is modeled by an FFD-based fractional-order Laguerre model, with $P = 0.49$, $M = 8$, $m = 3$ and various implementation lengths of the FFD approximation ($\overline{J}$). MSPE is used to evaluate the accuracy of modeling. Selected results are presented in Table 1.

Fig. 2 presents the results of modeling in terms of (indistinguishable) time plots of the actual and modeled outputs of the Hammerstein system for some random input signal.

It can be concluded from Fig. 2 and Table 1 that the introduced fractional-order Laguerre-Hammerstein model can be very effective in modeling of the class of block-oriented nonlinear systems. However, to obtain high modeling accuracies we have to use high implementation lengths of the FFD approximation. This inconvenience can be essentially reduced by making use of our computationally more efficient approximations to FD, that is AFFD,

Table 1 MSPE for Hammerstein system with the FFD-based Laguerre model

$\overline{J}$	50	200	500	1000	5000
MSPE	1.637	0.578	0.182	0.107	8.79e-2

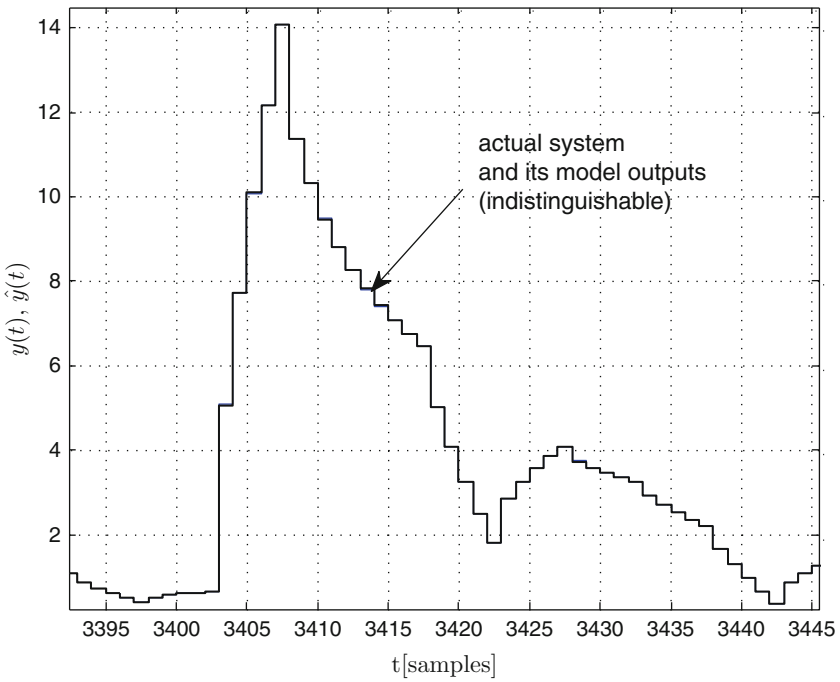


Fig. 2 Time plots of actual and modeled output of the Laguerre-Hammerstein system

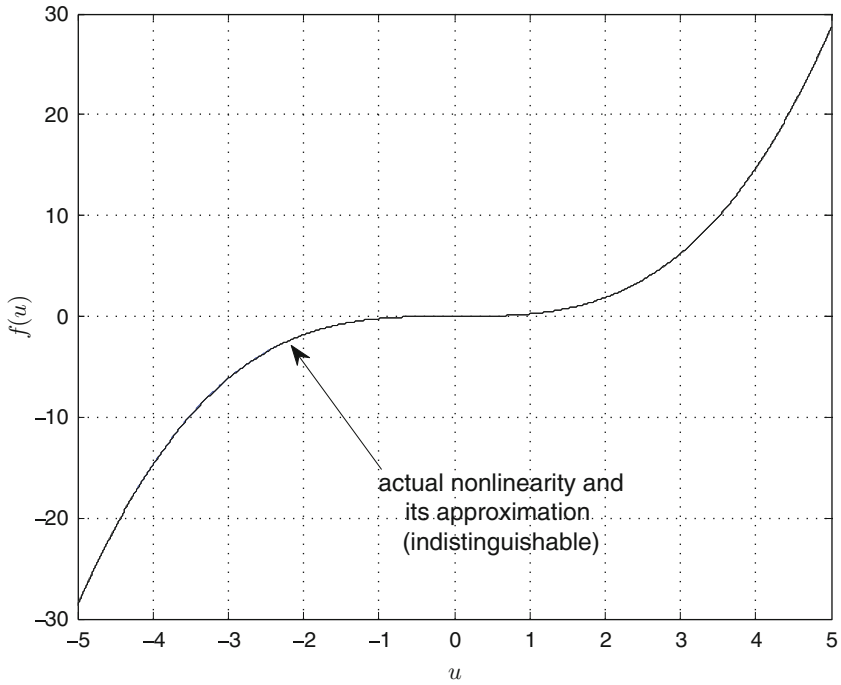


Fig. 3 Nonlinear static characteristic and its model

Table 2 MSPE for Hammerstain system with the FFD-based Laguerre model

$\overline{J}$	50	200	500	1000	5000
MSPE	534.0	533.4	532.9	532.7	532.7

PFFD [31], FLD and, in particular, FFLD [18]. Plots of the actual nonlinear static characteristic and its reconstruction presented in Fig. 3 confirm a very good identification performance. Right the same is with reconstruction of the linear part, in terms of indistinguishable respective impulse responses.

Example 2. Consider the fractional-order nonlinear system as in Example 1, with the zero mean disturbance $e_H(t)$ and $var(e_H(t)) = 0.1$.

Although the MSPEs are now visibly higher, the modeled nonlinear characteristic is, again, indistinguishable from the original one. However, a model of the linear dynamic subsystem is less precisely reconstructed and this is caused by the specific location of the noise $e_H(t)$. Still, the dynamic model accuracy is very good here.

6 Conclusion

The paper has presented a new simple, analytical solution to the nonlinear identification problem for the Hammerstein system using fractional-order Laguerre-based models. We have demonstrated that a combination of the inverse OBF modeling concept and fractional-order Laguerre filters can provide high-perfor-mance identification of fractional-order nonlinear systems. Simulation examples show that in both deterministic and stochastic cases, low prediction errors and accurate reconstructions of both nonlinear and linear parts of the system have been obtained for the introduced models.

References

1. Astrom, K.J., Bell, R.D.: A nonlinear model for steam generation process. In: Preprints IFAC 12th World Congress, Sydney, Australia. (1993)
2. Hasiewicz, Z.: Hammerstein system identification by the haar multiresolution approximation. Int J. Adaptive Control and Signal Processing **74** (1999) 191–217
3. Greblicki, W.: Stochastic approximation in nonparametric identification of Hammerstein system. IEEE Transactions on Automatic Control **47** (2002) 1800–1810
4. Ninness, B., Gibson, S., Weller, S.: Practical aspects of using orthonormal system parameterisations in estimation problems. In: Proc. 12th IFAC Symp. on System Identification (SYSID'2000), Santa Barbara, CA. (2000) 463–468
5. Latawiec, K., Marciak, C., Hunek, W., Stanisławski, R.: A new analytical design methodology for adaptive control of nonlinear block-oriented systems. In: Proceedings the 7th World Multi-Conference on Systemics, Cybernetics and Informatics. Volume 11., Orlando, Florida (2003) 215–220
6. Latawiec, K.J., Marciak, C., Stanislawski, R., Oliveira, G.H.C.: The mode separability principle in modeling of linear and nonlinear block-oriented systems. In: Proc. 10th IEEE MMAR Conference (MMAR'04), Miedzyzdroje, Poland. (2004) 479–484

7. Heuberger, P.S.C., Van den Hof, P.M.J., Wahlberg, B.: Modeling and identification with rational orthogonal basis functions. Springer-Verlag, London, UK (2005)
8. Stanisławski, R.: Hammerstein system identification by means of orthonormal basis functions and radial basis functions. In Pennacchio, S., ed.: Emerging Technologies, Robotics and Control Systems. Internationalsar, Italy (2007) 69–73
9. Kaczorek, T.: Selected Problems of Fractional Systems Theory. Springer-Verlag, Berlin, Germany (2011)
10. Matignon, D.: Stability results for fractional differential equations with applications to control processing. In: Computational Engineering in Systems and Applications Multiconference. Volume 2., Lille, France (1996) 963–968
11. Miller, K., Ross, B.: An Introduction to the fractional calculus and fractional differential equations. Willey, New York, NJ (1993)
12. Monje, C., Chen, Y., Vinagre, B., Xue, D., Feliu, V.: Fractional-order Systems and Controls. Springer-Verlag, London, UK (2010)
13. Ostalczyk, P.: Equivalent descriptions of a discrete-time fractional-order linear system and its stability domains. International Journal of Applied Mathematics and Computer Science **22**(3) (2012) 533–538
14. Oldham, K., Spanier, J.: The fractional calculus. Academic Press, Orlando, FL (1974)
15. Podlubny, I.: Fractional differential equations. Academic Press, Orlando, FL (1999)
16. Stanisławski, R., Latawiec, K.J.: Stability analysis for discrete-time fractional-order LTI state-space systems. Part I: New necessary and sufficient conditions for asymptotic stability. Bulletin of the Polish Academy of Sciences, Technical Sciences **61**(2) (2013) 353–361
17. Stanisławski, R., Latawiec, K.J.: Stability analysis for discrete-time fractional-order LTI state-space systems. Part II: New stability criterion for FD-based systems. Bulletin of the Polish Academy of Sciences, Technical Sciences **61**(2) (2013) 362–370
18. Stanisławski, R.: New Laguerre filter approximators to the Grünwald-Letnikov fractional difference. Mathematical Problems in Engineering **2012** (2012) Paper ID: 732917.
19. Stanisławski, R.: Advances in Modeling of Fractional Difference Systems - New Accuracy, Stability and Computational Results. Opole University of Technology Press, Opole, Poland (2013)
20. Stanisławski, R., Hunek, W.P., Latawiec, K.J.: Normalized finite fractional discrete-time derivative a new concept and its application to OBF modeling. Measurements, Automation and Monitoring **57**(3) (2011) 241–243
21. Stanisławski, R., Latawiec, K.J.: Modeling of open-loop stable linear systems using a combination of a finite fractional derivative and orthonormal basis functions. In: Proceedings of the 15th International Conference on Methods and Models in Automation and Robotics, Miedzyzdroje, Poland. (2010) 411–414
22. Latawiec, K.J., Stanisławski, R., Hunek, W.P., Łukaniszyn, M.: Laguerre-based modeling of fractional-order LTI SISO systems. In: Proceedings of the 18th International Conference on Methods and Models in Automation and Robotics, Miedzyzdroje, Poland. (2013) 64–69
23. Liao, Z., Zhu, Z., Liang, S., Peng, C., Wang, Y.: Subspace identification for fractional order Hammerstein systems based on instrumental variables. International Journal of Control, Automation, and Systems (2012) 947–953
24. Sierociuk, D., Dzieliński, A.: Fractional Kalman filter algorithm for states, parameters and order of fractional system estimation. International Journal of Applied Matchematics and Computer Science **16**(1) (2006) 101–112
25. Stanisławski, R.: Identification of open-loop stable linear systems using fractional orthonormal basis functions. In: Proceedings of the 14th International Conference on Methods and Models in Automation and Robotics, Miedzyzdroje, Poland. (2009) 935–985
26. Ljung, L.: System Identification. Prentice-Hall, Englewood Cliffs (1987)
27. Boukis, C., Mandic, D.P., Constantinides, A.G., Polymenakos, L. C.: A novel algorithm for the adaptation of the pole of Laguerre filters. IEEE Signal Processing Letters **13**(7) (2006) 429–432
28. Oliveira, S.T.: Optimal pole conditions for Laguerre and two-parameter Kautz models: a survey of known results. In: 12th IFAC Symposium on System Identification (SYSID'2000), Santa Barbara, CA (2000) 457–462
29. Stanisławski, R., Hunek, W.P., Latawiec, K.J.: Modeling of non-linear block-oriented systems using orthonormal basis and radial basis functions. Systems Science **35**(2) (2009) 11–18
30. Latawiec, K.J.: The Power of Inverse Systems in Linear and Non-linear Modeling and Control. Opole University of Technology Press, Opole, Poland (2004)
31. Stanisławski, R., Latawiec, K.J.: Normalized finite fractional differences: the computational and accuracy breakthroughs. International Journal of Applied Mathematics and Computer Science **22**(4) (2012) 907–919

A comparative Study of Model-Based and Data-Based Model Order Reduction Techniques for Nonlinear Systems

T. Aizad, O. Maganga, M. Sumislawska, and K.J. Burnham

1 Introduction

New technologies and multi-physical description of subsystems have forced designers to consider nonlinear effects for more accurate modelling leading to increased complexity of mathematical models [1]. Such complex models are non-trival to analyse and to develop control algorithms. Consequently, increasing complexity of circuit designs causes the need for model order reduction (MOR) techniques that are capable of reducing nonlinear models and decreasing computational cost of simulating nonlinear systems. MOR techniques for linear time invariant (LTI) systems are well established [2]. On the other hand MOR for nonlinear systems is an open problem [1]. There are several ways of obtaining reduced order model (ROM) for nonlinear systems via model-based approach such as linear approximation(LA) [3], bilinearisation, proper orthogonal decomposition (POD), quadratic approximation (QA) and trajectory piecewise linear (TPWL) approximation, etc. The LA of a nonlinear system and further reduction of linearised model provides poor approximation in most cases [4]. The bilinearisation method proposed in [5, 6], firstly approximates the original nonlinear system using bilinear model and then reduces the order of the bilinear model by using the Volterra series expansion. The main drawback of this technique is that it is valid only locally around the initial operating point of the nonlinear system. The POD is widely applied to fluid dynamic problems to obtain projection bases, however the cost of evaluating nonlinear operator remains high [7]. The QA typically expands the nonlinear operator about a single state and is only suitable for weakly nonlinear systems since the generated models are only accurate locally [4]. The TPWL approximation is one of the promising approach of obtaining ROM. This approach considers linearisation at multiple operating points along a trained state trajectory, hence gives better performance for strongly nonlinear system. This paper presents two comparative studies of methods suitable for reduction of weakly as well as strongly nonlinear systems. In this study model-based (LA, QA, TPWL approximation) and data-based (system identification based) methods are considered. Nonlinear transmission line model is used to demonstrate applicability of the MOR techniques. Moreover, the accuracy of the ROM as well as computational complexity of simulating the ROM are compared. The performance of MOR techniques is investigated by evaluating two efficacy indices: the coefficient of determination R_T^2, and the number of floating point operations (flops).

T. Aizad (✉) • O. Maganga • M. Sumislawska • K.J. Burnham
Control Theory and Applications Centre Coventry University, CV1 5FB, UK
e-mail: toheed.aizad2@coventry.ac.uk

2 Problem statement

2.1 Model Order Reduction

Model order reduction problem defined as follows. Consider a nonlinear system:

$$\dot{x}(t) = f(x(t)) + Bu(t) \tag{1a}$$

$$y(t) = C^T x(t) \tag{1b}$$

where $x(t) \in \mathbb{R}^N$ is the state vector, $B \in \mathbb{R}^N$ is input matrix, $C \in \mathbb{R}^N$ is an output matrix, whilst $u(t) \in \mathbb{R}^1$ and $y(t) \in \mathbb{R}^1$ are system input and output, respectively. The term $f(x) = [f_1(x), f_2(x), ..., f_N(x)]^T$ is a vector valued function. One can define complexity of the system as a number of its states, denoted N [8]. The task of model order reduction is viewed as finding such a reduced order model

$$\dot{z}(t) = f_r(z(t)) + B_r u(t) \tag{2a}$$

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_12, © Springer International Publishing Switzerland 2015

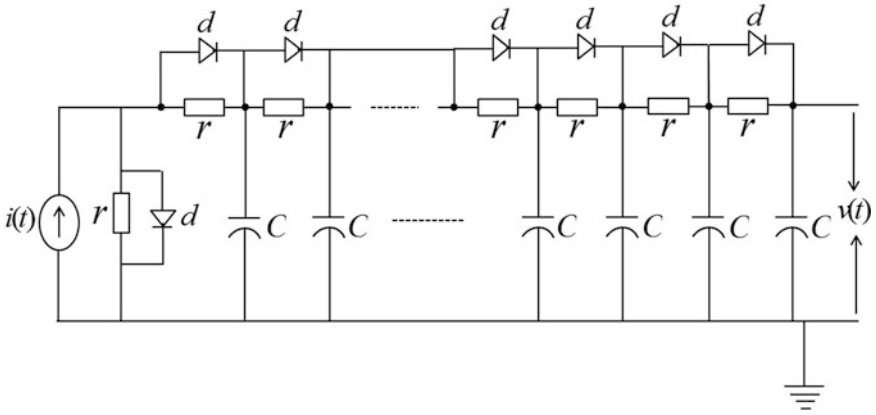


Fig. 1 Schematic diagram of transmission line model

$$y(t) = C_r^T x(t) \tag{2b}$$

such that $z(t) \in \mathbb{R}^q, q \ll N$ and the input-output relation of model (2) mimics the response of (1) [8]. For the sake of clarity, the time index (t) will be omitted in the rest of the paper.

2.2 Considered Case Study

In this paper a nonlinear transmission line model is considered [9, 10, 11, 12], see Fig. 1. The input to the system is the current source $u(t) = i(t)$, whilst output is the terminal voltage $y(t) = v(t)$. The circuit consists of $N + 1$ resistors (r), N capacitors (C) and $N + 1$ diodes (d). It is assumed that all diodes have the same resistance profile and the current passing through each diode, denoted i_d, is a function of the diode voltage

$$i_d(v_d) = e^{(40v_d)} - 1 \tag{3}$$

where v_d is the voltage between diodes terminals. All the resistors and capacitors have unit resistance and capacitance, respectively $(r = 1\Omega, C = 1F)$.

3 Model Order Reduction Methods

3.1 MOR via Quadratic Approximation

The QA method reduces a nonlinear model to a reduced order model of a quadratic nonlinearity. This approach, firstly, approximates the original nonlinear system using second order Taylor expansion and then reduces the order of the quadratic model [4, 11]. The quadratic approximation of system (1) about the origin is given by

$$\dot{x} = Ax + x^T W x + Bu \tag{4a}$$

$$y = C^T x \tag{4b}$$

where A is the Jacobian of $f(x)$ evaluated at the origin and W is a 3 dimension $(N \times N \times N)$ Hessian tensor given as

$$W_{i,j,k} = \frac{\partial^2 f_i}{\partial x_j \partial x_k} \tag{5}$$

Let $A_1 = A^{-1}$, then multiplying (4) by A_1 gives:

$$A_1 \dot{x} = x + A_1 x^T W x + A_1 B u \tag{6a}$$

$$y = C^T x \tag{6b}$$

The state x is projected onto a q-dimensional $(q << N)$ subspace $span\{V\}$, where $V \in \mathbb{R}^{N \times q}$ is an orthonormal matrix spanning the Krylov subspace $span\{[A_1 B, A_1^2 B, ..., A_1^q B]\}$. One can denote the projection of x onto the subspace $span\{V\}$ as:

$$z = V^T x \tag{7}$$

Recall equation (7). The motion equation of the projection of x onto span $\{V\}$, i.e $VV^T x = Vz$, is given by [4]

$$AV\dot{z} = Vz + Az^T V^T W V z + b_1 u \tag{8a}$$

$$y = C^T V z \tag{8b}$$

where $b_1 = A_1 B$.

Using $H = V^T A_1 V$ and multiplying both sides of (8a) by V^T one obtains:

$$H\dot{z} = z + V^T Az^T(t) V^T W V z + V^T b_1 u \tag{9a}$$

$$y = C^T V z \tag{9b}$$

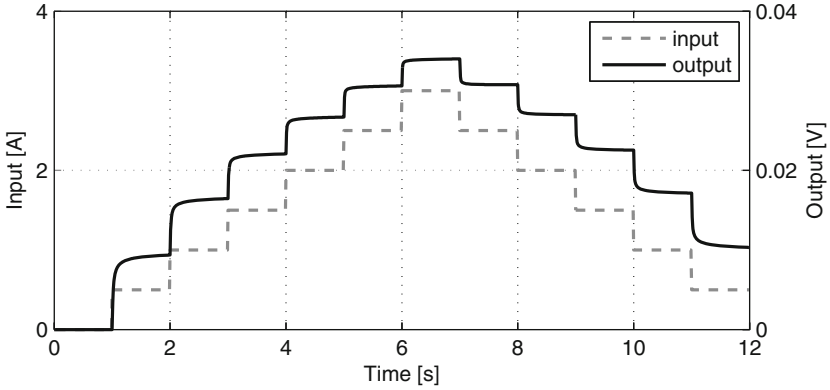


Fig. 2 Identification data set

Multiplying (9) by H^{-1} yields to:

$$\dot{z} = H^{-1}z + H^{-1}V^T A z^T V^T W V z + H^{-1}V^T b_1 u \tag{10a}$$

$$y = C^T V z \tag{10b}$$

Thus, the reduced order representation of (4) is given by:

$$\dot{z} = A_r z + z^T W_r z + B_r u \tag{11a}$$

$$y = C_r^T z \tag{11b}$$

where

$$A_r = H^{-1}, B_r = H^{-1}V^T b_1, C_r = V^T C \tag{12}$$

The term $H^{-1}V^T A_1 z^T V^T W V z$ can be written as $z^T W_r z$ for some reduced quadratic tensor denoted as W_r [4].

3.2 MOR via Trajectory Piecewise-Linear Approximation

Quadratic model evaluated around a single operating point often produces a poor approximation when dealing with a highly nonlinear system [11]. QA typically expands the nonlinear operator about a single state and therefore the generated models are only accurate locally. The TPWL approximation overcomes this problem by considering multiple linearised models about suitably selected states of the system, instead of relying on a single expansion around the initial state [12, 1, 13]. Consider s linearised models around states $x_0, \ldots, x_{(s-1)}$ of (1) [1]

$$\dot{x_i} = f(x_i) + A_i(x - x_i) + Bu \tag{13a}$$

$$y_i = C^T(x - x_i) \tag{13b}$$

where A_i are the Jacobians of $f(.)$ evaluated at states $x_i, i = 0, 1, ...(s-1)$.

Assuming q − th order ($q \ll N$) bases $V \in \mathbb{R}^{N \times q}$ (Note that bases V_i are different for different linearised models), yields a representation of reduced order system 6 [1, 10]:

$$\dot{z} = \left(\sum_{i=0}^{s-1} w_i(z) A_{ir}\right) z + \gamma . w(z) + \left(\sum_{i=0}^{s-1} w_i(z) B_{ir}\right) u \tag{14a}$$

$$y = C_r z \tag{14b}$$

where $A_{ir} = V_i^T A_i V_i$, $B_{ir} = V_i^T B_i$, $C_r = C^T V_i$, $\gamma = [V_i^T(f(x_0) - A_0 x_0), ..., V_i^T(f(x_{s-1}) - A_{s-1}x_{s-1})]$. The term $w(z) = [w_0(z), ..., w_{s-1}(z)]^T$ is a vector of state dependent weights $\left[\sum_{i=0}^{s-1} w_i(z) = 1\right]$. The selection of linearised points and weighting procedure is discussed further in this section.

Selection of linearised points. It is assumed that the linearisation of a nonlinear system at state x_i is accurate for a given state x if this state is close enough to the linearisation point x_i i.e. $\|x - x_i\|_2 < \varepsilon$ where $\|.\|$ is the euclidean vector norm. This means that x lies within a ball of radius ε and centered at x_i [12]. As it is not possible to cover the entire N-dimensional space with such linear models due to enormous memory and computational cost, the TPWL approximation proposes a collection of linearisation points x_i along a single fixed trajectory [1, 10]. The training trajectory is generated by simulating the response of the nonlinear system to a given training input signal. The generation of linearised models can be summarised in the following iterative procedure:

1. Generate a linearised model about the initial state and set $i = 0$.
2. Using the training input simulate the nonlinear system while $\underbrace{min}_{0 \le j \le i} \frac{\|x - x_j\|}{\|x_j\|} < \gamma$ where $\gamma(\gamma > 0)$ is a tuning parameter and $x_j, j = 0, 1, \ldots, i$ are selected linearisation points.

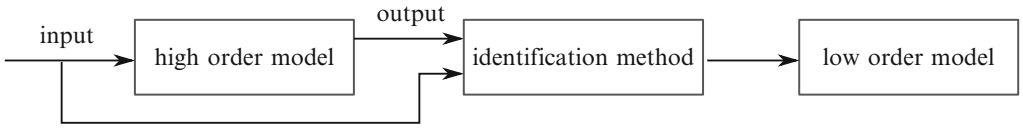


Fig. 3 Schematic diagram of data-based model order reduction

3. When $\underbrace{min}_{0\leq j\leq i} \frac{\|x-x_j\|}{\|x_j\|} > \gamma$, linearise nonlinear model about $x_{(i+1)} = x$, and set $i = i+1$ and go back to point 2.

Weighting Procedure: Linearised models are combined together using state dependent weights. The weighting coefficients $w_i(z)$ are given by $w_i(z) = \frac{\hat{w}_i(z)}{\sum_{(j=0)}^{(s-1)} \hat{w}_j}(z)$, where the term $\hat{w}_j(z)$ is function of the distance between z and the linearised point z_i. The coefficient $\hat{w}_j(z)$ is calculated as $\hat{w}_i = e^{\frac{-\beta d_i}{m}}$, where $d_i = \|z - z_i\|_2$, $m = min_{i=0,\ldots,(s-1)} d_i$ and $\beta > 0$ is a tuning parameter [1, 9, 10].

3.3 System Identification-Based MOR

MOR based on black box identification is schematically depicted in Fig. 3. Reduced order model is inferred from the response of the nonlinear high order model. Thus, the quality of the reduced order model entirely depends on the quality of the data used for the identification. The available methods allow for identification of both continuous-time and discrete-time models from sampled data [14, 15]. For the purpose of this research a discrete-time identification vial least squares method has been considered [15]. Analysis of the system response to a sinusoidal, impulse and step inputs indicates that a diagonal bilinear model structure suitable for modelling behaviour of the nonlinear system [16]. The selected model structure is described by

$$y_{k+1} = -\sum_{i=1}^{n} a_i y_{k-i} + \sum_{i=1}^{n_b} b_i u_{k-i} + \sum_{i=1}^{n_\eta} \eta_i u_{k-i} y_{k-i} \qquad (15)$$

The terms u_k and y_k are the system input and output, respectively, sampled at the time $t = kT_s$ where $T_s = 20$ ms is the sampling interval and k is the consecutive sample number. The terms a_i, b_i and η_i are model parameters, whilst n is the model order, and $n_b \leq n$ and $n_\eta \leq n$ are orders of the exogenous and bilinear terms. The model parameters have been identified using least squares method [17]. The data used for system identification are presented in Fig. 2.

Table 1 Number of flops required to simulate one discrete-time step of particular model

Model	Number of flops
Linear	$2(n_a + n_b) - 1$
QA	$2n^3 + n^2 + 4n - 1$
TPWL	$2n^2p + n$
bilinear	$2(n_a + n_b) + 3n_c - 1$

4 Efficacy Indices of Reduced Order Models

4.1 Consideration of Computational Cost

The computational cost of simulating a model can be calculated using the number of floating point operations (flops) [18]. One flop accounts for a single addition or multiplication operation [18]. Simulating a continuous time model requires use of a solver and various solvers result in different computational effort [19]. For the comparative study of the considered MOR methods, it is assumed that the structure-preserving Euler discretisation method [19] is used to simulate low order models obtained using the algorithms described in section 3. Computational effort of simulating a discrete-time interval using each of considered model structures is presented in Table 1, where n is the order of reduced model, n_a and n_b are the orders of the output and input polynomials, respectively, whilst n_c is the order of the bilinear polynomials.

4.2 Coefficient of Determination

The coefficient of determination is used to determine the accuracy of the model and is defined as [15]:

$$R_T^2 = 100\% \left(\frac{||\hat{y}_k - y_k||_2^2}{||y_k - \overline{y}_k||_2^2} \right) \qquad (16)$$

where $\hat{y}_k$ refers to the output simulated using a reduced order model, $\overline{y}_k$ is the mean value of the output.

5 Results

Study 1: This study is conducted in order to observe the performance of different MOR techniques for weakly nonlinear systems. A 100^{th} transmission line model is simulated using the input signal varying between 0 and 1. In order to select appropriate order of ROMs via model-based, Hankel singular values are computed. Right hand-side of Fig. 4 shows that, there are 20 dominant states therefore order of ROM selected to be 20 for model-based techniques. Nonlinear system is reduced to 12^{th} order using data-based technique. As expected, QA gives better performance compared to LA due to its ability of expansion of operating point. It has been observed that, performance of the TPWL mainly depends on tuning parameter γ. TPWL technique produces better approximation as it comprises of 7 ROMs. Bilinear model gives higher accuracy as compared to model-based techniques as well as less number of flops. Table 2 summarise efficacy indices of the ROMs for weakly and strongly nonlinear systems.

Study 2: In this study the input to the 100^{th} order model varies between 0 and 3, yielding a stronger nonlinearity than in study 1. Fig. 5 illustrate comparison of three different MOR techniques (QA,TPWL and bilinear). A $100th$ order model is reduced to $12th$ order bilinear model using system identification and $20th$ order using the TPWL and the QA. As shown in Table 2, compared to QA and TPWL, bilinear model has lower computational cost and high accuracy due to bilinearity type within nonlinear system. TPWL blends 33 linearised models, each of $20th$ order. Consequently, the computational cost of simulating TPWL model is higher compared to QA or bilinear model.

6 Conclusions

A comparative study of model-based and data-based techniques is conducted using a case study of transmission line model. Accuracy and complexity of these techniques are investigated by evaluating two efficacy indices, namely, $R_T{}^2$ and number of floating points operations required to simulate lower order models. It has been observed that, for this particular case of study, the bilinear model identification technique gives better performance in terms accuracy and number of flops as compared to quadratic approximation and trajectory piecewise linear approximation. The number of flops in trajectory piecewise linear approximation can be reduced by the use of fewer multiple quadratic models instead of many linear

Table 2 Efficacy for weakly and strongly nonlinear systems

	Weakly		Heavily	
	% R_T^2	flops	% R_T^2	flops
Linear	57	77	-	-
QA	88	16479	87.68	16479
TPWL	94.80	5620	92.43	26420
Bilinear	94.93	33	97.64	33

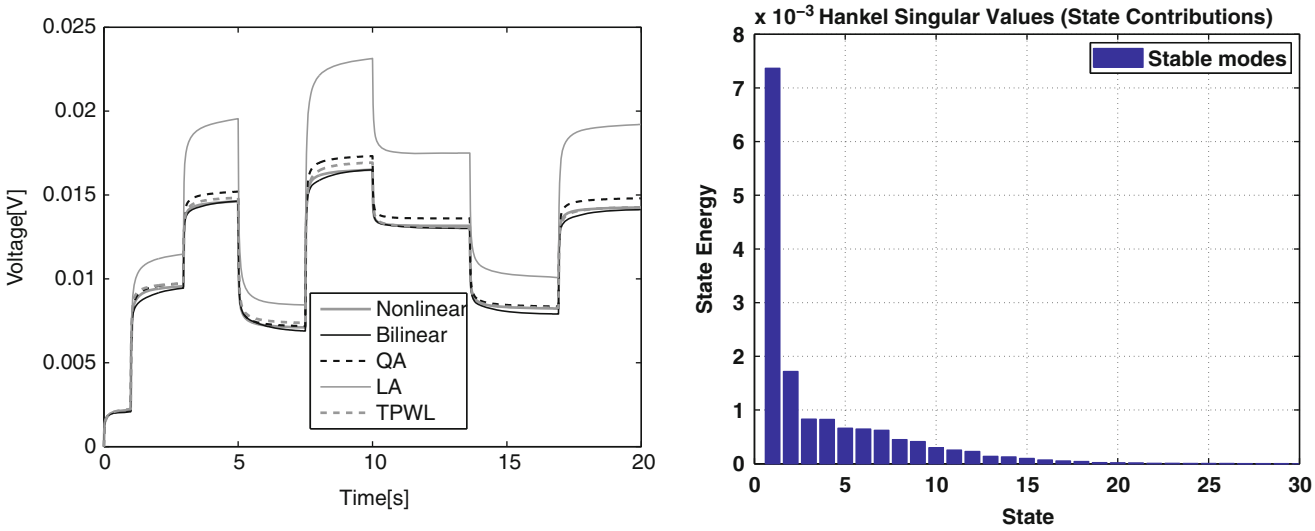


Fig. 4 MOR of weakly nonlinear system on left and Hankel singular values on right

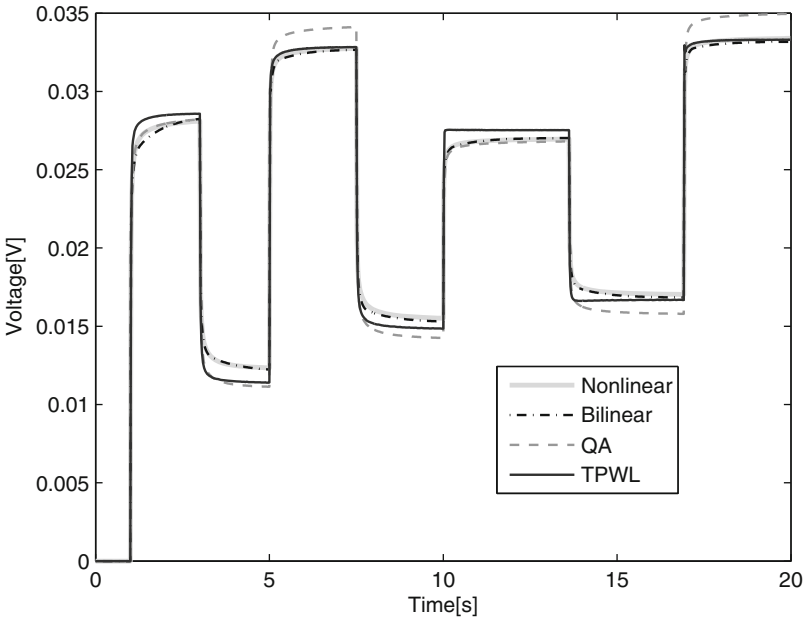


Fig. 5 Output results for different MOR technique

models or generating fewer reduced order models using multiple trajectories.

Acknowledgements This work is part of the EPSRC funded FUTURE Vehicles project (EP/I038586/1). The authors also acknowledge the support from project partners.

References

1. Rewienski, M.J.: A trajectory piecewise-linear approach to model order reduction of nonlinear dynamical systems. PhD thesis, Massachusetts Institute of Technology (2003)
2. Lai, X., Roychowdhury, J.: Tp-ppv: Piecewise nonlinear, time-shifted oscillator macromodel extraction for fast, accurate pll simulation. In: IEEE/ACM International Conference on Computer-Aided Design. (2006)
3. Aizad, T., Sumislawska, M., Maganga, O., Agbaje. O.and Phillip, N., K.J., B.: Investigation of model order reduction techniques: A supercapacitor case study. In: Proceedings of International conference on systems science. (2013)
4. Chen, Y.: Model order reduction for nonlinear systems. PhD thesis, Massachusetts Institute of Technology (1999)
5. Fillips, J.: Projection frameworks for model reduction of weakly nonlinear systems. In: Proceedings of the 37th Design Automation Conference. (2000)
6. Rugh, W.: Nonlinear System Theory. Johns Hopkins University Press (1981)
7. Sirovich, L.: Turbulence and the dynamics of coherent structures. Quart. Appl. Math **45** (1987) 561–571
8. Schilders, W., van der Vorst, H., Rommes, J.: Model Order Reduction: Theory, Research Aspects and Applications. Springer (2008)
9. Vesilyev, D.G., Rewienski, M., White, J.: A tbr-based trajectory piecewise-linear algorithm for generating accurate low-order models for nonlinear analog circuits and mems. In: Proceedings of the 40th Design Automation Conference. (2003)
10. Rewienski, M., White, J.: A trajectory piecewise-linear approach to model order reduction and fast simulation of nonlinear circuits and micromachined devices. IEEE transaction on computer-aided design of integrated circuits and systems **22** (2003) 155–70
11. Chen, Y., White, J.: A quadratic method for nonlinear model order reduction, department of electrical engineering and computer science. In: Proc of International Conference on Modelling and simulation of microsystems. (2000)
12. Rewienski, M., White, J.: A trajectory piecewise-linear approach to model order reduction and fast simulation of nonlinear circuits and micromachined devices. In: Proceedings of the 2001 IEEE/ACM international conference on Computer-aided design ICCAD '01. (2001)
13. Vob, T., Pulch, R., Maten, E.J.W., Guennouni, A.: Trajectory piecewise linear approach for nonlinear differential-algebraic equations in circuit simulation. Mathematics in Industry Volume **12** (2008) 167–173
14. Garnier, H., Wang, L., eds.: Identification of Continuous-time Models from Sampled Data. Number 978-1-84800-161-9. Springer (2008)
15. Ljung, L.: System Identification - Theory for the User. 2nd edn. PTR Prentice Hall Information and System Sciences Series. Prentice Hall, New Jersey (1999)
16. Pearson, R.: Discrete-time dynamic models. Oxford University Press (1999)
17. Hsia, T.: System identification: Least-squares methods. Lexington Books (1977)
18. Golub, G. H. Van Loan, C.F.: Matrix Computations. Johns Hopkins University Press, Baltimore and London (1996)
19. Butcher, J.: Numerical Methods for Ordinary Differential Equations. John Wiley & Sons (2008)

Optimised Job-Shop Scheduling via Genetic Algorithm for a Manufacturing Production System

Zhonghua Shen, Keith J. Burnham, and Leonid Smalov

1 Introduction

Job shop scheduling problem (JSSP) is one of most difficult discrete or combinatorial optimization problem in the planning and managing of manufacturing processes, which belongs to the class of (non-deterministic polynomial time) known as NP hard problems (Garey, Johnson and Sethi 1976), which is no closed from explicit solutions for JSSP. In this paper, it will be encountered with a micro-brewery production system. There is a set of orders to be produced in a set of vessels, each order is formed by a sequence of consecutive batch operations, and each batch operation can be produced in one vessel at a time without interruption until finish. The problem is how to schedule the batch operations to be produced in each vessel optimally. The further research is related to previous problem as defined by (Shen et al 2014). It has been formulated problems by mathematical model and simulated the scenario of a brewery production system by Matlab/Simulink. The further research is also concerned that to use heuristic methods to optimise the production system. Therefore, the purpose of this paper is to optimise the fermentation duration which is dependent on three beer products to be produced in the three different capacities of vessels.

Heuristic methods have been identified by (Seda 2008), which could be obtained an optimal solution for complex tasks that included genetic algorithm (GA), simulated annealing, tabu search, etc. In this paper, GA method will be implemented via Matlab to optimise a manufacturing production system as presented a micro-brewery production system. The paper is formed four sections. First of all, it reviews literatures of GA method. Furthermore, it assumes a complex production system in accordance with a real-life brewery production system and mathematic model will be formulated. What's more, it makes use of GA to optimise the production system. Finally, a conclusion and further work is provided.

2 Related Works

The job shop scheduling problem (JSSP) is well known as one of the most difficult NP-hard ordering problems. There are many approaches have been used to the JSSP, however, most of them cannot be obtained good solutions to solve large scale problems due to the computational time required, such as branch and bound and dynamic programming are only applicable to modest scale problems (Sun, Cheng and Liang 2010). It is also identified two successful main classes of meta-heuristics. One is the construction and improvement heuristic (Tabu search, simulated annealing, etc.), and another is the population based heuristic (GA, particle swarm optimization (PSO), artificial immune system and their hybrids, etc.).

Among the above methods, GA has been identified by (Chen et al 1999, Jia et al 2003, Ho and Tay 2004) that is one of most effective heuristic method to solve JSSP. It is used to identify approximate solutions by the principal of evolution for opmisation problem (Oprea and Nicoara 2005). It represents schedule as individuals or population. Each individual has its own fitness value which is evaluated by the objective function. The procedure of selection, crossover, and mutation, works iteratively to get better solutions until the stop criteria are satisfied in terms of required conditions and this iteration is a generation. The population size will remain constant from one generation to the next generation. The GA can be simplified as following Table 1.

Z. Shen • K.J. Burnham (✉) • L. Smalov
Control Theory and Applications Centre, Coventry University, Coventry, UK CV1 5FB
e-mail: Shenz3@uni.coventry.ac.uk; k.burnham@coventry.ac.uk; csx211@coventry.ac.uk

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_13,

3 Hypothesis

In a real-life brewery, there are n orders of various beer products have been arriving continuously to be formed a queue to wait for producing in the limited capacity in terms of fermentation vessels. Each order is to be accumulated for a batch production, each batch production can only be processed once in each vessel, and also each vessel has to be cleaned after each operation. Therefore, the beer production is time-based operation of brewing fermentation and other constraint conditions. In this paper, it is assumed that three beer products are to be produced simultaneously in the three parallel fermentation vessels of differing capacity and to obtain the minimum production time as following the Fig. 1.

Furthermore, three products can be denoted p_1, p_2 and p_3 separately. The production period of p_1, p_2 and p_3 are denoted t_{p_1}, t_{p_2} and t_{p_3}, where t_{p_1}, t_{p_2} and t_{p_3} are 72, 96 and 120 hours respectively. Three vessels can be denoted v_1, v_2 and v_3 separately. The maximum capacity of v_1, v_2 and v_3 are expressed in terms of barrels and denoted 20, 30 and 50 barrels respectively.

Subsequently, the operation of production is determined by the setting up time, fermentation time, cleaning time and changeover time as shown Table 2. The setting up time for p_1, p_2 and p_3 are denoted s_{p_1}, s_{p_2} and s_{p_3}, where s_{p_1}, s_{p_2} and s_{p_3} are 2, 3 and 5 hours respectively. The cleaning time for v_1, v_2 and v_3 are denoted c_{v_1}, c_{v_2} and c_{v_3}, where c_{v_1}, c_{v_2} and c_{v_3} are 2, 3 and 5 hours respectively. Furthermore, the changeover time might be occurred when the next batch production is to be changed in different vessels, and then it requires additional 5 hours for vessel cleaning. In addition, the due date for p_1, p_2 and p_3 are denoted d_{p_1}, d_{p_2} and d_{p_3}, where d_{p_1}, d_{p_2} and d_{p_3} are 5, 7 and 10 days respectively.

Table 1 Genetic Algorithm procedure

choose an initial population
determine the fitness of each individual
perform selection
repeat
perform crossover
perform mutation
determine the fitness of each individual
perform selection
until some stopping criterion applies

4 Modeling and Optimisation

4.1 Mathematical Model

The following notation is used as following Table 3.

4.2 Objective Function Formulation

Assume that three different types of beer are to be produced in three vessels separately, and no vessel changes at all, then the equation can be derived as follows:

$$T = \sum_{i=1}^{n} \sum_{j=1}^{n} x_{p_i v_j} m_{p_i v_j} \tag{1}$$

However, there are production process constraints, e.g. a due date of a product handover, a production delay, level of order priority. Therefore the following constraints are introduced

Table 2 Production of brewery

		p_1	p_2	p_2
t_{p_i} (hours)		72	96	120
s_{p_i} (hours)		2	3	5
d_{p_i} (days)		5	7	10
$c_{p_i v_j}$ (hours)	v_1	0	5	5
	v_2	5	0	5
	v_3	5	5	0

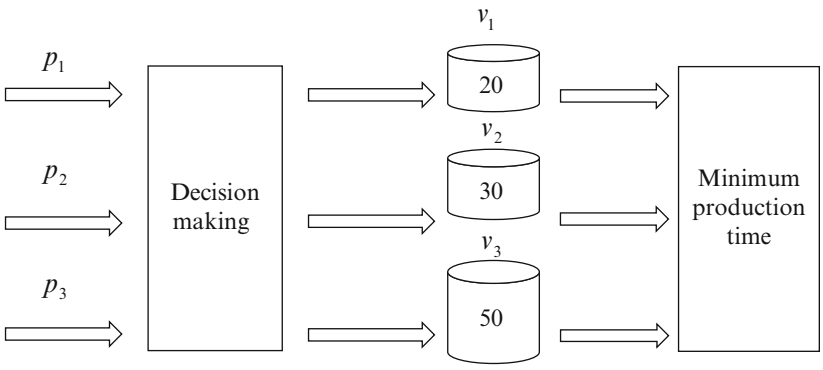


Fig. 1 Schematic model of a brewery production process

Table 3 Terminology

Abbreviation	Definition
p_i	Number of different products type i, $i = \{1, 2, 3\}$
v_j	Number of vessels j, $j = \{1, 2, 3\}$
$x_{p_i v_j}$	Number of occurrences of p_i made in vessel v_j
$m_{p_i v_j}$	Decision making: if coefficient is 1, working in the same vessel, if coefficient is 0, working in different vessel
t_{p_i}	The fermentation of p_i
n_{p_i}	Number of orders n to be scheduled for p_i, $n = \{1, 2, 3 \cdots n\}$
d_{p_i}	Due date for p_i
s_{p_i}	Setting up for p_i
$c_{p_i v_j}$	Changeover time for p_i and v_j
c_{v_j}	Cleaning time for v_j
v_{cap}	Capacity of v_j

to include the cleaning time, changeover time and setting up time. The objective function can be presented as follows:

$$T = t_{p_i}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{p_i v_j} m_{p_i v_j} + \sum_{j=1}^{m_1}\left(\sum_{i=1}^{n_1} s_{p_i} + \sum_{i=1}^{n_1} c_{v_j}\right) + \sum_{j=1}^{m_2}\left(s_{p_i} + \sum_{i=1}^{n_2} c_{v_j} + \sum_{i=1}^{n_2} c_{p_i v_j}\right) \quad (2)$$

4.3 Constraints

In the actual brewery production process, each vessel can produce only one batch of orders during a production process. This fact can be denoted as follows:

$$\sum_{i=1}^{n} m_{p_i v_j} = 1 \quad (3)$$

$$\sum_{j=1}^{n} m_{p_i v_j} = 1 \quad (4)$$

Where

$$m_{p_i v_j} \in \{0, 1\}, \forall_i \in \{1, 2, 3, \cdots, m\}, \forall \in \{1, 2, 3, \cdots, n\}$$

The capacity of vessels cannot be exceeded when the accumulated orders of the same type of beer are to be scheduled to process. This can be represented as

$$0 \leq n_{p_i} \leq v_{cap} \quad (5)$$

The due date of orders is also considered here, and production needs to be completed before or on the due date (the lateness will have an effect on customer satisfaction):

$$t_{p_i} + \left(\sum_{i=1}^{n_1} s_{p_i} + \sum_{i=1}^{n_1} c_{v_j}\right) \leq d_{p_i} \quad (6)$$

$$t_{p_i} + \left(\sum_{i=1}^{n_2} s_{p_i} + \sum_{i=1}^{n_2} c_{v_j} + \sum_{i=1}^{n_2} c_{p_i v_j}\right) \leq d_{p_i} \quad (7)$$

4.4 Genetic Algorithm (GA) Optimisation

Before the GA simulation, several essential conditions need to be considered as follows:

- Three different products working in three parallel vessels with limited capacity
- A set of fixed processing time for each product
- Each vessel must be processed one batch production only, once a vessel starts to process a batch of orders, no interruption is allowed, and then it needs to be cleaned until finish
- The arriving orders will be accumulated batches to meet the required capacity of vessel
- The production time cannot be exceeded the due date

In this case, it implements GA method based in the Matlab software to simulate and optmise the production system. A chromosome is represented as a list of orders for the three products, which is need to scheduled according to prescribed criteria and constraints in the production process. Each individual or population has a limited processing time and setting up time, and changeover time. The process of evolution is repeated as long as the stopping criteria are not satisfied. In Fig. 2 shows the result of decrease of the average makspan for the total production time of three products working in limited three vessels, it using the population size 100 and 400 generations. It is possible to note that GA improves the average makespan very fast and the optimal sequence of simulation is obtained, the best makespan is 850 after 95 generations.

5 Conclusion and Future Work

This paper presents a genetic algorithm (GA) for the resource constrained scheduling problem for a brewery production job-shop scheduling problem based on setting up time, priorities, delay time, and limited capacity of vessels. The objective function has been formulated by mathematical model and constraints have been defined. GA has been implemented to optmise the scheduling problem for a micro-brewery production system, the optimal sequence is obtained after 95 generations.

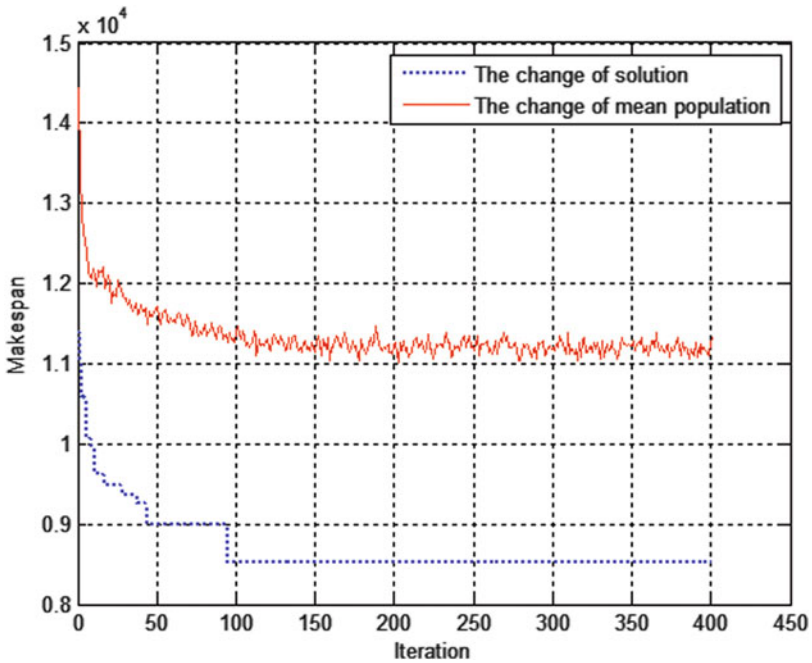


Fig. 2 Decreasing of the makespan

The future work will concentrate on the difference heuristic algorithms to optimise a more complex production system. It will also to compare them to find out the optimum solutions.

References

1. Chen.H, Ihblow.J, and Lehmann.C (1999) A genetic algorithm for flexible job-shop scheduling. IEEE International Conference on Robotics and Automation, Detroit, p1120–5.
2. Garey, E.L., Johnson, D.S and Sethi, R. (1976) The complexity of flow shop and job shop scheduling. Mathematics of Operations Research, p117–129.
3. Ho.N and Tay.J (2004) An efficient cultural algorithm for solving the flexible job-shop problem. IEEE International conference on robotics and automation, 1759–66
4. Jia.H, Nee.A, Futh.J and Zhang (2003) A modified genetic algorithm for distributed scheduling problems. International Journal of Intelligent Manufacturing,v14:351–62
5. Oprea.H and Nicoara.S (2005) Artificial intelligence, Petroleum-gas University of Ploiesti.
6. Seda, M. (2008) Mathematical models of flow shop and job shop scheduling problems. International Journal of Applied Mathematics and Computer Sciences 4, (4).
7. Shen.Z, Burnham.J.K, Samlov.L, and Amin.S (2014) Formulating scheduling problem for a manufacturing production system. Conference Proceeding of International Conference on Computer, Network Security and Communication Engineering, DEStech Pubilications.
8. Sun.L, Cheng.X and Liang.Y (2010) Solving job shop scheduling problem using genetic algorithm with penalty function. International Journal of Intelligent Information Processing, volume1, No.2.

3D CFD Simulation of the Thermal Performance of an Air Channel Solar Heater

Samir Moujaes and Jayant Patil

1 Introduction

Over last few years there has been a growing interest in the field of solar air heaters for solar assisted crop drying technique. Drying is a process of heat transfer to the product from a heating source. Solar energy allows various ways to construct a system and thereby it possesses a mode which can necessitate a simple technology for a particular usage and for a particular region. In this study at hand Las Vegas, Nevada is considered as potential region of interest for use of this type of heaters but for residential applications. Solar air heaters are the devices employed to gain useful heat energy from incident solar radiation. Solar collectors can be of concentrating or flat plate type. For solar energy crop drying applications or residential applications flat plate collectors provide the desired temperature ranges.

Several designs of the solar air heaters had been proposed in the past. Among the recent ones [1] presented the concept of applying different heat sources fro agricultural drying applications using six different natural circulation solar air collectors. Each of the collectors has combinations of glazing and different types of absorber plates such as zigzag, flat plate, front pass, back-pass etc. [2] performed the analysis of multi-tray crop drying attached to an inclined multi-pass solar air heater with in-built thermal storage. Another work [3] provided an optimization method for agricultural drying using forced convection hot air dryer. Another paper [4] reported on the design of a solar dehydrator for agricultural crops in India. The unit consists of solar air heart and a drying chamber. The study focuses on calculating the drying ratio and rehydration ratio. Finally [5] analyzed a solar air collector with different heat transfer coefficients in the collector and within the external environment. The study is aimed to calculate the fluid outlet temperature, efficiency as the function of wind speed, incident solar radiation and mass flow rate.

A review of these previous research works did not reveal any work using CFD simulations for the purpose of predicting the performance of any of these solar air heater designs which is the reason for this study and to show the flexibility and the power of using a computational fluid dynamics for the purpose of initial evaluation of scoping designs.

2 Physical Model

Fig. 1 shows the physical model simulated. The dimensions of the collector are 0.85 m x1.22 m (W x L). The height of the solar collector is 16.5 cm. The top cover of the collector is made of transparent glass (0.3175 cm thick). The insulation material for the collector is glass wool. The sides and the bottom of the collector are insulated with 5.0 cm thick layer of glass wool. The surfaces of this cavity heat up and eventually impart heat to the flowing air through that volume. This is a first attempt at the CFD simulation to double check and make sure that the model is giving reasonable data for exist temperatures of the air as it exists from the other end of the cavity. The inside of the cavity walls are assumed to be painted with a dark color paint to enhance the radiation exchange between the surfaces and the air.

3 Theoretical Model

In this study we have simulated a three dimensional, single pass solar air heater model using a computational fluid dynamic approach STAR-CD. The analysis is performed for a summer day (June 21) for Las Vegas region in Nevada.

The ASHRAE Clear Sky Model [9] is used to calculate the value of solar irradiation at the earth's surface for five different times of day for this simulation (6 AM,9 AM, 12 noon,3 PM,6 PM).

S. Moujaes (✉) • J. Patil
ME Department, University of Nevada Las Vegas,
Las Vegas, NV 89154-4027, USA
e-mail: samir.moujaes@unlv.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_14,

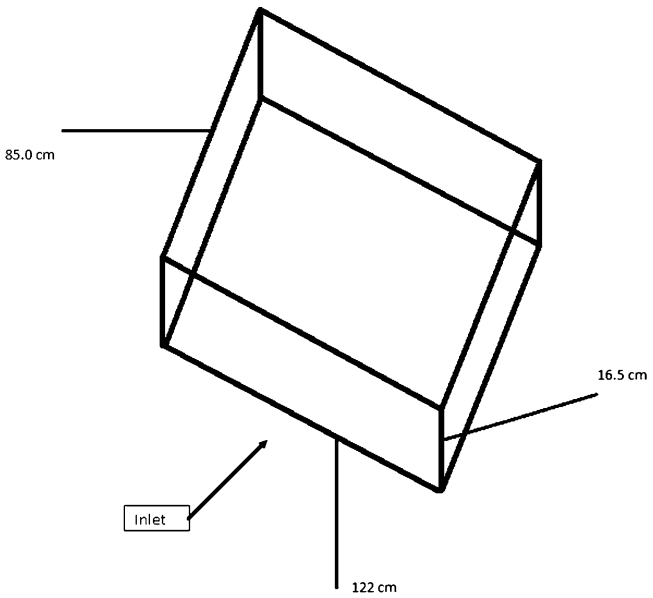


Fig. 1 Schematic Diagram of Solar air collector

The equations to calculate the total solar radiation on vertical and horizontal surfaces are given below[9]:

(a) Total solar radiation incident on a **vertical surface**

$$G_t = G_D + G_d + G_R \tag{1}$$

The direct radiation corrected for clearness is given by

$$G_D = G_{N^*D} * \max(\cos\theta) \tag{2}$$

For vertical surfaces the diffuse sky radiation is given by

$$G_d = \frac{G_{dV}}{G_{dH}} * C^* G_{ND} \tag{3}$$

The rate at which energy is reflected onto the wall is given by

$$G_R = G_{tH}{}^* \rho_g{}^* F_{wg} \tag{4}$$

The value of the solar irradiation at the surface of the earth on a clear day is given by

$$G_{ND} = \frac{A}{\exp\left(\frac{B}{\sin\beta}\right)} * C_N \tag{5}$$

The ratio of diffuse sky radiation on a vertical surface to that incident on a horizontal surface on a clear day is given by

$$\frac{G_{dV}}{G_{dH}} = 0.55 + 0.437^* \cos\theta + 0.313 * \cos^2\theta \tag{6}$$

The configuration or angle factor from wall to ground is

$$F_{wg} = \frac{1 - \cos\alpha}{2} \tag{7}$$

The rate at which the total radiation strikes the horizontal surface or ground in front of wall is given by

$$G_{tH} = (\sin\beta + C)^* G_{ND} \tag{8}$$

(b) Total solar radiation on a **horizontal surface** is

$$G_t = \left[\max(\cos\theta) + C^* F_{ws} + \rho_g{}^* F_{wg}{}^* (\sin\beta + C)\right] * G_{ND} \tag{9}$$

The fraction of energy that leaves the surface and strikes the sky directly is given by

$$F_{ws} = \frac{1 + \cos\alpha}{2} \tag{10}$$

4 Computational Model

The computational analysis is performed using a steady state solution method for each of the five times of the day mentioned above as it is assumed that the thermal capacitance for this system is rather small and hence steady state conditions are achieved rather quickly at each hour of solar insolation exposure.

The physical problem comprises of heat transfer between the collector side and outside air is considered on the top and bottom of the collector, conjugate heat transfer property of the simulation software is employed during the analysis. Glass wool is used as the material of insulation on the outside of the solar collector. The CFD model takes into account the calculation concerned with the glass cover. The STAR –CD package has in built glass property section which gives the scattering and absorption coefficient of glass. The solar radiation which falls on top of the glass cover is the total heat gain for the solar air collector to heat the moving air inside it. A turbulent k-ε model is used for the analysis because of the Reynolds number (~14,000) is considered turbulent for a rectangular channel. The channel is assumed to be oriented with its air flow moving in a N-S orientation. Some parts of the solar collector are subjected to shading due to the vertical walls and are considered as such at the appropriate times of the day. To account for this shading caused by the walls, separate calculations for the shading calculations are included in the model (for each run) which depicts the portion of the area that is under shade and the one which is in direct sunlight for any particular time of the day. The velocity of air at the inlet of the solar collector

is assumed to be 0.74 m/s. The model is analyzed with a solution tolerance limit of 10e-4. The numbers of nodes used in the model are 250,000 to insure grid independency. The cells of the computational model are chosen to be cubical in shape.

When the solution is converged the computational model generates an information file which predicts the enthalpy difference between the inlet and outlet of the collector. The final temperature (mean bulk temperature as calculated at the exit cross-section by STAR-CD) of the air moving exiting the collector is calculated as [8]:

$$\Delta H = \dot{m}^* C_p{}^* \Delta T \tag{11}$$

$$\Delta H = \dot{m}^* C_p{}^* (T_{in} - T_{out}) \tag{12}$$

The efficiency of the collector is calculated as

$$\eta_{th} = \frac{\Delta H}{W} \tag{13}$$

Where ΔH is the change in energy from inlet to outlet and W is the amount of heat flux calculated from the ASHRAE clear sky model incident for that particular time of the day on the glass cover.

The value of the heat transfer coefficient on the exterior of the collector is calculated as [8]:

$$h = \frac{x_1}{k_1} + \frac{x_2}{k_2} \tag{14}$$

Where x and k are the thickness and the thermal conductivity of solar collector material and insulation material respectively.

5 Results and Discussion

Fig. 2 shows the results of the bulk averaged temperature across the exit cross section of the air heater for the five times of the typical simulation day used in this study. The plot indicates a gentle rise in that temperature initially till about 9 AM and then a much steeper rise till about 3:00 PM where that temperature is maximum and then a rapid decline of its value as the late afternoon hours progress. Although the insolation values are usually expected to maximize at around noon the ambient temperature is still increasing till late in the afternoon which is why it partially explains the time shift in that maximum value from solar noon.

Fig. 3 shows the variation of the thermal efficiency as defined in equation (13). It shows that the value of η_{th} it shows that the efficiency is small at the start of the day indicating that the magnitude of the losses is large with respect to the total insolation on the air heater. AS the day progresses these losses although the grow larger but don't exceed the gains of heat into the air heater medium. That efficiency is maximized and stays flat over a period of roughly 9:00 AM till 3:00 PM in the afternoon indicating a stabilization between the heat losses and inputs to the system. Towards the latter part of the afternoon a steady decrease is seen from a valued of about 76 % till around 30 % due to the fact that in the late afternoon the sun's rays are at a higher inclined angle than at noon hence more reflection and ambient temperature also start decreasing as well increase the total heat losses relatively speaking.

Fig. 4 shows a localized plot of the axial velocity along the central axis of the channel's flow. It shows a slight increase of that velocity due to the effects of boundary

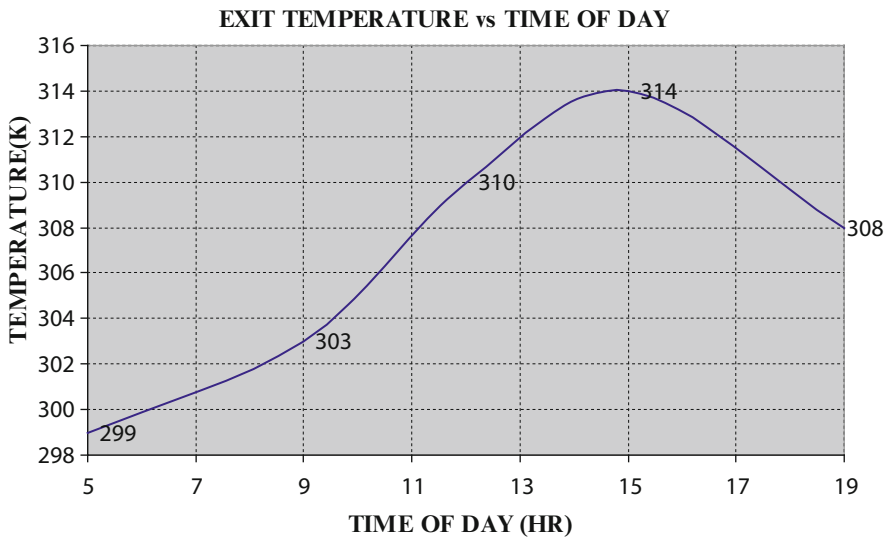


Fig. 2 Exit air temperature at five different times of the day (6 AM,9 AM,12 noon,3 PM,6 PM)

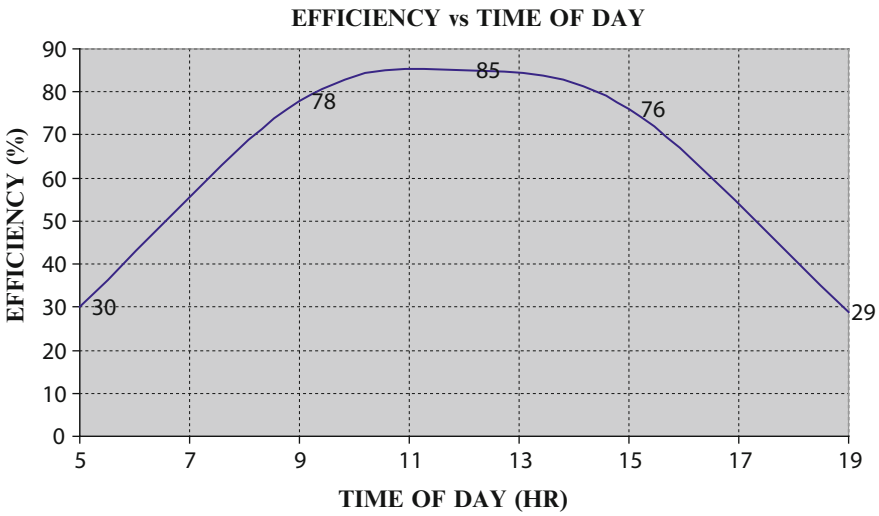


Fig. 3 Efficiency of the solar air collector for the five different times of the day

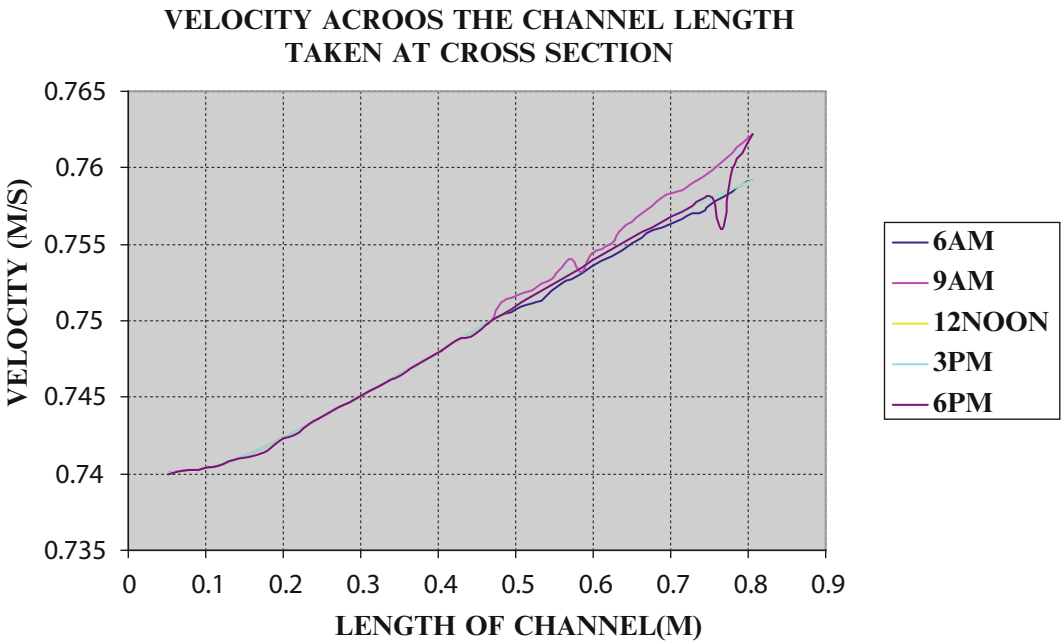


Fig. 4 Velocity profile taken at cross section of channel from inlet to exit

layer flow on the wall boundaries as well as some instability toward the exit locations probably due to the exit conditions affecting some of the upstream flow. Again for most of the plots during the different times of day the velocity along this axis does not show major variations as the air is assumed to be an incompressible fluid at these temperatures and pressure conditions. As mentioned earlier in the paper the assumed inlet velocity was taken to be a uniform value across the inlet at 0.74 m/s.

Finally Fig. 5 shows a plot of the pressure variation at the same axial location mentioned in Fig. 4. As one would expect there are no significant pressure drops in the air flow at these average velocities as air has a relatively low density and viscosity that would not incur any major pressure drops. His is shown in Fig. 5 as a horizontal line for all the runs indicating a small variation as expected.

6 Conclusion and Future Work

The results of the CFD study shown above shows the potential of using this approach to adequately describe the trends and performance of a solar air heater adequately. Various design features can be added to the basic design to

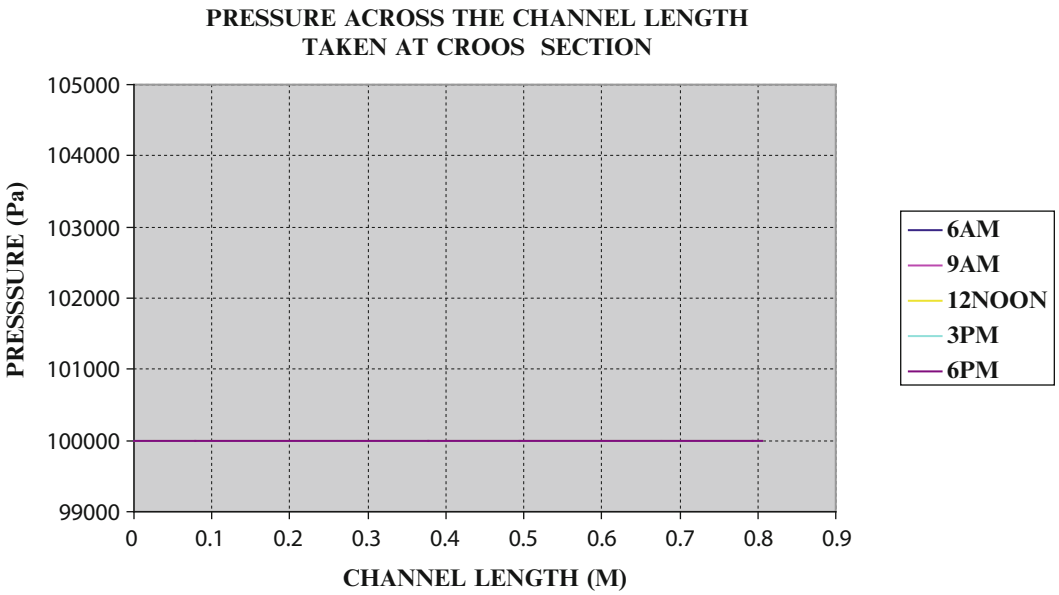


Fig. 5 Pressure profile taken at cross section of channel from inlet to exit

accommodate the different unique features the various designer might envision to increase its thermal efficiency. The CFD results of this study match closely with the experimental work done by Othman, Yatim, et. On a double pass solar air collector.

The double pass solar air collector has curve shaped solar plates which are black in color where as in the current CFD study only plane baffles are used. The solar collector plates in double pass solar collector are located on either side of the medium separating the flows which inurn increases the efficiency. Still with the single pass, CFD study shows that the results are in close agreement with the double pass solar air collector.

Future work would include a more automated method of calculating the solar insolation at various times of the day for different locations and would incorporate the varying reflected sunlight at those times of day so that the net transmitted solar heat is calculated at each different time. In a future development a study of how naturally convected air through the solar air heater can be also incorporated which can be used to study a variety of flow situations.

Nomenclature

G_t	Total solar radiation incident on surface (w/m^2)
G_D	Direct radiation corrected for clearness (w/m^2)
G_d	Diffuse sky radiation (w/m^2)
G_R	Rate of reflected energy (w/m^2)
G_{ND}	Normal direct irradiation (w/m^2)
θ	Angle of incidence
G_{dV}/G_{dH}	Ratio of diffuse sky radiation for a vertical wall to horizontal surface
C	Clearness number
G_{tH}	Rate at which total radiation strikes surface in front of wall (w/m^2)
ρ_g	Reflectance
F_{wg}	view factor
A	Apparent solar irradiation (w/m^2)
B	Atmospheric extinction coefficient
β	Solar altitude
α	Tilt angle
F_{ws}	Fraction of energy that leaves surface and strikes sky directly (w/m^2)
ΔH	Enthalpy difference
m	Mass flow rate(kg/s)
h	Heat transfer coefficient(w/m^2.k)
x	Thickness of insulation(m)
k	Thermal conductivity(w/m^2.k)
η_{th}	Thermal efficiency as defined in eq. (13)
1,2	solar heater material index and insulation material index

References

1. Koyunchu, T: Performance of various designs of solar air heaters for crop drying, applications *Renewable Energy*, v 31, n 7, p 1055-1071, June 2006
2. Dilip, J: Modeling the system performance of multi-tray crop drying using an inclined multi-pass solar air heater with in-built thermal storage, *Journal of Food Engineering*, v 71, n 1, p 44-54, November 2005

3. McDoom, I.A.; Ramsaroop, R.; Saunders, R.; Tang Kai, A.:Optimization of solar crop drying, *Renewable Energy*, v 16, n 1-4 -4 pt 2, p 749-752, January/April 1999
4. Garg, H.P, Mahajan, R.B.; Sharma, V.K.; Acharya, H.S..: Design and Development of a Simple Solar Dehydrator for Crop Drying, *Energy Conversion and Management*, v 24, n 3, p 229-235, 1984
5. Shemski, S; Bellagi F., Ahmed, S.:Drying of agricultural crops by solar, *Desalination*, v 168, n 1-3, p 101-109, August 15, 2004
6. Supranto S., K.; Daud, W.R.W.; Othman, M.Y.; Yatim, B. Design of an experimental solar assisted dryer for palm oil fronds, *Renewable Energy*, v 16, n 1-4 -4 pt 2, p 643-646, January/April 1999
7. Kreider F. J.;Kreith F.: Solar heating and cooling, Hemisphere Publishing Corporation and McGraw-Hill Book Company (1976)
8. Bergman T., Lavine A., Incropera F., DeWitt D.; Introduction to Heat Transfer, 6th Edition, 2011, Wiley &sons
9. McQuiston F., Parker J., Spitler J.: Heating, Ventilating and Air Conditioning: Analysis and Design, 6th Edition, Wiley &sons 2005

A Concept Study for a Compact High-Speed Rotation Heat Pump

Haakon Karlsen and Tao Dong

1 Introduction

As a well-studied topic and a mature technology, heat pumps are systems capable of moving heat from low temperature mediums to higher temperature mediums towards the direction of spontaneous heat transfer [1]. They can be an energy efficient way of using energy in energy-intensive processes as cooling/heating and can reuse heat from the environment and other heat sources. There are several types of heat pumps, where the general types tends to rely on external mechanical work like mechanically driven vapor-compression heat pump systems [2], or using absorption cycles involving geothermal heat or solar assistance [3–5]. In this work a new concept of rotational heat pump concept is introduced which takes advantage of the high centripetal force from high-speed rotation to make a compact device combining compression and intermediate heat exchange in one unit. The device utilizes two separated closed gas cycles interacting with each other through a heat exchanger, in such a way that it can be viewed as being a combination of two Brayton cycles, one reversed and one normal, where compression is partially driven by the rotation itself and partially by external compressor.

2 Theory

2.1 Rotation

Individual sections which the concept system comprises are envisioned as u-shaped tubes where inlet and outlet is located close to the axis of rotation, extends radially outwards a length H and is connected by an axially extending tube, see Fig. 1. The u-shape is rotated together with the coordinate axes about the z-axis, and observed in a rotating reference frame. Acceleration in a non-rotating reference frame and a rotating reference frame is related through [6]:

$$a_i = a + 2(\vec{\Omega} \times u) + \vec{\Omega} \times (\vec{\Omega} \times r) + \dot{\Omega} \times r \qquad (1)$$

Where $\vec{\Omega} = \Omega \mathbf{e} = \Omega(e_x, e_y, e_z)$ is angular frequency vector, $\boldsymbol{u} = (u, v, w)$ is velocity vector and $\boldsymbol{r} = (x, y, z)$ is displacement vector from axis of rotation. For an inertially stationary system with constant rotation frequency about the z-axis, Eq. 1 when multiplied by density ρ can be written as the sum of Coriolis and Centrifugal force. Assuming that $H/b \gg 1$, the forces are simplified as: $F_{cor} = 2\rho\Omega u$, $F_{cent} = \rho\Omega^2 r$, where r is distance from the rotation axis in the rotating reference frame, and u is mean radial velocity. Their ratio is defined as the Rossby number $Ro = u/\Omega r$ [7].

Rotation Frequency is set to a certain level where $g_c/g \gg 1$ the centrifugal acceleration dominates over gravity. The u-shaped tube contains a barotropic fluid which is affected in response to the rotation. Assuming (initially) insulated walls and that transient velocity have decayed (leaving the coriolis force effectively zero), the internal equilibrium requires the force sum on all fluid elements to be zero, giving the hydrostatic equilibrium equation [8] for the rotating system:

$$\frac{dp}{dr} = \rho\Omega^2 r \qquad (2)$$

Considering the 2nd law of thermodynamics and no heat transfer, Eq. 2 can be solved for the isentropic density and pressure relationship giving:

$$\frac{p}{p_0} = \left(\frac{\rho}{\rho_0}\right)^{\gamma} = \left(\frac{T}{T_0}\right)^{\frac{\gamma}{\gamma-1}} = \left(1 + \frac{\Omega^2(r^2 - r_0^2)}{2c_p T_0}\right)^{\frac{\gamma}{\gamma-1}} \qquad (3)$$

H. Karlsen (✉) • T. Dong (✉)
Department of Micro and Nano Systems Technology, Faculty of Technology and Maritime Sciences, Buskerud and Vestfold University College, 2243, N-3103 Tønsberg, Norway
e-mail: tao.dong@hbv.no

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_15,

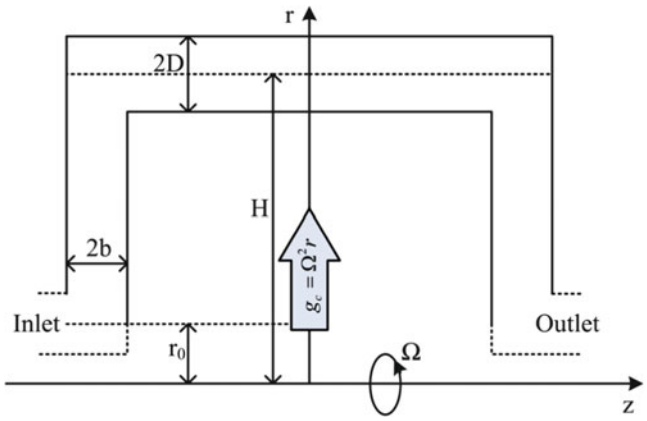


Fig. 1 Geometry for u-shaped tube

Where γ is specific heat capacity ratio and c_p is specific heat capacity at constant pressure, and subscript '0' is ambient condition.

2.2 Heat transfer section

Two sets of u-shaped tubes forming one sector of the complete part are cylindrically distributed with an angle θ, ensuring no intersection between radial tubes. Radial tubes are still considered thermally insulated, but the axial tubes of both u-shaped tubes are defined and treated as a heat transfer section

The heat transfer between point 2 and 3 is described with the Energy balance/ε -NTU method [9, 10]. We define the heat capacity rate $= \dot{m}c_p$, Minimum heat capacity rate $C_{min} = min[C_1, C_2]$ and the non-dimenional parameter $\varepsilon = q/q_{max}$ describing the ratio of actual transferred heat over maximum transferred heat, where maximum transferred heat $q_{max} = C_{min}(T_{hi} - T_{ci})$. Subscript '*hi*' and '*ci*' indicate hot and cold inlet respectively.

Assuming the depth averaged pressure in the axial tube is approximately constant:

$$\overline{P} = \frac{H}{2D}\int_{1-D/H}^{1+D/H} pd(r/H) \tag{4}$$

For the isobaric heat transfer between location 2 and 3, the temperature at location 3 is given by:

$$T_{13} = T_{12} - \varepsilon\frac{C_{min}}{C_1}(T_{12} - T_{22}) \tag{5}$$

$$T_{23} = T_{22} + \varepsilon\frac{C_{min}}{C_2}(T_{12} - T_{22}) \tag{6}$$

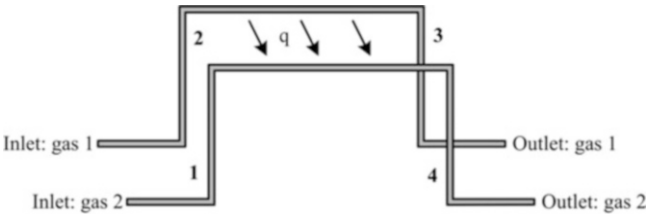


Fig. 2 Two separate u-shaped tubes. Four locations are defined: **1**) Inlet state entering radial tube, **2**) compressed state entering axial tube, **3**) heat exchanged state entering radial tube, **4**) outlet state exiting radial tube

First subscript refers to which u-shape in Fig. 2, second subscript refers to location in Fig. 2. Subscript '12' is defined as hot inlet '*hi*', and subscript '22' is defined as cold inlet '*ci*'.

2.3 Pressure balance

For a single rotating u-shaped tube in equilibrium filled with an ideal gas without heat transfer, with no difference in conditions between inlet and outlet, the force acting on fluid columns $1 \rightarrow 2$ and $4 \rightarrow 3$ are equal. By imposing a steady mass flow ($\dot{m}$ in the heat capacity rate) with sufficiently low velocity comparing to the tangential velocity to ensure that the Ekman-number defined as $Ek = \mu/\rho\Omega b^2 \approx 0$, Mach-number defined as $Ma = u/c \approx 0$, and Rossby-number $Ro \approx 0$, so that terms related to viscosity, convection and coriolis force, all involving velocity, are of negligible magnitude allowing (for isentropic conditions) both the momentum equation and the energy equation to be reduced to the hydrostatic equation. Thus the pressure differences between $1 \rightarrow 2$ and $3 \rightarrow 4$ cancels out making the necessary driving pressure on the order of what would be expected for a non-rotating tube. Velocity relative to initial velocity along the radial tube can be found through the steady continuity equation to be inversely proportional to the density ratio, $u/u_0 = (\rho/\rho_0)^{-1}$. And the axial tube velocity can be expressed approximately through the continuity equation for tubes with varying crossection $u_{ax}/u \approx (b/D)^2$.

If there is a temperature difference between '*hi*' and '*ci*' and heat is allowed to exchange between fluids in u-shape 1 and u-shape 2, the '*ci*' fluid being heated from location $2 \rightarrow 3$, becomes warmer and lighter causing an imbalance between the force on the two hydrostatic fluid columns thus encountering a lower force from $3 \rightarrow 4$ than from $1 \rightarrow 2$, $\Delta F = F_{3\rightarrow4} - F_{1\rightarrow2} < 0$, and is ideally not dependent upon a driving force (depending on the magnitude of ΔF). The warmer fluid being cooled from location 2-3, encounters a larger force from $3 \rightarrow 4$ than from $1 \rightarrow 2$,

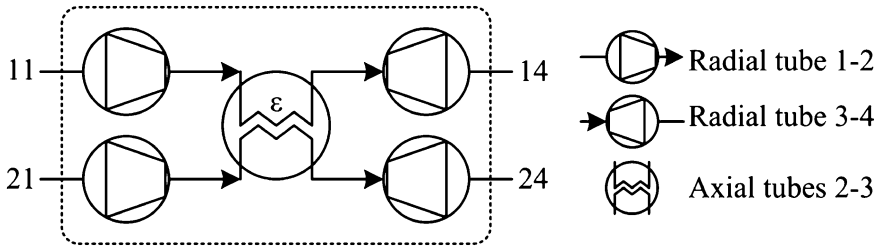


Fig. 3 Schematic drawing of full rotating part as lumped element with 2 fluid inputs and 2 outputs

$\Delta F = F_{3\to4} - F_{1\to2} > 0$, and then relies on a driving pressure to overcome this pressure difference.

2.4 External components

A minimum of necessary external components to close the gas loops and complete the system are a compressor, a turbine, and two heat exchangers, as depicted in Fig. 4. A compressor is necessary to drive the loop that releases heat from 2 →3, a turbine for the loop that receives heat from 2 → 3, and two heat exchangers for transferring heat to and from the environment of the respective sides and bring the temperature back to the respective ambient conditions.

For the compressor and turbine, the work (-rate) is defined as conventional, with adiabatic efficiency. For the compressor:

$$W_c = \frac{1}{\eta_c} C\Delta T_{cs} \tag{7}$$

For the tubine:

$$W_t = \eta_t C\Delta T_{ts} \tag{8}$$

Where subscript 'c' and 't' is compressor and turbine respectively, 's' is necessary theoretical to achieve wanted pressure ratio. The theoretical temperature difference for a given pressure ratio is calculated from entropy conservation. $\Delta s = 0;\ p = KT^{\gamma/(\gamma-1)}$. The external heat exchangers are considered to be sufficiently large to operate at close to 100 % efficiency, and transfer heat to/from infinite reservoirs.

3 Results

3.1 Working fluids

This system is intended to work with barotropic fluids; the present analysis is intended for ideal gases, with an operating temperature above critical temperature, and operating pressure below critical pressure, to ensure that the ideal gas assumption is a valid approximation. But the methods can easily be extended to handle more complex fluids and real behavior through more accurate equations of state and departure functions [11, 12].

For a hydrostatic temperature increase on the order of $\Delta T = (\Omega r)^2/2c_p$, the main important factor apart from tangential velocity is the specific heat capacity. The purpose of the hydrostatic temperature increase is to overcome the source temperature difference. To achieve this the temperature increase for the cold environment gas must be larger than for the warm gas, and needs a lower c_p. Possible candidates can be for example Argon for cold environment working fluid, and air for warm environment working fluid.

3.2 Combine equations and components

Combining the equations for the internal sections of the rotating part without external components, Fig. 3, gives outlet temperatures:

$$T_{14} = T_{11} - \varepsilon \frac{C_{min}}{C_1}(\Delta T_s + \Delta T_\Omega) \tag{9}$$

$$T_{24} = T_{21} + \varepsilon \frac{C_{min}}{C_2}(\Delta T_s + \Delta T_\Omega) \tag{10}$$

Where $\Delta T_s = T_{11} - T_{21}$ is the source temperature difference (with '11' indicating the ambient cold environment temperature), which for all practical heat pump applications are defined negative, and $\Delta T_\Omega = \Delta T_1 - \Delta T_2$ is the difference in rotation driven temperature increase between loop 1 and 2. For outlet pressures, the imbalance is calculated by: $p\_4/p\_1 \approx (p\_2/p\_1)(p\_4/p\_3)$, for $\overline{p}_2 \approx \overline{p}_3$, and is depicted in Fig. 4 for a specific situation.

$$\frac{p_{14}}{p_{11}} = \left[\frac{1 + \frac{\Delta T_1}{T_{11}}}{1 + \frac{\Delta T_1}{T_{11} - \varepsilon\frac{C_{min}}{C_1}(\Delta T_s + \Delta T_\Omega)}}\right]^{\frac{\gamma_1}{\gamma_1 - 1}} (\leq 1) \tag{11}$$

$$\frac{p_{24}}{p_{21}} = \left[\frac{1 + \frac{\Delta T_2}{T_{21}}}{1 + \frac{\Delta T_2}{T_{21} + \varepsilon\frac{C_{min}}{C_2}(\Delta T_s + \Delta T_\Omega)}}\right]^{\frac{\gamma_2}{\gamma_2 - 1}} (\geq 1) \tag{12}$$

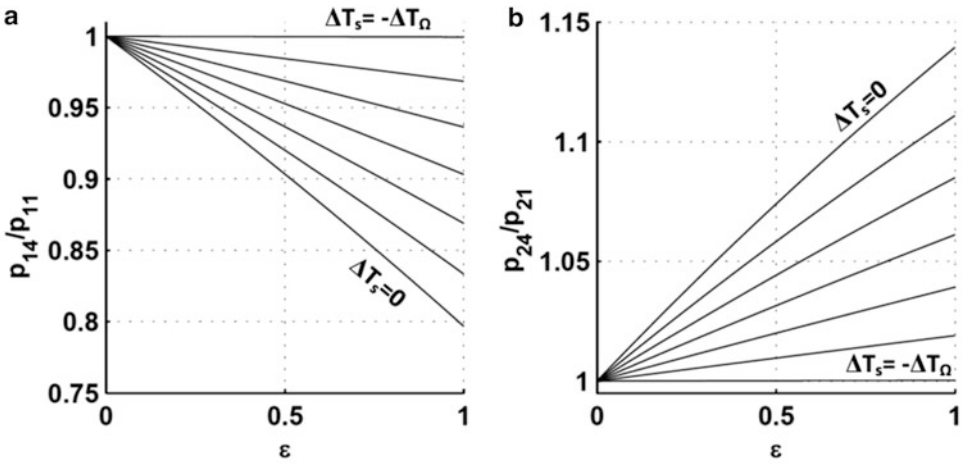


Fig. 4 Outlet-inlet pressure ratio resulting from imbalance between hydrostatic fluid columns, for $T_{11} = 263.15\,K$, $\Omega H = 360\,m/s$, $\Delta T_s \in [-\Delta T_\Omega, 0]$ in steps of 10 K as a function of counter-flowing internal heat exchange efficiency ε for: a) Argon cooling cycle, b) air heating cycle

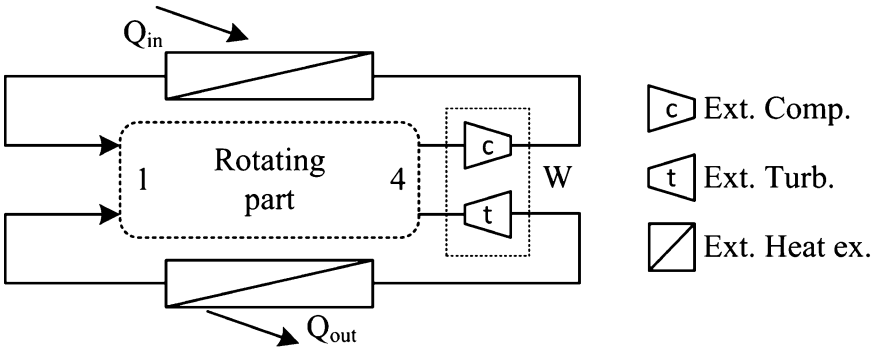


Fig. 5 Full system with external components

For the full system setup, external components can be placed in different locations, but generally, the best is to perform compression at low temperature, and use a turbine at high temperatures. The compressor and turbine is then placed immediately after position 4 before the external ambient heat exchanger, depicted in Fig. 5. The pressure ratio for the compressor and turbine is calculated to be sufficient to bring the pressure back to initial pressure. $(p\_4/p\_1)(p\_5/p\_4) \approx 1$, giving a (necessary) compressor and (achievable) turbine work of

$$|W_c| = \frac{\varepsilon q_{max}}{\eta_c}\left[\frac{\Delta T_1}{T_{11} + \Delta T_1}\right] \tag{13}$$

$$|W_t| = \eta_t \varepsilon q_{max}\left[\frac{\Delta T_2}{T_{21} + \Delta T_2}\right] \tag{14}$$

And an input/output heat, for restoring temperature to ambient temperature:

$$Q_{in} = C_1(T_{11} - T_{15}) = \varepsilon q_{max}\left[\frac{T_{11} + \Delta T_1(1 - \eta_c^{-1})}{T_{11} + \Delta T_1}\right] \tag{15}$$

$$\begin{aligned} Q_{out} &= C_2(T_{21} - T_{25}) \\ &= -\varepsilon q_{max}\left[\frac{T_{21} + \Delta T_2(1 - \eta_t)}{T_{21} + \Delta T_2}\right] \end{aligned} \tag{16}$$

This together giving a (for the sake of argument) frictionless Coefficient Of Performance (COP) for heating of:

$$COP_h = \frac{Q_{out}}{|W_c| - |W_t|} = \frac{1 - \eta_t\left(\frac{\Delta T_2}{T_{21}+\Delta T_2}\right)}{\frac{1}{\eta_c}\left(\frac{\Delta T_1}{T_{11}+\Delta T_1}\right) - \eta_t\left(\frac{\Delta T_2}{T_{21}+\Delta T_2}\right)} \tag{17}$$

The temperature span in which this system is theoretically capable of functioning as a heat pump is $\Delta T_s + \Delta T_\Omega > 0$, when the temperature generation from rotation is able to overcome the source temperature difference. The turbine and compressor efficiency determines the deviation of the maximum theoretical COP (in the point where $\Delta T_s + \Delta T_\Omega = 0$) from the Carnot limit. For optimization of output heat versus COP, plots on the form of Fig. 6ab can be used to determine the optimal position of tangential velocity.

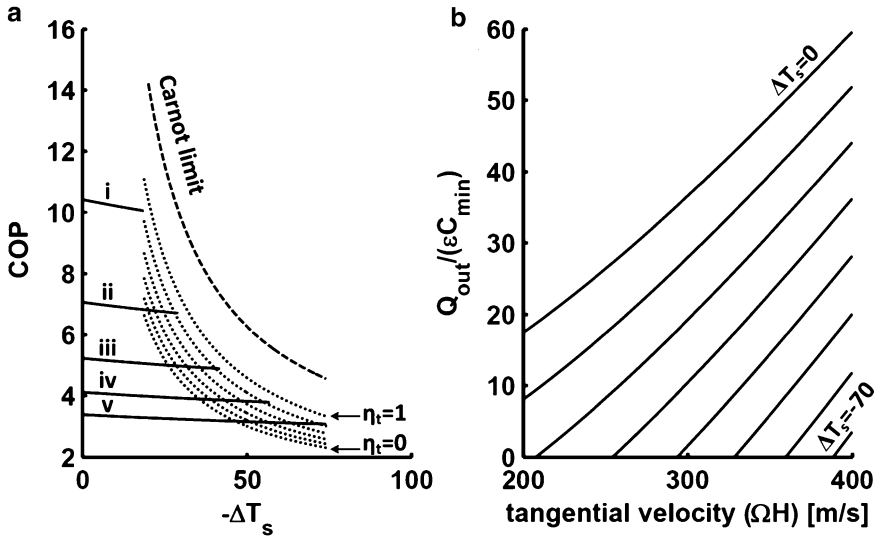


Fig. 6 (**a**) Theoretical heating COP as a function of source temperature difference for $T_{11} = 263.15K$ with $\eta_c = 0.85$, for Argon and Air. Linetypes: Dashed line - Carnot limit. Dotted lines - correspond to COP values for $\eta_t \in [0, 1]$ in steps of 0.2 at the point of $\Delta T_s + \Delta T_\Omega = 0$. Continuous lines – COP value for lines of constant tangential velocity as a function of ΔT_s (i=200, ii=250, iii=300, iv=350, v=400). (**b**) $Q_{out}/\varepsilon C_{min}$, theoretical achievable output heat as a function of tangential velocity for lines of constant source temperature difference

4 Conclusion

The whole system is two separated closed gas loops with different working fluids connected only through a heat exchanger. This can be described as a heat exchanger serving as a connection between a Brayton cycle and reversed Brayton cycle, where compression is partially driven by rotation and partially by external compression. By shifting the entropy axis the two cycles can be drawn in the same plot, Fig. 7 (idealized) where we can see that the heating cycle cancels part of the cooling cycle, leaving the bottom half as the effective cycle. For this system work is performed in the cooling cycle from 4-5 to drive the gas allowing it to receive heat from 5-1 and transfer heat to the heat cycle from 2-3. The heating cycle receives heat from the cooling cycle in 2'-3', work is extracted from 4'-5' to alleviate some of the compressor work (in 4-5) and heat is released to the environment from 5'-1'.

Acknowledgement Supported by "Regionalt bedriftsprosjekt fra Oslofjordfondet" in Norway, for project number: 226001, "Rotobooster and RotoHeatPump, phase 2 (OFF-2)"

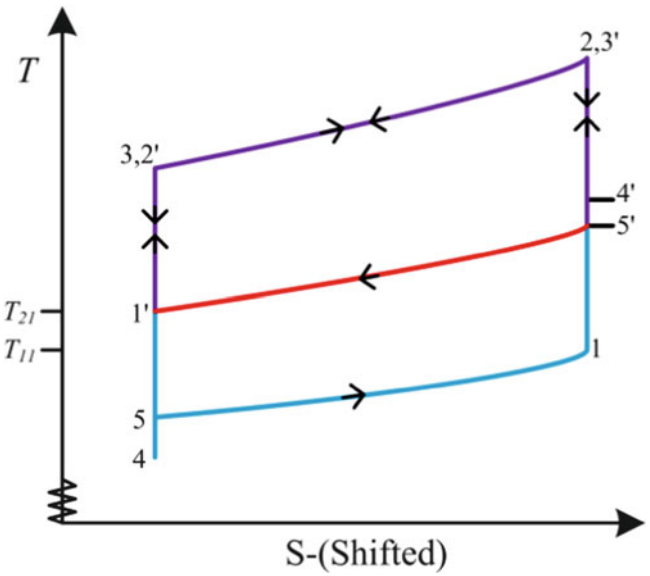


Fig. 7 TS-diagram of the two cycles plotted on top of each other without external intermediate heat exchanger and all other efficiencies set to 100 %, with equal heat capacity rates for both cycles. Cooling cycle 1-5, heating cycle 1′-5′

References

1. Chua, K.J., S.K. Chou, and W.M. Yang, *Advances in heat pump systems: A review*. Applied Energy, 2010. **87**(12): p. 3611–3624.
2. Sarkar, J., *Ejector enhanced vapor compression refrigeration and heat pump systems—A review*. Renewable and Sustainable Energy Reviews, 2012. **16**(9): p. 6647–6659.
3. Self, S.J., B.V. Reddy, and M.A. Rosen, *Geothermal heat pump systems: Status review and comparison with other heating options*. Applied Energy, 2013. **101**(0): p. 341–348.
4. Eslami-nejad, P. and M. Bernier, *Coupling of geothermal heat pumps with thermal solar collectors using double U-tube boreholes with two independent circuits*. Applied Thermal Engineering, 2011. **31**(14–15): p. 3066–3077.
5. Hepbasli, A., et al., *A review of gas engine driven heat pumps (GEHPs) for residential and industrial applications*. Renewable and Sustainable Energy Reviews, 2009. **13**(1): p. 85–99.
6. Kundu, P.K. and I.M. Cohen, *Fluid mechanics 4th ed*. 4 ed, 2008.

7. Childs, P.R.N., *Rotating Flow*, 2011, Oxford: Butterworth-Heinemann. 389.
8. Marshall, J., et al., *Hydrostatic, quasi-hydrostatic, and nonhydrostatic ocean modeling*. Journal of Geophysical Research, 1997. **102**(C3): p. 5733–5752.
9. Kuppan, T., *Heat exchanger design handbook*. Mechanical engineering, 2000, New York: Marcel Dekker.
10. Incropera, F.P. and D.P. DeWitt, *Fundamentals of heat and mass transfer sixth ed*., 2007, Hoboken NJ: John Wiley.
11. Prausnitz, J.M., *Isentropic Compression of Nonideal gases*. Industrial and engineering chemistry, 1955.
12. Poling, B.E., J.M. Prausnitz, and J.P. O'Connell, *The Properties of Gases and Liquids*. 5 ed, 2001: McGraw-Hill.

Experimental Investigation of Developing Spray Boiling on a Flat Flake Surface with Constant Heat Flux

Zhongyuan Shi and Tao Dong

1 Introduction

Spray evaporation cooling is to deliver liquid droplet, of which the diameter is usually stated in micrometers, to the heated surface where the boiling droplets works as heat sink with significantly high heat flux, to which the field synergy [1] of spray droplet velocity and temperature gradient is believed to contribute. This technology has been widely utilized in many applications, e.g. metallurgy, machining, chemical industry, aerospace engineering, nuclear power plant safety, etc.

Konpchikov [2] is among the first authors who conducted experiments on spray cooling. It was reported that the critical heat flux is more than four times of that from pool boiling (up to 500 W/cm^2 with a superheat of 20°C), pressurized water atomized when leaving the nozzle outlet, of which the diameter is 0.02-0.2mm. Toda [3][4] proposed a curve of wall superheat vs. heat flux by fitting the experimental data. It was claimed that the spray droplets with a mean diameter of 117μm (estimated) and a mean velocity of 72.4m/s (estimated) impinged onto a circular wall surface of which the diameter and superheat are 15mm and 50°C respectively, yielding a heat flux of 400W/cm^2. The critical heat flux rises with the increasing subcooled temperature, spray mass flow rate and droplet velocity individually. Bonacinna [5] investigated the spray cooling characteristics under low flow rate distribution density and low excess temperature. When the superheat is 35°C and only 19% of the heated surface is covered by spray droplets, the heat flux is 215W/cm^2. In the experiment, the droplet diameter is reported as 89.5μm with a velocity of 2m/s. Also, the droplets tend not to rebound when its impacting velocity is low. Monde [6] analyzed the effect of spray flow rate and droplet velocity. Bubbles get formed on the heated surface with splashing droplets while the heat flux increases with spray flow rate. Tilton and Pais [7] claimed a much higher critical heat flux of 1100W/cm^2 compared to all the previously reported values when the superheat is 60°C, laser granulometry applied in the investigation. Yao and Choi [8] reported that the heat flux increases with spray mass flow rate with a droplet velocity range of 2.72-5.84m/s and a droplet diameter range of 0.407-0.530mm. According to their experimental result, ,the effect of impacting velocity of droplets on heat transfer is not significant within the film boiling region whereas in the nucleate and the transitional boiling region, the impact of impinging velocity is considerable. Using the same spray system and test piece, Choi and Yao [9] further explored the effect of spray direction on heat transfer characteristics. In the film boiling region, vertical spray (VS) yielded obviously higher heat transfer coefficient than that from horizontal spray (HS), which is caused by the secondary impact of the rebounded droplets. However, in the transitional boiling region, HS resulted in higher heat transfer coefficient, which is probably due to the fact that it is easier for bubbles to get detached from the test piece surface with HS. Yang [10] used gas-liquid mixing nozzle in the experimental study. The heat transfer coefficient was found increasing with spray flow rate. Increasing gas velocity within the nozzle enhances the heat transfer by generating droplets of smaller size and thinner boundary layer.

In boiling regions, surface condition is one of the important factors to the distribution and number of boiling nuclei. It is believed that rough surface facilitates the generation of boiling nuclei and therefore the heat transfer [11]. On the other hand, surfaces that cannot get wetted easily is also believed to be helpful to the formation of boiling nuclei, which led to higher heat flux when the wall temperature is between 100°C and the temperature corresponding to critical heat flux [12]. Nevertheless, no consensus has been reached toward the effect of surface roughness. Sehmbey [12]

Z. Shi • T. Dong (✉)
Norwegian Center of Expertise on Micro- and Nano Technologies, Department of Micro and Nano Systems Technology (IMST), Faculty of Technology and Maritime Sciences (TekMar), Buskerud and Vestfold University College (HBV), Borre N-3199, Norway
e-mail: Zhongyuan.Shi@hbv.no; Tao.Dong@hbv.no

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_16,

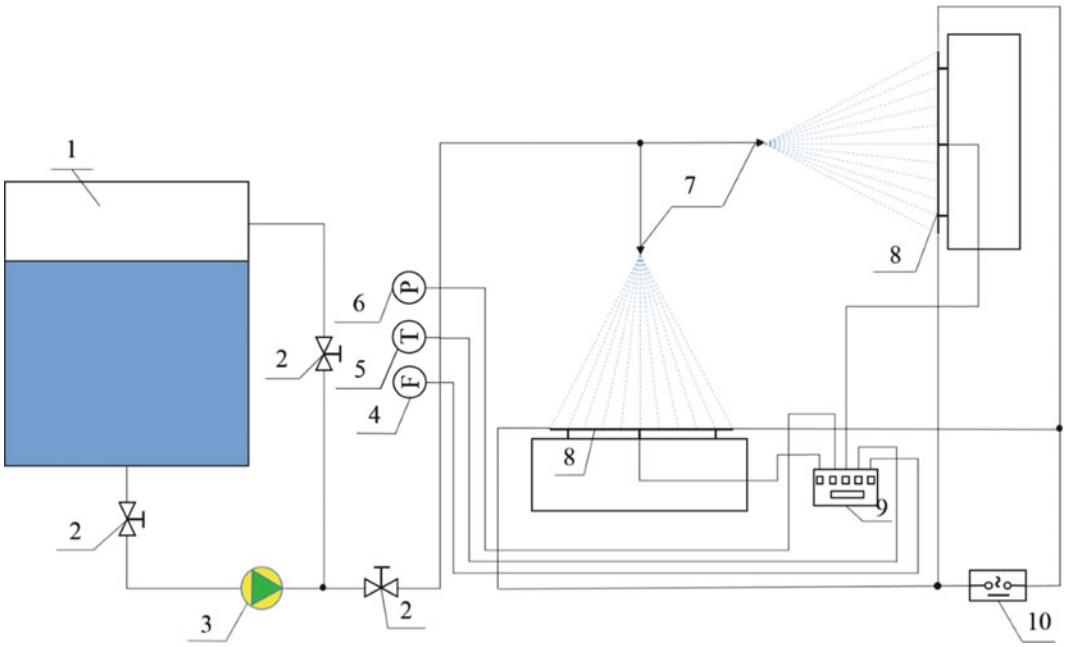


Fig. 1 Schematic representation of experiment system with 1 – water tank, 2 – needle valve, 3 – centrifugal pump, 4 – turbine flowmeter, 5 – platinum resistance thermometer, 6 – pressure gauge, 7. solid cone spray nozzle, 8. test piece – a stainless steel flake, 9 – multimeter, 10 – electric power supply

reported an apparent enhancement on heat transfer coefficient for smoother surfaces (averaged roughness less than 0.1μm) with gas-liquid mixing nozzle. But for normal pressure nozzle, rougher surfaces resulted in higher heat flux. Bernardin [13] also found that rougher surfaces intensifies the heat transfer in nucleate boiling and part of the transitional boiling region.

Most of the relevant work are presented concerning established boiling regions, i.e. nucleate boiling, transitional boiling and film boiling whereas cases for subcooled or developing nucleate boiling are rarely seen in published literature. In the present work, the effect of such premature boiling will be investigated in regard of both horizontal and vertical spray droplets impacting on a flat flake with constant heat fluxes. A new data reduction method will be proposed for a more reasonable interpretation of the heat transfer performance variation in this region.

2 Test Setup

Fig. 1 shows the test setup mainly consisting of the piping system, the electric power for heating the test piece, the data acquisition set and the support frame with the distance from nozzle exit to the test piece surface (nozzle-to-surface distance, L) adjustable. The connecting line between the nozzle exit and the geometric centroid of the test piece is normal to the droplet-impacted surface plane of the test piece. The spray fluid, i.e. water in the present study, is delivered to the nozzle by a centrifugal pump with high head and low flow rate. For the spray flow rate (V) to be tuned more precisely, two needle valves are installed in the system, one on the bypass pipeline at the outlet of the pump, the other on the main branch leading to the nozzle.

The adjustable electric power is connected to the brass poles fixed at two ends of the test piece (170mm× 20mm×0.05mm, stainless flake), together with the gel layer for thermal insulation on the back side of the test piece to ensure the controllable heat generation within the test piece is removed by impacting droplets.

Since the spray flow rate is not always uniformly distributed within the cross sectional plane of the spray jet [14], two thermocouples are fixed on the back side of the test piece, aligned with the longitude center axis and circular-symmetric to the geometric centroid where a third thermocouple is fixed. The spray temperature, pressure (p) and flow rate are acquired respectively from the platinum resistance thermometer, pressure gauge and the turbine flow meter before the nozzle.

Before the experimental tests, the supply water temperature in the water tank was adjusted to an appropriate value. To evacuate the stagnant air in the piping system, the pump was working with relatively higher speed for a period of time. Spray mass flow rate was adjusted to a desired value before the test piece got heated by input electric power. To ensure the accuracy, each datum point was obtained by

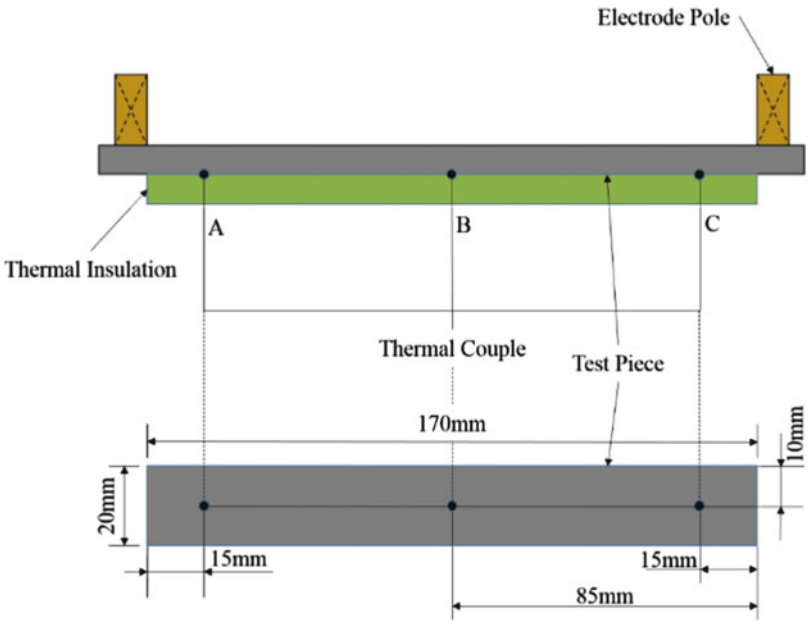


Fig. 2 Local details of test piece with temperature measurement

averaging auto-acquisition scans at steady-state. The multimeter was connected to a computer program for data logging and reduction which is discussed below.

3 Data Reduction

As aforementioned the test piece heated by the input electric power will be regarded as a heat source with constant wall flux [15] which yields

$$q = \frac{Q}{A} \tag{1}$$

in which Q (W) – total power input to the test piece, A (m^2) –area of the test piece surface, q (W/m^2) – wall heat flux on the test piece surface, note that test piece surface hereinafter refers to that in direct contact with incoming droplets.

The heat transfer coefficient is thereafter obtained by,

$$\begin{cases} h_{sp} = \dfrac{q}{T_w - T_f}, \; for \; T_w \le T_s \\ h_b = \dfrac{q}{T_w - T_s}, \; for \; T_w \ge T_s \end{cases} \tag{2}$$

where h_{sp} (W/m^2K) and h_b (W/m^2K) represent the heat transfer coefficient for single phase impinging convection (SPIC) and boiling while T_w, T_f and T_s are for temperature of the test piece surface, spray fluid temperature and saturation temperature corresponding to the ambient pressure in which spray impinging evaporation takes place, respectively.

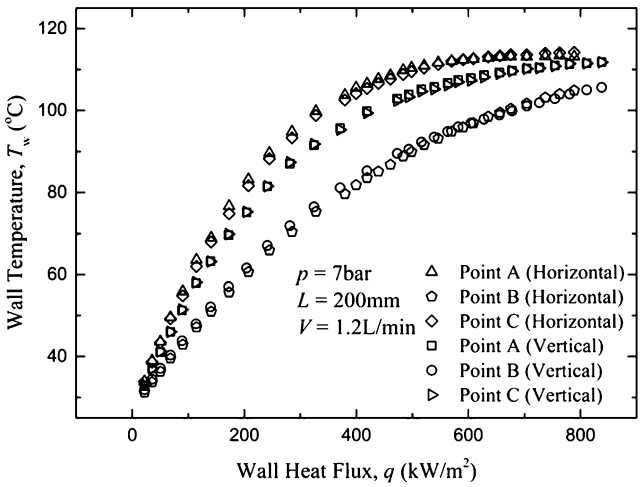


Fig. 3 Variation of wall heat flux with temperature at different locations

4 Result and Discussion

Fig. 3 shows the performance of spray cooling with a volumetric flow rate of 1.2L/min and a nozzle-to-surface distance of 200mm. The temperature at Point B is the lowest compared by those from Point A and C, which overlap with each other, given the same wall heat flux. It is therefore presumably inferred that the local mass flux is larger at the location near to the spray jet axis (SJA) and decreases symmetrically in the radial direction perpendicular to SJA.

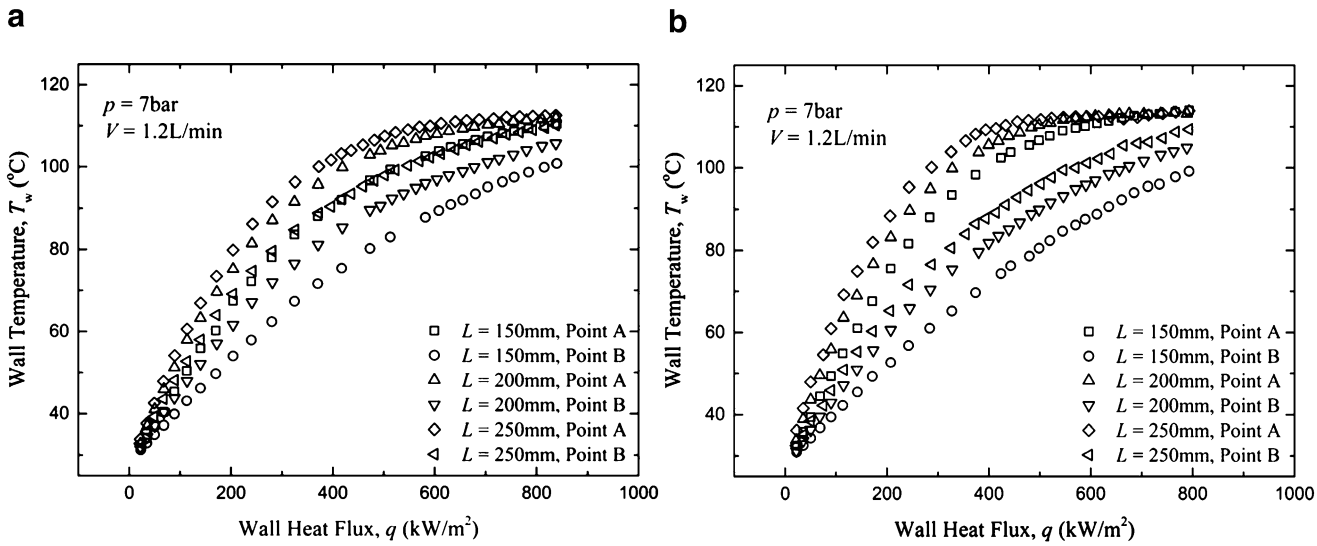


Fig. 4 Variation of wall heat flux with temperature under different nozzle-to-surface distances for (**a**) VS cases and (**b**) HS cases

The slope of wall temperature starts to decrease significantly with rising heat flux in a wall temperature region near to the saturation point, i.e. 100°C in the present study especially for locations (Point A and C) with less impinging mass flux, whereas for the near-to-SJA region (Point B) the slop decreases more smoothly and tends to continue decreasing even after the counterpart at Point A or C stalls (temperature reaching a constant value). This is probably because of the trade-off between two major mechanisms dominating heat transfer, the SPIC and the boiling. When the local spray flux is low, the effect of SPIC gets dominated by that of boiling more "quickly" with the increasing wall heat flux.

Compared to cases for HS, the VS cases yielded higher heat flux at Point A and C where the local spray mass flux is lower, indicating a possible enhancement from rebounded droplet [9]. For Point B, no obvious difference is observed between the VS and the HS cases. On the other hand, a smoother slope decreasing can be identified for all the VS cases (Point B included), again the effect of SPIC is supposedly retarding the transition to boiling dominated regime.

Fig. 4(b) shows the effect of spray distance on spray cooling performance. Since the temperature profile at Point A simply overlap with that at Point C (like in Fig. 4(a)), only one of these two profiles will be involved in the following discussion (see Fig. 4(b)) to avoid confusion. For a certain location under measurement, either point A or B, the local surface temperature (LST) increases with increasing nozzle-to-surface distance whereas the increasing magnitude (IM) is not uniform. Before the temperature measured at a certain point reaches the saturation temperature, the IM increases with increasing heat flux at different nozzle-to-surface distances. When LST goes beyond the saturation

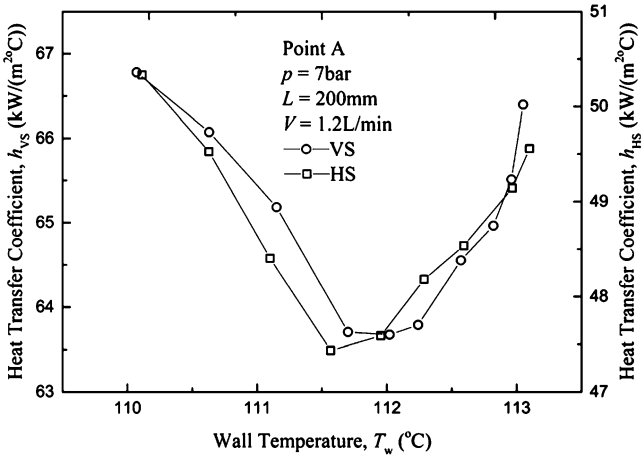


Fig. 5 Variation of heat transfer coefficient with wall temperature in the boiling region

temperature, the IM decreases with increasing heat flux. This trend is more obvious in the location (Point A) with less impinging mass flux [18], which is reasonable compared to the location (Point B) where the retarding effect of SPIC is significant and the IM decreases slower.

A similar trend is observed for the HS cases in Fig. 4(b). Nevertheless, the effect of SPIC is more dominating in horizontal cases since the temperature difference between locations with different impinging mass flux is larger compared to that of the VS cases given all other conditions unchanged.

The heat transfer coefficient in the boiling region, h_b calculated by Eq. (2) with $T_w > 110$°C is shown in Fig. 5. A decreasing-increasing trend with a local minimum (at

$T_w = 111.6°C$ and 112°C for HS and VS respectively) is observed for h_b with increasing T_w. However, according to the conventional boiling heat transfer theory, it is expected that the heat transfer keeps getting intensified with an increasing wall heat flux before the critical heat flux is reached, which indicates that the conventional data reduction method for heat transfer coefficient in Eq. 2 may need modification.

The existing report [16][17] on subcooled pool boiling involves two major characteristic turning points. One is the onset of nucleate boiling(ONB) corresponding to a wall temperature of T_{ONB} below which heat transfer is principally determined by single phase convection. Bubbles start to form on and get attached to the wall surface when the wall temperature keeps rising to T_{ONB} and beyond, accompanied by a substantial increase in wall heat flux and heat transfer coefficient. The other is the onset of fully developed boiling (OFDB) marked by a wall temperature of T_{OFDB} beyond which bubbles get detached from the heated wall surface. The concept can be analogously adopted for spray impinging evaporation.

The wall heat flux is comprised of two components induced by SPIC (q_{SPIC}) and nucleate boiling (q_{nb}) individually [18] where

$$q_{SPIC} = h_0(T_w - T_f)\left(\frac{\mu_f}{\mu_w}\right)^n \tag{3}$$

where h_0 and n are constants from regression, μ_f and μ_w are dynamic viscosity of water based on the spray fluid temperature and the wall temperature respectively and

$$q_{nb} = q - q_{SPIC}. \tag{4}$$

Following Eq. (3) and Eq. (4), the SPIC and the nucleate boiling heat transfer coefficient are redefined as

$$h_{SPIC} = h_0\left(\frac{\mu_f}{\mu_w}\right)^n \tag{5}$$

and

$$h_{nb} = \frac{q_{nb}}{(T_w - T_s)} \tag{6}$$

respectively.

Fig. 6 shows the comparison of the experimental wall heat flux and the calculated SPIC wall heat flux. For both HS and VS cases, the experimental heat flux the deviates from the calculated value to a considerably higher magnitude starting from approximately T_w=100°C, which resembles the ONB in pool boiling. Back to Fig. 5, heat transfer coefficient decreases before the local minimum, which is

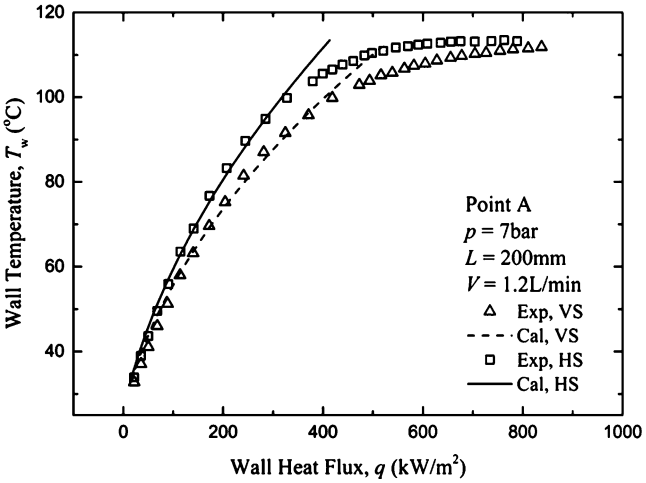


Fig. 6 Boiling curve, experimental vs. calculated from regression

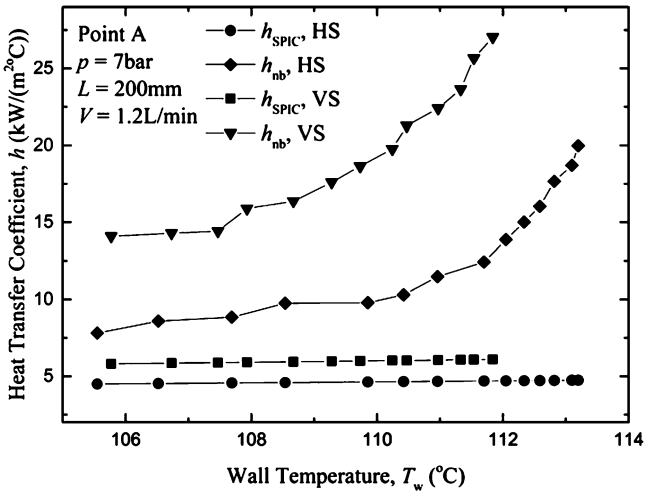


Fig. 7 Heat transfer performance in transition to fully developed boiling

probably caused by the fact that attached bubbles are limiting the space for single phase convection in the near-to-wall region within the liquid film layer on the heated surface. T_w increasing after the local minimum, the heat transfer coefficient begins to increase which is probably due to the enhanced single phase convection by the "stirring" effect of bubbles' motion leaving the heated surface [16] which rationalizes marking the local minimum as OFDB.

As shown in Fig. 7, the nucleate boiling component of the heat transfer coefficient h_{nb} increases monotonically with increasing T_w, which avoids the anomalously posed trend in Fig. 5. On the other hand, the SPIC component also increases with ascending T_w but a negligible slope compared to its counterpart.

When the wall heat flux reached a certain value in correspondence with that greater than the rightmost wall temperatures for HS or VS cases individually, the test piece would be burnt into two parts from the region close to the centerline normal to the longitude direction, which is similar to the critical point of pool boiling that droplets evaporated almost immediately after contacting the heated surface before the next batch of droplets arrives so that any effort to achieve a higher wall heat flux will result in a growth spurt of temperature and burn the test piece apart.

However, the difference is for pool boiling, the heated surface usually get burnt after a significant heat transfer coefficient decrease which is not observed in the present investigation, which might be caused by the fact that the thin steel flake get burnt too fast (within the minimum increasing step of power input to the test piece) before any data could be collected.

5 Conclusion

In the present study, the heat transfer performance of water spray impinging on a heated steel flake is investigated with respect to the effect of spray direction, nozzle-to-surface distance and wall heat flux. The VS cases yield higher heat transfer coefficient than that of the HS cases. The nozzle-to-surface distance has influence on the local impinging mass flux, which affects the heat transfer coefficient considerably. The heat transfer performance is dominated by SPIC and nucleate boiling sequentially with increasing wall heal flux.

Conventional data reduction method yields an irrational heat transfer variation with increasing wall temperature when the enhancement mechanism is moving forward to the fully developed boiling region in analogy to pool boiling. The proposed decomposition of wall heat flux separates the nucleate boiling heat transfer coefficient from the conventional all-in-one heat transfer coefficient based on which a physically more credible variation trend is achieved.

Acknowledgment The present work is primarily supported by Forskningsradet Nærings-Ph.d (Project No.: 231952), Scientific Research Project (Project No. A2620110012), VRI bedriftsmidler (Project No.: 235042) and Regionale Forskningsfond Oslofjordfondet (Project No.: 235809). The Research Council of Norway is acknowledged for the support through the Norwegian Micro- and Nano-Fabrication Facility, NorFab (197411/V30). The RotoBoost AS which initiates and facilitates the present work is hereby acknowledged.

References

1. Guo, Z. Y., Tao, W. Q., Shah, R. K.: The Field Synergy (Coordination) Principle and its Applications in Enhancing Single Phase Convective Heat Transfer. Int J Heat Mass Transfer 48(9), 1797–1807 (2005)
2. Kopchikov, I. A., Voronin, G. I., Kolach, T. A., Labuntsov, D. A., Lebedev, P. D.: Liquid Boiling in a Thin Film. Int. J. Heat Mass Transfer 12(7), 791–796 (1969)
3. Toda, S.: A Study of Mist Cooling : 1st Report, Experimental Investigations on Mist Cooling by Mist Flow Sprayed Vertically on Small and Flat Plates Heated at High Temperatures. Trans. JSME 38(307), 581–588 (1972)
4. Toda, S.: A Study of Mist Cooling (2nd Report : Theory of Mist Cooling and its Fundamental Experiments). Heat Transfer Japanese Research 3, 1–44 (1974)
5. Bonacina, C., Comini, G., Del Giudice, S.: Dropwise Evaporation. J. Heat Transfer 101(3), 441–446 (1979)
6. Monde, M., Inoue, T.: Critical Heat Flux in Saturated Forced Convective Boiling on a Heated Disk With Multiple Impinging Jets. J. Heat Transfer 113(3), 722–727 (Aug 1991)
7. Pais, M., Tilton, D., Chow, L., Mahefkey, E.: High-Heat-Flux, Low-Superheat Evaporative Spray Cooling. In : 27th Aerospace Sciences Meeting (January 1989)
8. Yao, S. C., Choit, K. J.: Heat Transfer Experiments of Mono-Dispersed Vertically Impacting Sprays. Int. J. Multiphase Flow 13 (5), 639–648 (1987)
9. Choi, K. J., Yao, S. C.: Mechanisms of Film Boiling Heat Transfer of Normally Impacting Spray. Int J Heat Mass Transfer 30(2), 311–318 (1987)
10. Yang, J. D., Pais, M. R., Chow, L. C.: High Heat Flux Spray Cooling. In : Proc. SPIE 1739, High Heat Flux Engineering (1993)
11. Rohsenow, W. M., Hartnett, J. P., Cho, Y. I.: Handbook of Heat Transfer 3rd edn. McGraw-Hill, New York (1998)
12. Sehmbey, S. M., Pais, R. M., Chow, C. L.: Effect of surface material properties and surface characteristics in evaporative spray cooling. In : 5th Joint Thermophysics and Heat Transfer Conference, AIAA and ASME, Seattle, pp.505–512 (1990)
13. Bernardin, J. D., Stebbins, C. J., Mudawar, I.: Effects of Surface Roughness on Water Droplet Impact History and Heat Transfer Regimes. Int J Heat Mass Transfer 40(1), 73–88 (1996)
14. Yuan, W. X., Liu, C. X., Zhao, X. S., Zeng, T.: Development and Application of System for Measurement on the Spray Cooling Feature of Nozzle (in Chinese). Steelmaking 20(1), 36–39 (May 2004)
15. Lei, S. Y.: Investigation on Spray Cooling Mechanism and Nonboiling Model (in Chinese). (2002) National Natural Science Foundation of China (No. 50176025).
16. Bergman, T. L., Lavine, A. S., Incropera, F. P., DeWitt, D. P.: Fundamentals of Heat and Mass Transfer 7th edn. Wiley, Hoboken, USA (2011)
17. Maprelian, E., Castro, A. A. d., Ting, D. K. S.: Onset of Nucleate Boiling and Onset of Fully Developed Subcooled Boiling Detection Using Pressure Transducers Signals Spectral Analysis. In : 15th Brazilian Congress of Mechanical Engineering, São Paulo, Brazil (1999)
18. Chen, W. K.: Research on Efficient Spray Impinging Evaporation Heat Transfer Technology under Small Temperature Differences (in Chinese). Master Thesis, University of Shanghai for Science and Technology, Shanghai (2007)

Life Cycle Assessment of Circulating Fluidized Bed Combustion with CO_2 Post-Combustion Capture

Cristian Dinca, Adrian Badea, Vladimir Tanasiev, and Horia Necula

1 Introduction

Today, climate change is one of the most discussed topics worldwide. The main greenhouse gases with a direct impact on climate change are: CO_2, CH_4 and N_2O. Corresponding to the CML methodology proposed by Leiden University in the Netherlands, the contribution of greenhouse gases vary depending on the analyzed study period. Thus, for a period of 100 years (often chosen in such studies), the contribution of greenhouse gases relates to CO_2 as the reference gas as follows [1]:

- $GWP\_CO_2 = 1kg\ CO_2\ eq./kg\ CO_2$;
- $GWP\_CH_4 = 21kg\ CO_2\ eq./kg\ CH_4$;
- $GWP\_N_2O = 310kg\ CO_2\ eq./kg\ N_2O$

Although, qualitatively, the emission of N_2O, respectively CH_4 have a contribution of 310, respective 21 times higher than CO_2, in terms of their quantitative contribution it becomes insignificant compared to that of the carbon dioxide. In this context, by taking into account simultaneously both the quality and the quantity factor, Equation (1) can be used:

$$GWP = \sum_{i=1}^{n} GWP_i \cdot m_i \qquad (1)$$

where: GWP_i – represents the contribution of pollutant "i" to climate change, values presented above; m_i – mass of pollutant "i" inventoried in the combustion process, in kg pollutant i.

C. Dinca (✉) • V. Tanasiev • H. Necula
Politehnica University of Bucharest, Splaiul Independentei, 313, Romania 060042, Romania
e-mail: crisflor75@yahoo.com

A. Badea
Politehnica University of Bucharest, Splaiul Independentei, 313, Romania 060042, Romania

Romanian Academy of Scientists, Bucharest, Romania

N.A. Odeh, (2008) presented in his paper work the contribution to the greenhouse effect of the fossil primary resources on their life cycle. Thus, in the life cycle of coal is generated in the environment around 960 g / kWh CO_2 equivalent in which 870 g of CO_2 per each kWh. Moreover, 855 g / kWh of CO_2 are generated only in the stage of electricity generation (in the combustion process). In other words, the emission of CO_2 is responsible for about 90 % of the total GWP indicator in the coal life cycle. The electricity generation stage is responsible for 89 % of the greenhouse gases emissions generated throughout the life cycle.

Compared with the coal life cycle, in the case of the oil and natural gas life cycle, the value of GWP indicator is 660 respectively 410 g / kWh [2, 3].

Considering these aspects it is obviously that in order to achieve a powerful engineering industry based on coal is needed a development of the actual technologies for converting the chemical energy into electricity and / or heat and for reducing the CO_2 emissions.

2 Life Cycle Assessment Methodology

Life cycle analysis is a methodology that permits the global evaluation of the environmental impact of a product, process, system or subsystem started from raw material extraction to final disposal. By definition, the life cycle analysis consists of four steps: a) the objectives definition and the study boundaries, b) inventory analysis, c) impact analysis, and, d) solutions for reducing the overall environmental impact [4].

In this article, we have analyzed the effects in terms of the environmental impact of the post-combustion CO_2 capture process integration in the circulating fluidized bed combustion technology.

In Figure 1 is presented the schema of the power plant revamped based on circulating fluidized bed combustion with CO_2 capture process. Flue gas treatment includes the retention of dust particles in an electrofilter; flue gas

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_17,

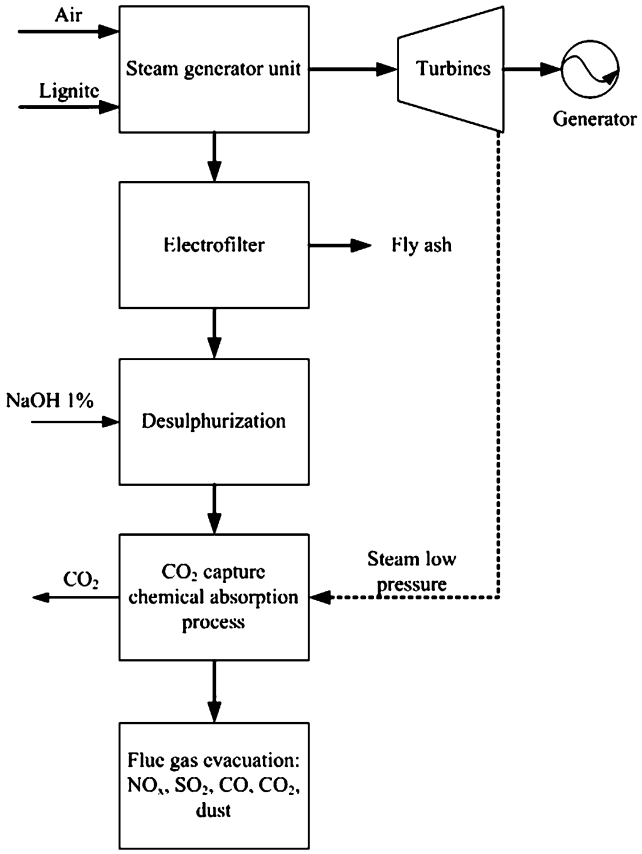


Fig. 1 Schematic diagram of the retrofitted circulating fluidized bed combustion technology

desulphurization using a solution based on NaOH at a concentration of 1 %, and CO_2 separation from flue gases using chemical absorption process based on primary amines – monoethanolamina (MEA).

The revamping of the CFBC power plant includes only the integration of the CO_2 separation process.

Table 1 shows the values of the main parameters that define the existing thermal power plant considered as the reference case.

2.1 Functional unit definition

All emissions and the energy consumption were reported at the functional unit. Functional unit was considered equal to the amount of net annual electricity - 1 TWh.

2.2 Circulating fluidized bed combustion with and without CO_2 post-combustion process

Regarding the combustion process it was considered the following elemental composition of lignite: C = 21.55 %; H = 1.25 %; O = 2.55 %; N = 0.65 %; S = 1 %; W = 36 %; A = 37 %. Using the Mendeleev semi empirical equation it has been determined the low heating value of lignite: 7881 kJ/kg. For excess air exhaust of 1.5, the concentration of CO_2 in flue gas before entering the absorption unit is 12.37 %.

In figures 2 and 3 it is presented the thermal power plant of circulating fluidized bed coal combustion with and without post-combustion CO_2 capture.

In this article it has been taken into account only the generated emissions in the coal life cycle (extraction, treatment, transport and electricity generation), neglecting the emissions generating during the construction and the decommissioning of the power plant.

Consequently, the cases studied in the paper are:

- Case I – Circulating fluidized bed combustion without CO_2 capture (reference case);
- Case II - Circulating fluidized bed combustion with CO_2 capture (revamping case).

For a better understanding of the CO_2 chemical absorption process, Table 2 presents the main parameters.

The optimal operating parameters of the chemical absorption process were established according to the flow rate of flue gases and the ratio L/G. For the ratio L/G of about 1.2 kg_liquid / kg_flue_gases, and a fuel flow rate of 85 kg/s, it results a flow rate of MEA solution of 2 300 kg/s and a MEA concentration in solvent of 30 %.

The efficiency of CO_2 separation process was determined using the Equation (2):

$$\varepsilon_{CO_2} = \frac{CO_{2in} - CO_{2out}}{CO_{2in}} \quad (2)$$

Where: CO_{2in}, CO_{2out} – represents the concentration of CO_2 in the flue gases at the in and out of the absorption unit, in %.

For the chemical solvent regeneration used in the CO_2 separation process it used the steam extracted from the low pressure steam turbine. Therefore, the difference between

Table 1 Main parameters of the CFBC coal power plant

Plant capacity	300 MW	Low heating value of lignite	7.9 MJ/kg
Net electrical energy	1566 GWh	Loading rate	75 %
Overall efficiency	40 %	Yearly coal consumption	1 785 000 tons
Annual life time	5800 h/y	Power plant life	30 years

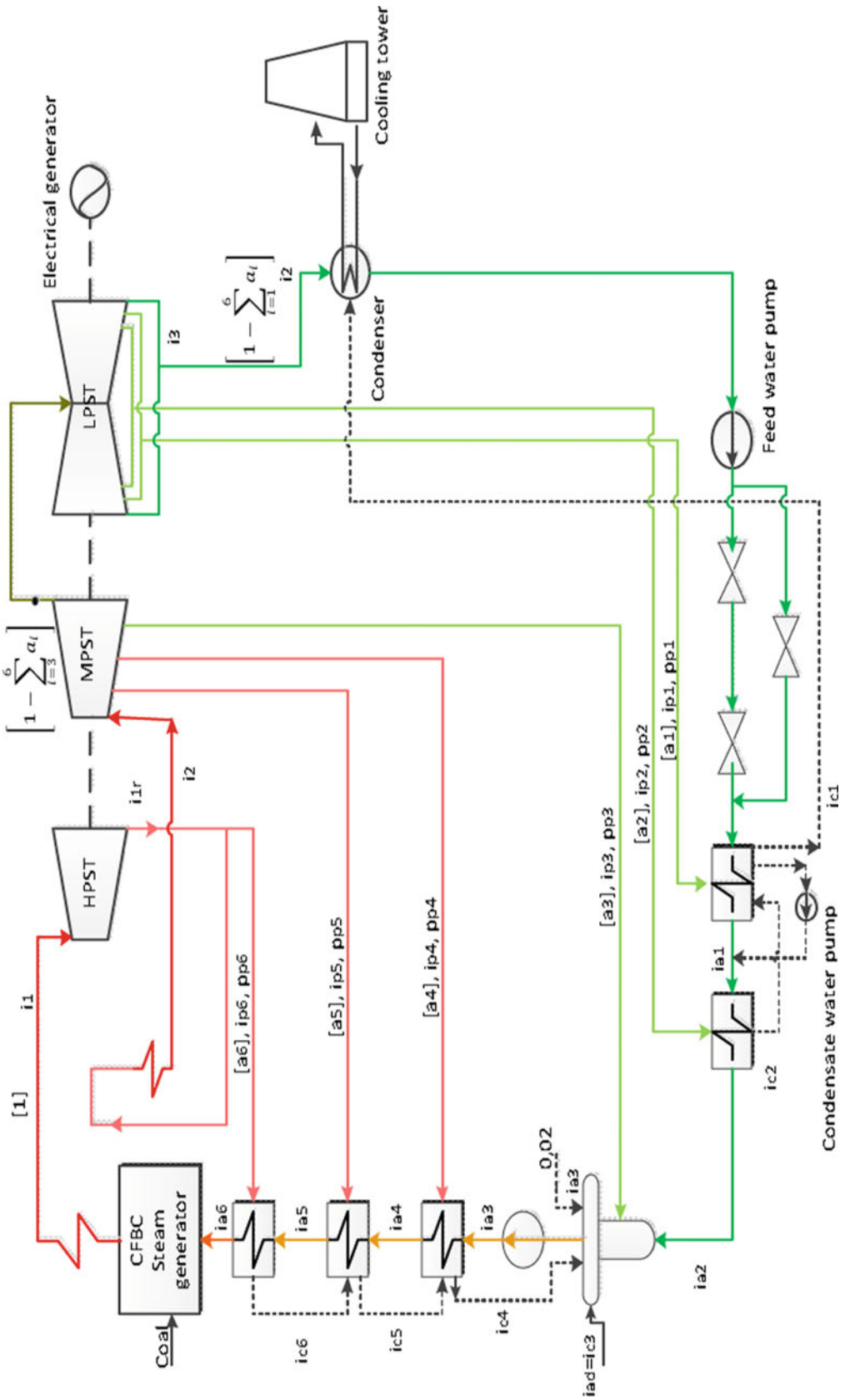


Fig. 2 CFBC technology without CO_2 capture unit [5]

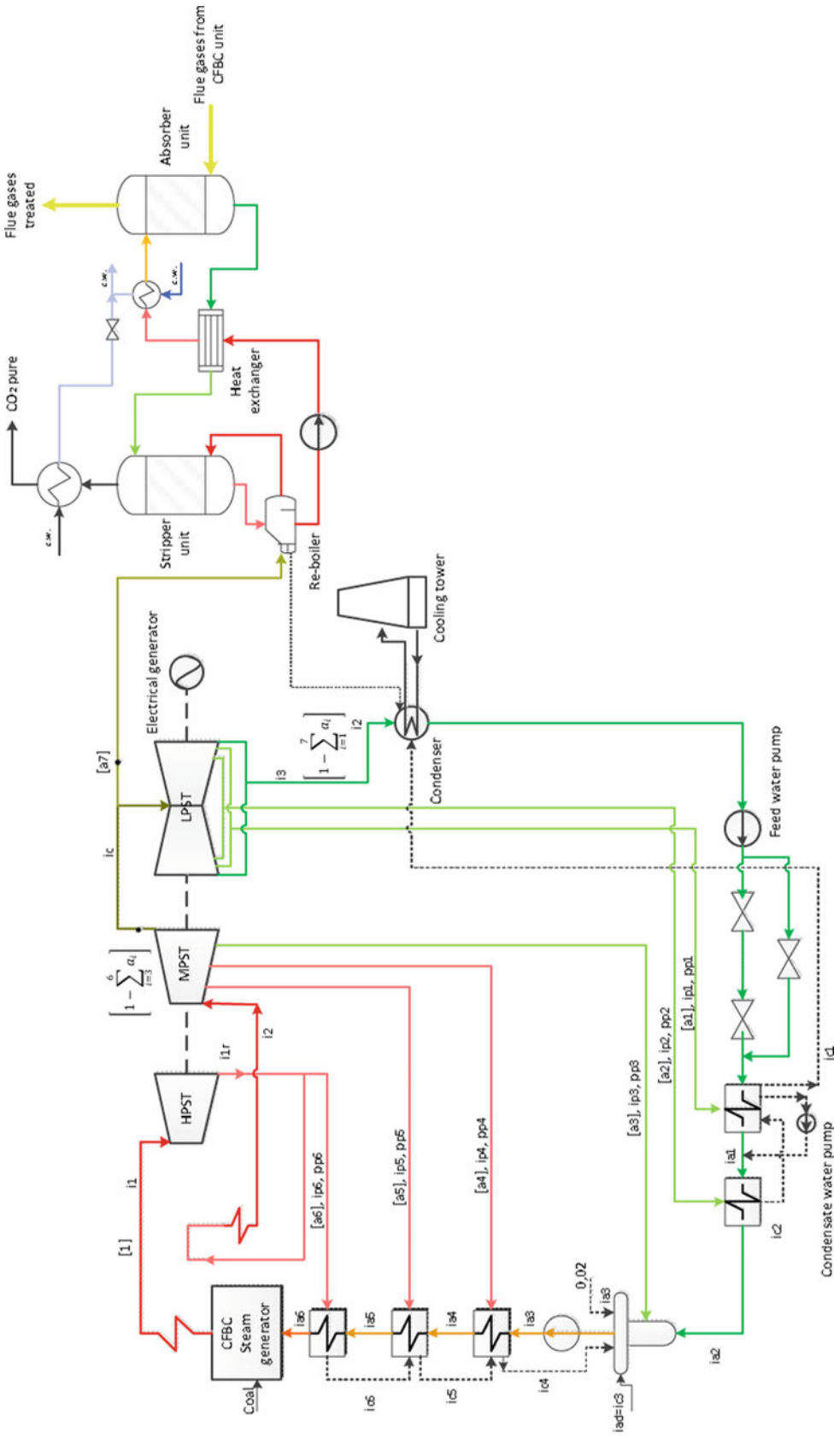


Fig. 3 CFBC technology with CO_2 post-combustion capture process [5]

Table 2 The process parameters for CO_2 separation by chemical absorption

Parameter	Range of variation	Process conditions
Solvent flow rate, kg/s	2 300	Absorber pressure – ct.; Process efficiency – ct.; Amine concentration – ct.
CO_2 partial pressure, atm	~ 0.124	-
Process absorption temperature, °C	31 .. 49	Absorber pressure – ct.; Process efficiency – ct.; Amine concentration – ct.
Process absorption pressure, atm.	1.2	-
L/G rapport, kg_l/kg_fg	~ 1.2	Absorber pressure – ct.; Process efficiency – ct.; Amine concentration – ct.
CO_2 efficiency process, %	~ 85	Absorber pressure – ct.; Amine concentration – ct.
Stripper process temperature, °C	~ 105	Absorber pressure – ct.; Process efficiency – ct.; Amine concentration – ct.
Solvent concentration in solution, % wt.	~ 30	Absorber pressure – ct.; Process efficiency – ct.;

the two cases of mechanical work $(P_{wthc} - P_{wc})$ is determined by the following equation:

$$P_{wthc} - P_{wc} = \dot{m} \cdot a_7 \cdot (i_c - i_3) \tag{3}$$

3 Impact assessment of the CFBC technology. Results and discussion

Comparative analysis of the two cases was performed considering the environmental impact indicators presented in Table 3 and the techno-economic indicators presented in the Table 4 for considering the thermal energy required by the chemical absorption process.

The environmental impact indicators were calculated using the equation (4) where the amount of pollutant according to the functional unit was weighted with its contribution to the environmental impact class. All the pollutants that contribute for each class of environmental impact are presented in Table 3.

$$\mathrm{I} = \sum_{i=1}^{n} I_i \cdot m_i \tag{4}$$

In Figures 4–6 it is presented a comparative analysis between the two cases analyzed considering only emissions that contribute to the greenhouse gases class. The results calculated for the other indicators are presented in Table 4.

Table 3 Environmental impact assessment indicators [6]

Class impact	Emissions/Primary energy consumptions	Reference pollutant	Impact scale
Global warming potential (GWP)	CO_2, CH_4, N_2O	CO_2 – equiv.	Global
Acidification potential (AP)	SO_x, NO_x, HCl, HF, NH_3	SO2 – equiv.	Regional, local
Eutrophication potential (EP)	NO, NO_2, NH_3, PO_4^{3-}	PO_4^{3-} - equiv.	Local
Photo-oxidant formation potential (POCP)	NMHC	C_2H_6 – equiv.	Local
Human toxicity potential (HTP)	Dust, Hg, H_2S, NO_2, NH_3, SO_2	1,4 DCB equiv.	Global, Regional, Local
Abiotic resources depletion potential (ADP)	Coal, natural gas	Antimoniu eqiv.	Global, Regional, Local

Table 4 The impact evaluation of the coal life cycle stage

	Coal life cycle stage							
	Extraction		Treatment		Transport		Power generation	
Impact indicator	Case I	Case II	Case I	Case II	Case I	Case II	Case I	Case II
GWP, ton CO_2 – equiv.	36650	40722	6962	7735	60	66	761246	131306
AP, SO_2 – equiv.	298	331	106	118	0.7	0.7842	4626	4503
EP, ton PO_4^{3-} - equiv.	42	47	19	21	0.08	0.091	188	178
POCP, ton C_2H_6 – equiv.	14	15.5	2.03	2.25	0.165	0.184	197.8	192.4
HTP, ton 1,4 DCB equiv.	83.8	93.11	44.7	49.7	0.857	0.952	15720*	17467*
ADP, ton Antimoniu equiv.	14156	15729	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

* - the values are presented in 1000 tons 1,4 DCB equiv.

Table 5 Technical and economical evaluation of the power plant with and without CO_2 capture

Process Parameter	Power plant without CO_2 capture	Power plant with CO_2 capture
Steam cycle parameters		
Power of steam generator turbine, kW	300,000	255,543
CO_2 removal steam generator output, kW	-	44,457
Heat consumed for main steam, kW	649,265	649,265
CO_2 removal system parameters		
Type of solvent	-	MEA
Solvent concentration, %	-	30
Solvent regeneration energy, GJ/tonneCO_2	-	3.35
Steam flow in the LPST, kg/s	185.02	111.01
Steam extraction flow for MEA regeneration, kg/s	-	74.01
CO_2 generated, kg/s	58.28	-
CO_2 captured, kg/s	-	52.45
CO_2 emissions, kg/s	-	5.83
Plant performance parameter		
Net plant efficiency, %	40	36
Energy penalty, %	-	10
Incremental capital		
Total investment cost, mil €	320.4	387.6
Fixed O&M cost, mil €/year	2.04	2.65
Variable O&M cost, mil €/year	14.64	19.18
CO_2 capture cost, €/tonne	-	53

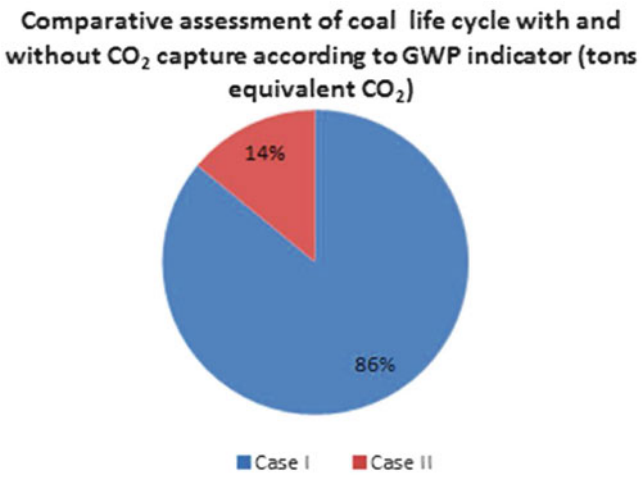


Fig. 4 Comparative assessment of coal life cycle with and without CO_2 capture according to GWP indicator

By extraction of the steam in the low pressure steam turbine (Figure 3), it was reduced the amount of electricity generated considering the same primary energy in the boiler. In the second case, for achieving the functional unit (the amount of electricity produced), additional combustible rate is required. This has led to increased ADP indicator at 15 700 ton antimony equiv. compared with 14 150 ton antimony equiv. obtained in the referential case. Natural resource consumption occurs in the extraction stage, other resources not being consumed during the coal life cycle.

On the other hand, by integrating the chemical absorption process in the CFBC power plant the NO_x and SO_2 emissions were reduced. This is highlighted by the diminishing of the AP and EP impact indicators compared with the referential case.

4 Conclusions

The scope of this paper consisted to analyze the global environmental impact of the coal life cycle used in circulating fluidized bed combustion for energy generation. In this paper two cases were analyzed: circulating fluidized bed combustion with and without carbon dioxide chemical absorption process.

The integration of the CO_2 chemical absorption process in the CFBC technology conducted to reduce eight times the global warming potential comparative to the referential case. In order to obtain a high efficiency of the CO_2 chemical absorption process, a steam flow was extracted from the low pressure steam turbine for amine solvent regeneration. Considering that the steam flow was 74 kg/s, the global efficiency of the power plant was10 % smaller than the referential case.

Beside, by using the primary amine in the chemical absorption process not only the CO_2 is captured but also the NO_x and SO_2. In that case, the environmental indicators acidification (AP) and eutrophication (EP) were reduced with 2.5 % and 5.3 % according to referential case considering only the combustion stage.

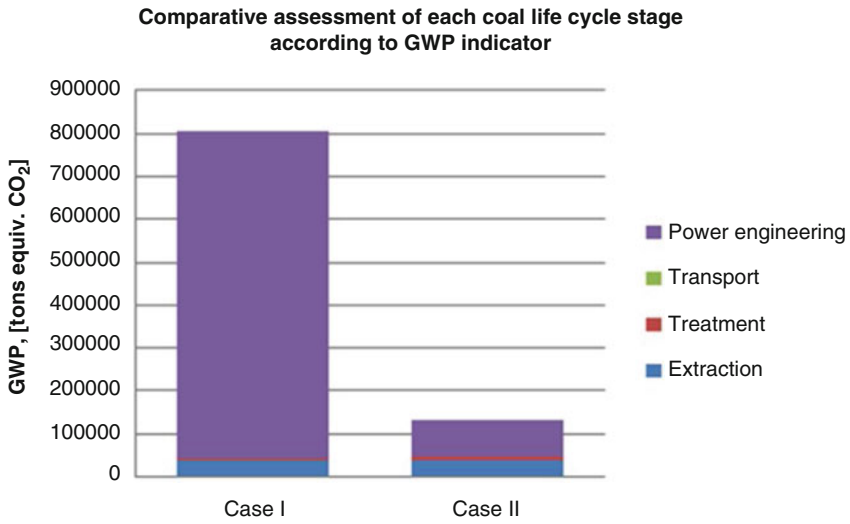


Fig. 5 Comparative assessment of each coal life cycle stage according to GWP indicator

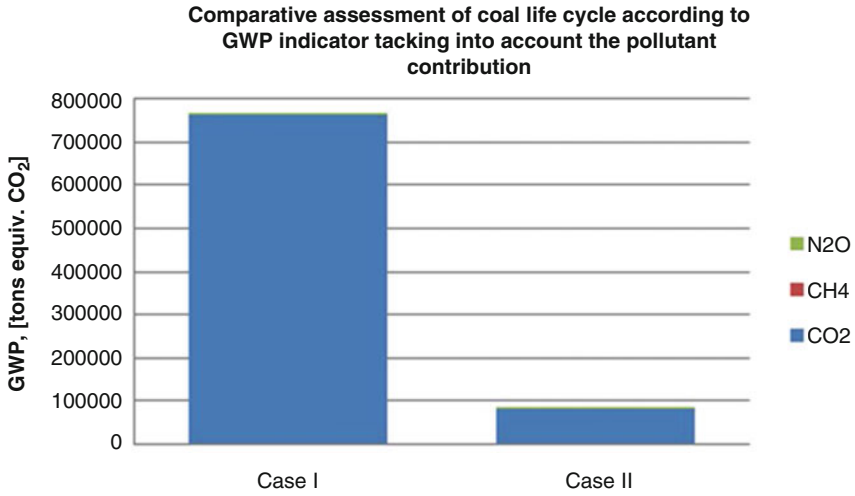


Fig. 6 Comparative assessment of coal life cycle according to GWP indicator tacking into account the pollutant contribution

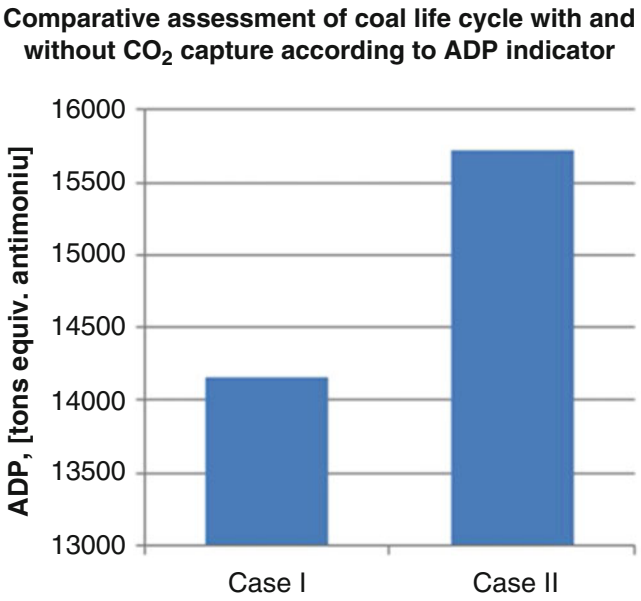


Fig. 7 Comparative assessment of coal life cycle with and without CO_2 capture according to ADP indicator

In terms of economic, the CO_2 chemical absorption process integration in the CFBC technology has led to increase the initial investment with about 21 % compared with the referential case. Also, the maintenance and operation fixed and variable costs are increased with 33 %.

Acknowledgements The study has been funded by the UEFISCDI within the National Project number 38/2012 with the title: „Technical-economic and environmental optimization of CCS technologies integration in power plants based on solid fossil fuel and renewable energy sources (biomass)" – CARBOTECH. Also the work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/134398.

References

1. Dinca, C., Rousseaux, P., Badea, A., A life cycle impact of the natural gas used in the energy sector in Romania. Journal of Cleaner Production, Vol. 15, Issue 15, 2007, 1451 - 1462 pg., ISSN: 0959-6526

2. Odeh, N.A., Cockerill, T., Life cycle analysis of UK coal fired power plants. Energy Conversion and Management, 49, 2008, 212-220
3. Odeh, N.A., Cockerill, T., Life cycle GHG assessment of fossil fuel power plants with carbon capture and storage, Energy policy journal, 36, 2008, 367-380
4. International Organization for Standardization. Environmental management - life cycle assessment - goal and scope definition and inventory analysis [ISO 14041]. Geneva, Switzerland: ISO; 1997.
5. Dinca, C. Technical, environmental and economic assessment of CO_2 absorption chemical process integration in the power plant technologies. Book: Materials and processes for energy: communicating current research and technological developments. Editor: A. Méndez-Vilas, Formatex Research Center, Spain, 2013, (935-945 pp) 11 pg. ISBN: 978-84-939843-7-3
6. Rousseaux P, Apostol T. Valeur environnementale de l'énergie. Lyon, France: INSA de Lyon; 2000.

Suggested Simulation of the First Copper-Chlorine Reactor Step for Solar Hydrogen Generation Process

Samir Moujaes and Mohamed Yassin

1 Introduction

Interest in hydrogen as a clean fuel has surged in the recent past as concerns over the costs of fossil fuels to the economy and environment have become paramount. Quantities of hydrogen gas on earth are limited, so it must be chemically derived from other sources. Thermochemical water-splitting cycles coupled to a solar heat source to drive the thermal reactions are considered a feasible and possibly advantageous method of generating hydrogen without greenhouse gas emissions. Various studies on thermochemical processes have been reported in the literature [1], [2], [3], [4], [5]. GA has described and analyzed the S-I (sulfur iodine cycle) thermochemical process with flowsheet and simulation results [6]. The UT-3 cycle has also been studied and analyzed by [7], [8], [9]. In the present study, the copper-chlorine (Cu-Cl) reactor portion of the cycle has been studied as part of a promising cycle which can produce hydrogen at a lower temperature than the other cycles. The cycle is closed since all materials are recycled with the exception of water which is split into hydrogen and oxygen. The process involves three main separate reactions: two thermal steps driven by heat and an electrochemical step driven by electric energy. Argonne National Laboratory (ANL) has recently initiated exploratory research to develop a Cu-Cl cycle that operates at its highest temperature of 550°C.

Simulation of the first reactor system in this cycle is introduced based on the CuCl cycle including the arrangement of the receiver and how it couples to provide the necessary heat to the reactors. The MATLAB numerical analysis package was used to solve the partial differential equations resulting from the simulation model. The thermochemical process in this simulation is considered to run semi-continuously. The high temperature solar energy is used directly during sunshine hours and alternatively stored in a thermal storage system. The thermal energy is supplied to the process during night operation from the thermal storage system.

The results of the simulation gives information on the reactor design such as the relations between the changes of the solids reactants radius i.e. (solids mass content) with time and length of the reactor.

2 The Copper Chlorine Cycle

Copper-chlorine (Cu-Cl) cycle is a good alternative for hydrogen gas generation at low temperature in which chemicals are combined with water and heated to cause chemical reactions that produce hydrogen (and oxygen) at 550°C, a temperature compatible with current power plant technologies. The chemicals are not consumed, and are recycled. Argonne National Laboratory's Chemical Engineering Division is currently working on this cycle. The cycle is referred as Argonne Low Temperature Cycle-1 (ALTC-1). This cycle is considered promising over other cycles for the following reasons [11]:

- The maximum cycle temperature (<550°C) allows the use of multiple and proven heat sources
- The intermediate chemicals are relatively safe, inexpensive and abundant
- All reactions have been proven in the laboratory and no significant side reactions have been observed.

The Cu-Cl thermochemical cycle was first proposed in a study and was designated H-6 in a Gas Research Institute (GRI) report [10]. In that study, H-6 consisted of four reactions, three thermal and one electrochemical. ANL's preliminary study indicated that two additional reactions should be added to the original H-6 cycle. So the proposed ALTC-1 cycle consists of six reactions. Reaction-1 is the hydrogen generation reaction and Reaction-6 is oxygen generation reaction [10]. The other reactions close the cycle.

S. Moujaes (✉) • M. Yassin
University of Nevada Las Vegas, Las Vegas, NV, USA

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_18,

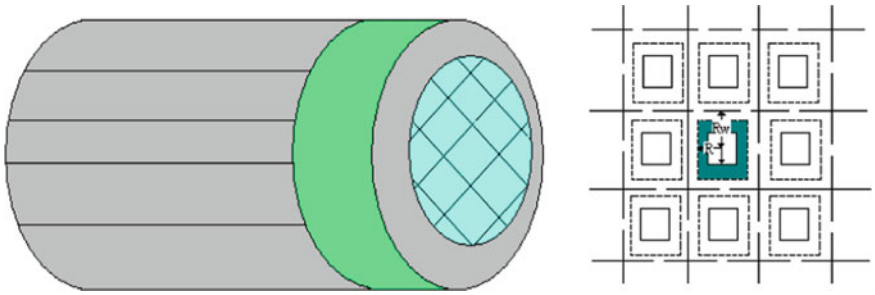


Fig. 1 Model of Honeycomb Shaped Solid Reactants

The simulated reactor of the present study IS for the hydrogen and is summarized as:

(a) 2Cu(s)+2HCl(g) → 2CuCl(s)+H_2(g) which occurs at 450C. The copper (Cu) is in the solid phase and that of the hydrochloric acid (HCL) is in the gaseous phase (g).

For reactor design in general the heat transfer through the reactor as well as mass of the reactants must be conserved.

3 Simulation of First Reactor

The suggested reactor geometry is shown in Fig. 1 and illustrates a honeycomb-shaped solid reactants and the reactor [12]. The aim of this simulation is to predict the performance of the reactor using this honeycomb shaped solid reactant whose solid content of copper forms the walls of this multicell shaped structure. The gaseous HCL enters this reactor through the honeycomb structure and reacts with the solid copper walls which eventually consume the copper material of the walls due to that reaction and eventually has to be replaced.

First the following assumptions are made in the model:

- The reaction proceeds as shown by dotted lines (Fig. 1.) and follows the shrinking core model [13] i.e. as time progresses the void in the core will increase.
- The temperature distribution in the radial direction is negligible.
- The honeycomb shaped reactor is a bundle of independent tabular sections (illustrated by dot-dash lines in Fig. 1).
- Turbulent heat transfer between the gas reactant and the inside wall of solid reactant occurs.
- Heat transfer in the axial direction takes place by conduction for solid and by convection for gas.

Based on these assumptions, the mass and heat balance of both gas and solid reactants are expressed in the following dimensionless form:

- Mass balance for gas reactant

$$\frac{\partial \phi}{\partial \tau} + K_1 \frac{\partial \phi}{\partial l} = (K_2 \gamma)\phi \tag{1}$$

Where:

$$K_1 = \frac{G_0 t_0}{L_0 \rho_{g0}}$$

$$K_2 = \frac{2 t_0 k_s C_{so}}{R_0}$$

$$\gamma = \frac{k_r}{k_s C_{so}} = \frac{1}{k_s C_{so}} \times \left[\frac{R \ln(R/R_0)}{D_e} + \frac{R_0}{R k_s C_{so}} \right]^{-1}$$

- Mass balance for solid reactant

$$\frac{\partial r}{\partial \tau} = \frac{K_3}{r + R_0/(R_w - R_0)} \times \gamma \times \phi \tag{2}$$

Where:

$$K_3 = \frac{\beta R_0 t_0 \, C_{go} \, k_s}{(R_w - R_0)^2}$$

- Heat balance for gas reactant

$$\frac{\partial \theta_g}{\partial l} = K_4 (\theta_s - \theta_g) \tag{3}$$

Where:

$$K_4 = \frac{2 h_w L_o}{R_0 G C_{pg}}$$

- Heat balance for solid reactant

$$\frac{\partial \theta_s}{\partial \tau} = K_5 (\theta_s - \theta_g) + K_6 \frac{\partial^2 \theta_s}{\partial^2 l} - K_7 \times \gamma \times \phi \tag{4}$$

Where:

$$K_5 = -\frac{8 R_0 h_w t_0}{(R_w^2 - R_0^2) C_{ps}}$$

$$K_6 = \frac{k_e t_0}{C_{ps} L_0^2}$$

$$K_7 = -\frac{8 R_0 \Delta H_R t_0 \, C_{go} \, k_s \, C_{so}}{(R_w^2 - R_0^2) T_0 C_{ps}}$$

- Boundary and Initial conditions are given by

$$\theta_g = 1,\ \theta_s = 1,\ \phi = 0\ for\ \ \tau = 0(l \leq 0)$$
$$\theta_g = 1,\ \theta_s = 1,\ \phi = 0\ for\ \ l = 0(\tau > 0)$$

After specifying all the required equations, and calculating all the constants, the next step will be solving these partial differential equations numerically. MATLAB numerical analysis package has been used to solve the resulting PDEs. Table 1 summarizes the data used in this study.

4 Results of Simulation

First, the hydrogen generation step is analyzed, and later in this chapter the second step (Oxygen generation step) is simulated. The plots of the results are given below:

Table 1 Data Used in the Simulation of Reactor

Parameter	Value	Units
C_{pg}	8.5	cal mol^{-1}C^{-1}
C_{ps}	0.31	cal cm^{-3} C^{-1}
C_{so}	3.36*10^-3	mol cm^{-3}
D_e	8.0*10^-3	cm^2 s^{-1}
Lo	600.0	cm
ΔH_R	-8.26*10^-3	cal mol^{-1}
Kc	1.80*10^-3	cal cm^{-1} s^{-1} C^{-1}
ks	308	$cm^4\ mol^{-1}\ s^{-1}$
h_w	3.11*10^-3	$cm^4\ mol^{-1}\ s^{-1}$
To	500	C
β	1.0/8.0	–
P	20.0	atm

Fig. 2 illustrates the relation between the non dimensional radius, the time, and the length of the reactor. As seen from the figure and as expected the radius of solid is decreasing as reaction time increases, the value of the radius goes to zero at $\tau = 1.37$ which means that the copper material has been completely used up.

Fig. 3 shows a 2-D plot which illustrates the decay of the copper material at a chosen location of the reactor which is midway of the length of the reactor l = 0.5 (non-dimensional)

As seen from Fig. 3., r = 0 at $\tau = 1.37$,

From the definition of τ,

$$\tau = \frac{t}{to} \qquad where \qquad t = \ reaction\ \ time$$
$$to = 3600\,s$$
$$t = 4932 \quad \text{sec} = 82.2 \ \text{min}$$

This means that the reaction will be completed within the time (in minutes) of 82.

Fig. 4 illustrates the change of the gas concentration as a function of time and length. As seen from the figure, when the time increases the concentration of the gas increases, this means that the generation of the hydrogen is increasing with time. The non dimensional temperature of gas and solid as a function of distance and time can be seen in Fig. 5 and 6. It is observed that the temperature history of the gas is slightly faster than that of the solid which could be partially explained by the smaller specific heat of the gas than that of the solid. Of course the flow of the gas is removing the heat generated from the reaction by convection which keeps that increase under reasonable control.

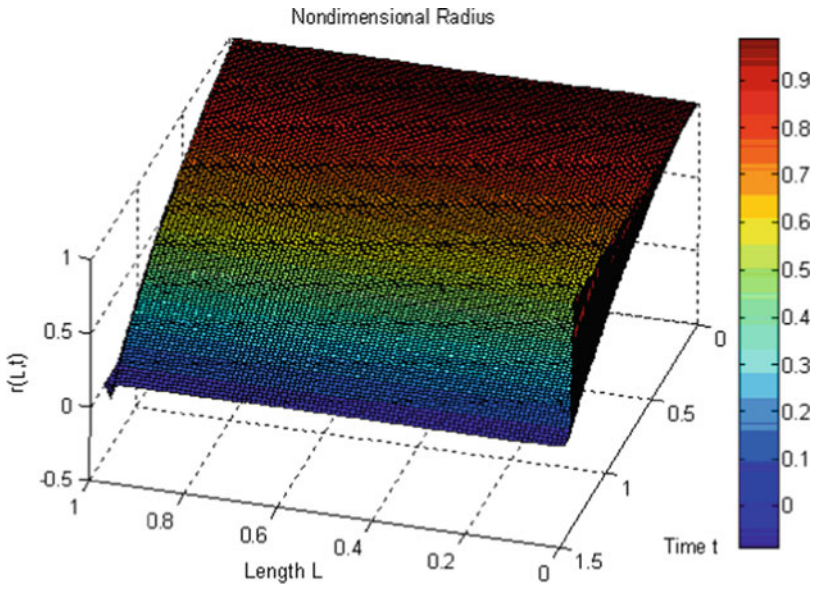


Fig. 2 Non Dimensional Radius as a Function of Time and Length

5 Conclusions and Suggestions

A suggested design of a reactor for one of the reaction steps to low temperature hydrogen using CuCl has been presented and simulated. Based on this conceptual design the dynamic performance response of this solid/gas reactor was performed. The relationship between the reactor length versus gas concentration and the relation between the reactor length versus gas and solid temperatures were obtained. Also, as expected the simulation showed that the solid reactants shrink with time. Initially that shrinkage envisioned by an increase of the radius where the gas is flowing is relatively slow but that rates increases as the solid copper is getting to its last mass content remaining in the honeycomb lattice Based on these results, the amount of solids consumed in the reactor can be specified and a reasonable duration to allow the length of reaction time for the reactants must be designed for

- The highest temperature for this cycle is around 550C which is much lower than other cycles available for producing hydrogen namely the S-I cycle and that of a pure thermal dissociation of the components of water.
- Though the simulation model presented for this reactor is preliminary, it will be very helpful for producing insights for future improvement of this reactor.
- The cost analysis of Cu-Cl cycle has not been done. But it is very necessary to conduct cost analysis and estimation for the economic feasibility of the cycle.
- The solar input must be matched to the chemical process such that high thermal efficiency is obtained, but not at the expense of sacrificing the operability of the combined plant.
- The matching must be done in a way that promotes operational stability of the chemical process.

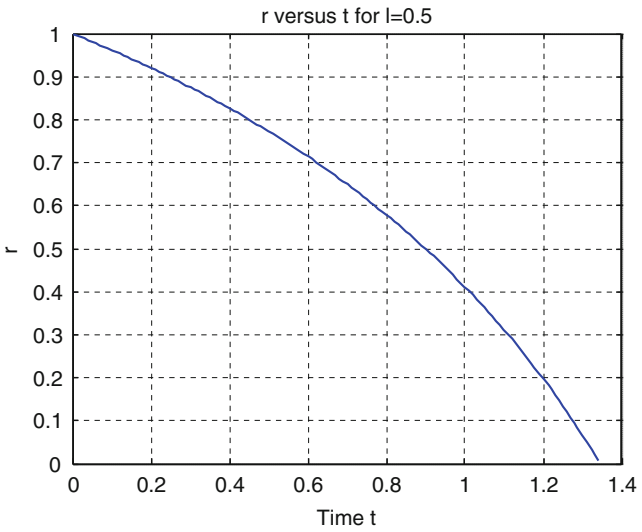


Fig. 3 Non Dimensional Radius as a Function of Time

Nomenclature

C_{go}	Initial molar concentration of gas reactant [mol cm ^-3]
C_g	Concentration of gas reactant [mol cm^-3]
C_{pg}	Average heat capacity of gaseous substances [cal mol^-1 C^-1]

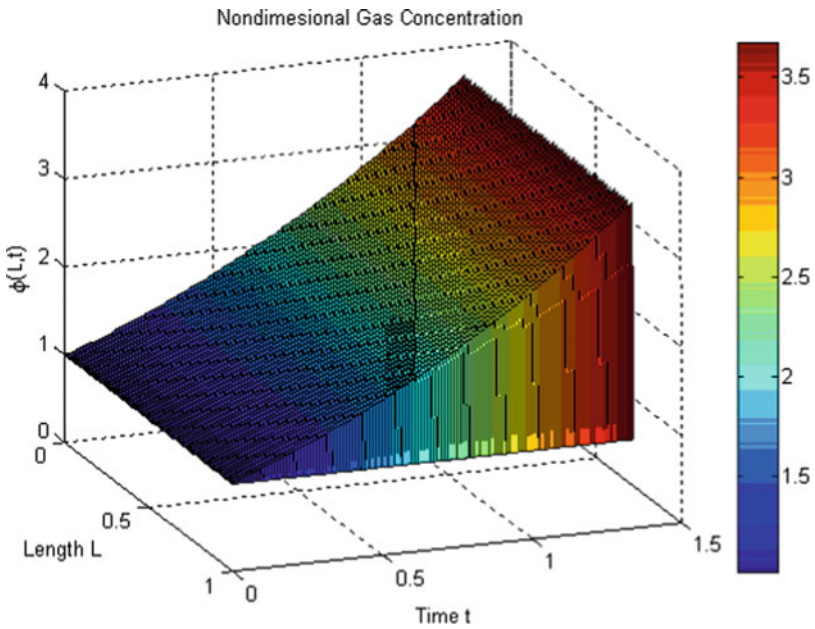


Fig. 4 Non Dimensional Gas Concentration as a Function of Length and Time

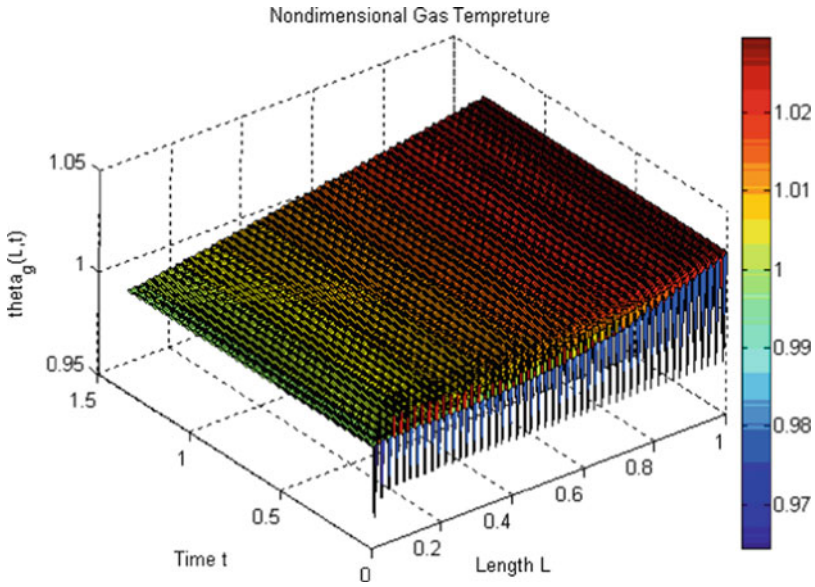


Fig. 5 Non Dimensional Temperature of Gas as a Function of Length and Time

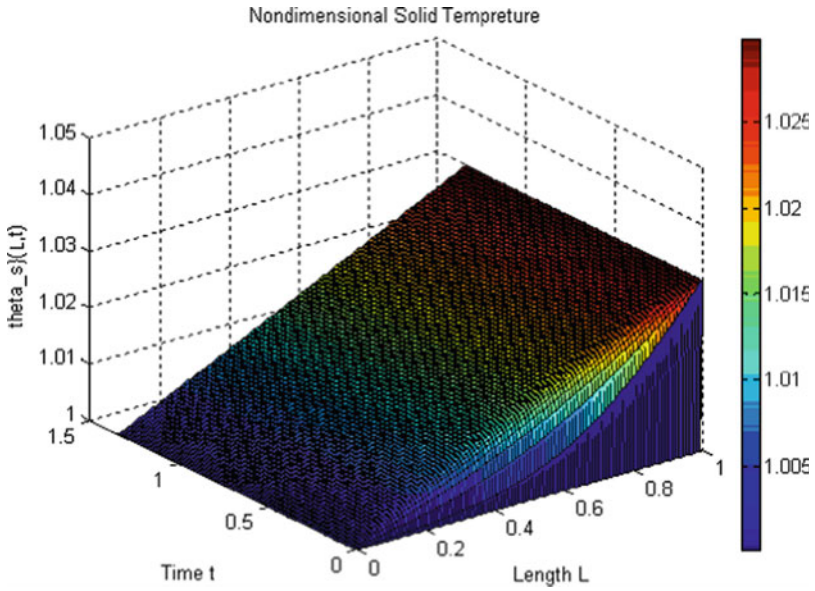


Fig. 6 Non Dimensional Temperature of Solid as a Function of Length and Time

C_{ps}	Average heat capacity of solid substances [cal cm^-3 C^-1]
C_{so}	Initial molar density of solid reactant [mol cm ^-3]
Cs	Molar density of solid reactant [mol cm ^-3]
D_e	Effective diffusion coefficient in product layer [cm^2 s^-1]
f	Conversion of gas reactant [-]
Go	Initial molar velocity of gas [mol s^-1 cm^-2]
G	Molar specific velocity of gas [mol s^-1 cm^-2]
h_w	Film coefficient of heat transfer [cal cm^-2 s^-1 C^-1]
ΔH_R	Heat of reaction [cal mol^-1]
Kc	Effective heat conductivity of solid [cal cm^-1 s^-1 C^-1]
kr	Overall reaction rate constant [cm s^-1]
ks	Reaction rate constant based on unit surface [cm^4 mol^-1 s^-1]
l	Non dimensional length L/Lo [-]
l_s	Length of reaction zone [cm]
Lo	Total length of solid reactant [cm]
L	Distance from the inlet of reactor [cm]
P	Reaction Pressure [atm]
r	Non dimensional radius [(R-Ro)/(Rw-Ro)] [-]
Ro	Inner radius [cm]
Rw	Outer radius [cm]
R	Distance from the center of hole to reaction surface [cm]

t_o	Resident time of reactant in reactor [s]
t	Reaction time [s]
To	Inlet temp of gas reactant [C]
Tg	Temp of gas reactant [C]
Ts	Temp of solid reactant [C]
v_s	Advancement rate of reaction zone [cm s^-1]
X	Molar fraction of gas reactant at inlet [-]
β	Ratio of mole of solid to that of gas [-]
γ	Non dimensional reaction rate term
θ_g	Non dimensional temperature of gas Tg/To [-]
θ_s	Non dimensional temperature of solid Ts/To [-]
ρ_{go}	Inlet molar density of gas [mol cm^-3]
ρ_g	Molar density of gas [mol cm^-3]
τ	Non dimensional time t/to [-]
φ	Non dimensional concentration of gas Cg/Cgo [-]

References

1. Huang, C.; Raissi, A., Analysis of sulfur-iodine thermochemical cycle for solar hydrogen production. Part-I: decomposition of sulfuric acid, Solar Energy 2004.
2. Ozturk, I.T.; Hammache, A.; Bilgen, E., An improved process for H_2SO_4 decomposition step of the sulfur-iodine cycle, Energy Conversion Management, 36 (1995), 11–21.
3. Roth, M.; Knoche, K. F., Thermochemical water splitting through direct HI-decomposition from $H_2O/HI/I_2$ solutions, International Journal of Hydrogen Energy, 14 (1989) 545–549
4. Normana, J. H.; Myselsb, K. J.; Sharpa, R.; Williamsonc, D., Studies of the Sulfur-Iodine thermochemical water splitting cycle, International Journal of Hydrogen Energy, 7 (1982), 545–556
5. Bilgen, C.; Bilgen, E., An assessment on hydrogen production using central receiver solar systems, International Journal of Hydrogen Energy, 9 (1984) 197–204.
6. Brown, L.C.; Besenbruch, G. E.; Lentsch, R.D.; Schultz, K.R.; Funk, J. F.; Pickard, P.S.; Marshall, A.C.; Showalter, S.K., High efficiency generation of hydrogen fuels using nuclear power, Final Technical Report, August 1, 1999 through September 30, 2002, General Atomics Report GA – A24285 (2003).
7. Sakurai, M.; Miyake, N.; Tsutsumi, A.; Yoshida, K., Analysis of a reaction mechanism in the UT-3 thermochemical hydrogen production cycle, International Journal of Hydrogen Energy, 21 (1996), 871–875.
8. Aihara, M.; Umida, H.; Tsutsumi, A.; Yoshida, K., Kinetic study of UT-3 thermochemichal hydrogen production cycle, International Journal of Hydrogen Energy, 15 (1990), 7–11.
9. Tadokoro, Y.; Yamaguchi, T.; Sakai, N.; Yoshida, K.; Kameyama, H.; Aochi, T.; Nobue, M.; Aihara, M.; Amir, R.; Kondo, H.; Sato, T., A simulation study of the UT-3 thermochemical hydrogen production process, International Journal of Hydrogen Energy, 15 (1990), 171–178.
10. Lewis, M. A.; Serban, M.; Basco, J., Hydrogen production at 550°C using a low temperature thermochemical cycle, Proceedings of the OECD/NEA Meeting, Argonne National Laboratory, 2003
11. Lewis, M. A.; Serban, M.; Basco, J., Kinetic study of the hydrogen and oxygen production reactions in the Copper-Chlorine Thermochemical Cycle, AIChE 2004 Spring National Meeting, New Orleans, LA, April 25–29, 2004.
12. H. Kameyama and K. Yoshida, Reactor Design for the UT-3 Thermochemical Hydrogen Production Process, Int. J. Hydrogen, Vol. 6, No. 6, pp. 567–575, (1981).
13. O. Levenspiel, Chemical Reaction Engineering, 2nd Edition, chapter 12, Wiley, New York (1972).

Voltage Regulation in Resonant Coupled Systems for Near Field Power Transfer

Hema Ramachandran and G.R. Bindu

1 Introduction

Nikola Tesla in the early 19th century used his infamous Tesla coil to send streaks of wireless power to moderate distances using high density electric and magnetic fields. The necessity of using wireless energy was not thought of until its resurgence at MIT in 2007 [1]. The recent advances in the communication industry through enabling of a wide variety of portable devices have accelerated the need of wireless power transfer. Most of the portable electronic devices have to be manually plugged in at the time of charging. The connectors for plugging the devices occupy a major volume compared to the equipment size. The wireless power transfer system experimented at MIT involved two five turn copper coils each of 60 cm diameter and were wound on a plastic bobbin. The sending coil was attached to a high frequency sinusoidal power source. The receiving coil was attached to a 60 W bulb, and placed at a distance of 2 m from the sending unit. Using fine tuning mechanisms, the source and receiving coils were tuned to the same resonant frequency. Power was transferred to the receiving coil at 2 m distance with 45 % efficiency when the source and receiving coils were tuned to resonance at 9.90 MHz.

Wireless power transfer systems are categorized into three types namely inductive coupled power transfer systems (ICPT), resonant coupled power transfer systems (RCPT) and radio frequency systems (RF) [2]. ICPT systems require high coupling between the source and receiving coil to achieve high efficiency and works only over distances of a few millimeter. In RCPT systems the source and receiving coils are loosely coupled and they exchange energy efficiently when tuned to the same frequency. The effect of coupling is seen to be very weak at other frequencies. Both ICPT and RCPT systems are used for powering devices in the near field [3] at room scale distances. Near field is the electromagnetic field which is omnidirectional and propagated up to a distance of 0.159λ, where λ is the wavelength of the source emitting the field. The near field effects are evanescent and die down with increasing distance from the source. The magnetic field strength is proportional to the inversecube of the distance and electric field strength proportional to inverse-square of distance. The electromagnetic fields essentially are evanescent and dies with distance [4]. RF systems uses frequencies greater than the near field and spans to distances greater than $2D^2/\lambda$ to infinity where D is the diameter of the source sending the field [5]. The fields are radiative and require highly directional antennas for targeted military systems. Mostly these systems use microwaves and lasers for achieving line of sight. RCPT systems using near field power transfer has been used to power multiple receivers terminated by lumped capacitors [6]. It is also seen that efficiency of resonant power transfer can be increased using variable coupling between the coils [7]. Matching networks have been proposed to deal with distance variations, but it adds loss to the system [8].

This paper is organized as follows: The various topologies of resonant coupled power transfer between the source and receiver coils is presented in Section 2; the system modeled using distributed capacitance and external capacitances with the receiver equivalent circuit is presented in Section 3; Section 4 analyses the voltage regulation for varying loads and correction schemes are proposed. Experimental verification on RCPT resistive load and measurement of voltage regulation is performed in Section 5.

H. Ramachandran (✉)
Speed-IT Research Fellow, Department of Electrical Engineering, Government College of Engineering, Thiruvananthapuram, Kerala, India
e-mail: hemaachuth@gmail.com

G.R. Bindu
Department of Electrical Engineering, Government College of Engineering, Thiruvananthapuram, Kerala, India
e-mail: bgr100@gmail.com

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_19,

2 Topologies of Resonant Coupled Power Transfer

The basic topologies of RCPT systems are classified into four categories. The basic topology in Fig. 1(a) is modified to represent all the other categories as in Fig 1(b)-(d).

The source coil is inductively coupled to the receiving coil by a high power, high frequency driving circuit. Since the receiver coil is placed at varying distances from the source coil, the mutual inductance M is related to distance of separation of the coils. In Fig. 1(b), the drop across the source during wireless power transmission is attributed as an impedance drop Z_M, which is reactive but very small, compared to the reactance of the source circuit. The impedance drop does not change rapidly in the range of power transfer and hence power transfer remains identical with frequency tuning of the coils. In Fig. 1(c), large impedances of resistive, inductive and capacitive nature are inserted between the source and receiving coils, where power transfer curves are not symmetrical with same frequency tuning of the coils. In Fig. 1(d), capacitive reactances form the coupling impedances. Most of the power transfer systems opt for the basic arrangement in Fig. 1(a).

3 Model of Rcpt Using Parasitic and External Capacitances and Approximate Receiver Equivalent Circuit

We used a variation in basic arrangement as in Fig. 2, in which source coil inherent characteristics are represented where resistance R_s is in series with inductance L_s which together is in parallel with parasitic capacitance. The coil inductance was measured considering that the coils formed a cylindrical current sheet of infinite length [9] where μ_0 is the permeability of free space, A(m) and l (m) are the area and length of the solenoid and is represented by (1).

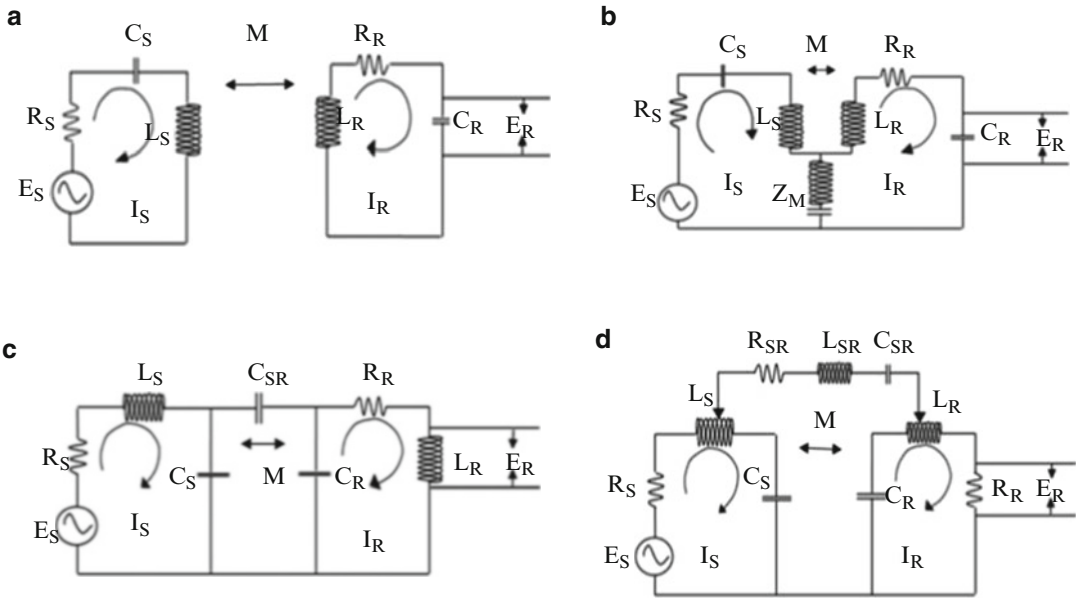


Fig. 1 (**a**) Basic topology of a resonant coupled wireless power transfer system (**b**). Drop across source coil represented as impedance between the source and receiver. (**c**) Insertion of inductance, resistance and capacitance between the source and receiver coils. (**d**) Insertion of coupling capacitance between the source and receiving coils

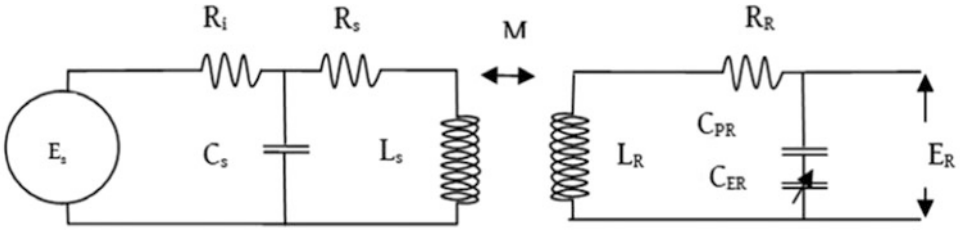


Fig. 2 Variation in basic arrangement of RCPT incorporating parasitic and external capacitances

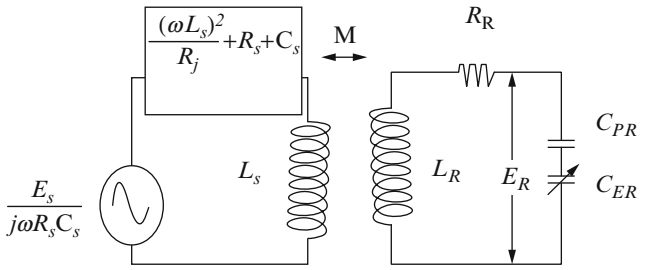


Fig. 3 The transformed RCPT circuit after current source and voltage source transformation

$$L = \frac{\mu_0 A N^2}{l} \tag{1}$$

The parasitic capacitance of coils C_{PS} and C_{PR} is not negligible [10] and has been well represented by equation (2) where D is the diameter of the coil (m) and l is the length of the coil (m).

$$C = \left(\frac{4\varepsilon_0 l}{\pi}\right) \times \left(1 + 0.717439\left(\frac{D}{l}\right) + 0.933048\left(\frac{D}{l}\right)^{\frac{3}{2}} + 0.106\left(\frac{D}{l}\right)\right) \tag{2}$$

We utilize this parasitic capacitance C_{PS} along with external variable capacitors C_{ES} inserted in parallel with the coil to achieve resonance. The powering circuit to the source coil is represented as a resistance R_i. The receiver coil also is represented in a similar manner with resistance R_R, inductance L_R and effective capacitance due to parasitic behavior C_{PR} and insertion of external variable capacitances C_{ER}. The resistances of coils takes into account the skin effect and proximity effect in AC power transmission. The coil resistances where a is the radius of the coil, N the number of turns, σ is the conductivity of copper and r the radius of the wire and is calculated by (3).

$$R = \frac{aN}{\sigma r \delta} \tag{3}$$

The coil quality factors are represented as $Q_{S,\,R=(\omega}\,L_{S,,R})\,/\,R_{S,,R}$. The mutual inductance between the coils measured using the Neumann formula [11] as in (4), D is the distance between dl_1 and dl_2.

$$M = \frac{\mu_0}{4\pi} \oint_{c1c2} \frac{dl_1 dl_2}{D} \tag{4}$$

The transformed circuit of the source coil using current source transformation and voltage source transformation along with the receiver coil can be represented as in Fig. 3.

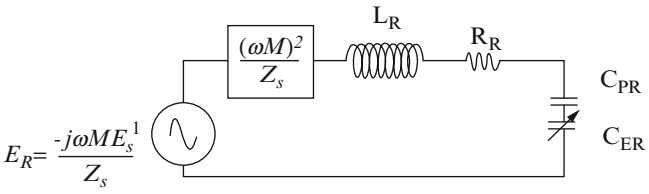


Fig. 4 Approximate equivalent circuit of the receiver coil with the thevenin open circuit voltage across its terminals

By applying Thevenin's theorem, when the receiver circuit is opened, the open circuit voltage that appears across the receiver coil is $E_R = \frac{-j\omega M E_s^{\,l}}{Z_s}$ where $E_s^l = \frac{E_s}{j\omega R_s C_s}$ and $Z_s = R_s + \frac{(\omega L_s)^2}{R_i} + j\omega L_s + 1/j\omega(C_{PR}+C_{ER})$ and I_R is the no load current. When the terminals are shorted, the equivalent impedance viewed from the receiver terminal is $ZR + (\omega M)^2/Z_s$,where $Z_R = R_R + J\omega L_R + 1/(j\omega(C_{PS} + C_{ES}))$. The receiver coil circuit consists of the actual receiver coil impedance Z_R and an additional impedance $(\omega M)^2/Z_s$ from the source coil often referred as coupled impedance. The reduced equivalent circuit viewed from the receiver terminals is shown in Fig. 4.

4 Voltage Regulation For Varying Load Conditions

Power factor of the load is a determining factor in the load voltage, it influences voltage regulation. In resonant coupled power transfer systems, at maximum power transfer, the circuit power factor is unity. High quality factor and efficient coupling are prerequisites for attaining this condition. The voltage regulation of the near field system can be represented similar to a transformer as the percentage change in the output voltage from no-loading of the receiver coil to full-loading of the receiver coil. The voltage regulation as in a transformer is dynamic and load-dependent.

Resistive loads as in Fig. 5(a), when used at the output of the receiver coil will produce a voltage drop across the resistance and but will not affect the overall circuit power factor. The output voltage across the load is less than the receiver coil voltage and accounts for positive voltage regulation as shown in Fig. 5(b). The percentage voltage regulation is given by (5) where

$$R_eq = \frac{(\omega M)^2 R_s^1}{(R_s^1)^2 + (X_s^2)} + R_R, \quad R_s^1 = R_s + \frac{(\omega L_s)^2}{R_i} \text{ and } X_s = \omega L_s - 1/\omega C_s. \frac{E_R - V_L}{V_l} \times 100\% = \frac{I_L R_{eq} + I_L X_{eq}}{V_L} \times 100\% \tag{5}$$

When connected to an inductive load such as lighting ballast or an induction furnace as in Fig. 6(a), the load

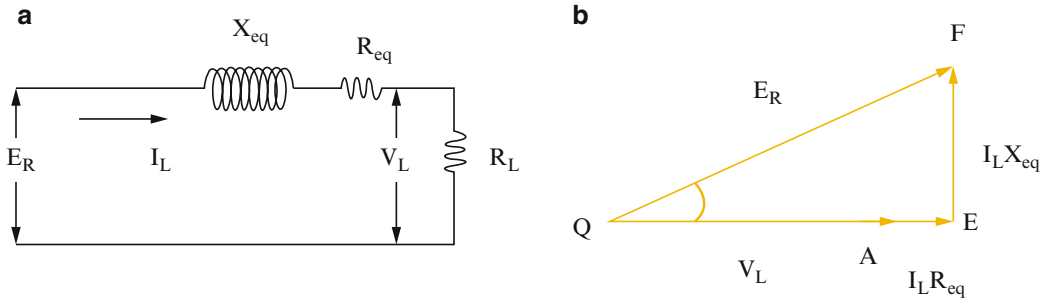


Fig. 5 (**a**) The receiver coil of RCPT feeding a resistive load. (**b**) Voltage regulation phasor diagram for resistive load condition

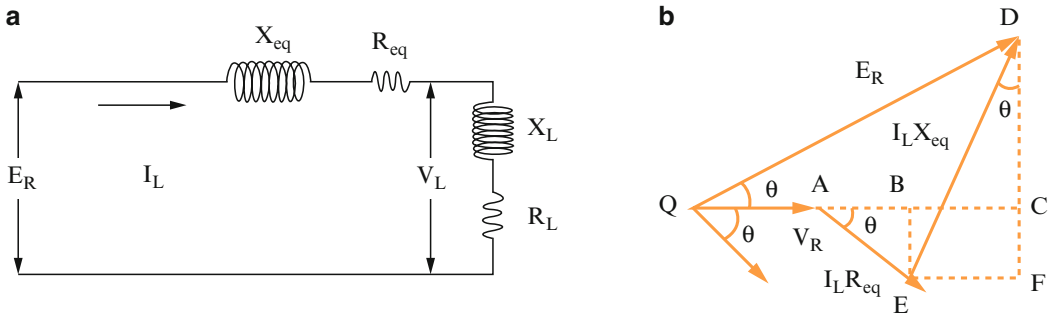


Fig. 6 (**a**) The receiver coil of RCPT feeding an inductive load. (**b**) Voltage regulation phasor diagram for inductive load condition

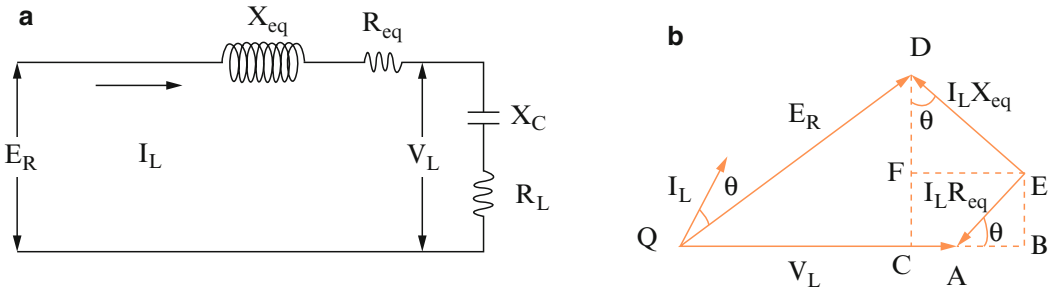


Fig. 7 (**a**) The receiver coil of RCPT feeding a capacitive load. (**b**) Voltage regulation phasor diagram for capacitive load condition

current lags the load voltage, lowering the power factor of the entire circuit and thereby the resonant frequency. Reduction in power factor leads to an increase in current drawn by the transmitter coil from the driving circuit and the size of the transmitter coil have to be increased. To ensure maximum power transfer, a power factor correction scheme is proposed by incrementing the external capacitance added to the receiver circuit C_{ER} so that the inductive reactance is compensated by capacitive reactance due to variable external tuning capacitance.

The effect of such a correction scheme shifts the regulation curve from Fig. 6 to 5(b). The percentage voltage regulation before correction is given by (6) and after power factor correction it reduces to (5).

$$\frac{E_R - V_l}{V_l} \times 100\% = \frac{I_l\, R_{cq} \cos\theta + I_l\, X_{cq} \sin\theta}{V_l} \times 100\% \quad (6)$$

When the RCPT system is connected to a capacitive load such a flash on camera as in Fig. 7(a), the load current leads the voltage which results in increase of power factor of the entire circuit and a shift in resonant frequency. The regulation phasor diagram is presented in Fig. 7(b).

The power factor correction scheme proposed here is a lowering of external capacitance C_{ER}, thereby retuning the

circuit to Fig. 5(b). The percentage voltage regulation before power factor correction is given by (7) and after correction is reduced to (5).

$$\frac{E_R - V_l}{V_l} \times 100\% = \frac{I_l R_{cq} \cos\theta - I_l X_{cq} \sin\theta}{V_l} \times 100\% \quad (7)$$

5 Experimental Verification Of Voltage Regulation On Rcpt

The experimental schematic of load regulation on a resonant coupled power transfer system is shown in Fig. 8. Regulated power supply of 230 V, 50Hz was fed to a Bridge rectifier followed by a Buck converter using IGBT followed by a Full Bridge Inverter using MOSFETs. The output of the inverter was fed to a load matching transformer with capacitors inserted in parallel to get the desired resonant frequency. The maximum power fed from the power converter circuit was decided by the input impedance of the supply and impedance of the source-receiver RCPT model. The voltage across the source coil was set to 500 V. The source coil and the receiver coil used Litz conductors of cross sectional area 2.5sqmm. Both the coils were wound around a plastic bobbin of diameter 15 cm and length 2 cm. Identical source and receiver coils were taken so that the effect of loading can be determined precisely, since power transfer is determined by the geometry of the coils. Both the source and receiver coil was connected to fixed valued external capacitors in parallel for tuning them to the resonant frequency of 1.175 MHz. The extracted coil parameters of the identical source and receiver coil extracted are presented in Table I.

The receiver coil was attached to a 60 W, 230 V bulb as the resistive load with a resistance of 881.96 Ω. The experimental setup of wireless power transfer is shown in Fig. 8(b). The receiver coil can be set to varying coil positions in coaxial orientation from the source coil at distance 1, 10, 20 and 30 cm thereby varying the power transferred to the load and efficiency of power transfer. The open circuit voltage across the receiver coil was measured with a potential transformer before and after loading. The current flowing through the load was measured using a current transformer. The experimental plots are presented in Fig. 9 (a) and (b). Maximum efficiency is obtained is 15.73 % and power transferred to the load is 57.16 W, when the receiver coil is kept at 1 cm from the centre of the source coil. The resistive light load glows with maximum brightness. The open circuit voltage across the receiver coil at no load is 224.85 V and with load it drops to 223.46 V corresponding to a regulation of 0.62 %. As distance increased, the current through the load drops as the evanescent near field begins to die down. At a distance of 20 cm, the open circuit voltage drops to 74.95 V while the voltage across the load becomes 74.92 V. The efficiency at 20 cm distance was 5.66 %. The efficiency dropped to 2 % at a distance of 30 cm at which the power transferred to the load was only 6.37 W, where the lamp showed a slight flicker. The regulation was very minimal corresponding to a load current of 0.04A.

6 Conclusion

Voltage regulation study conducted on the experimental RCPT system shows that the regulation increases with increase of load current and remains positive. The experiment can be repeated with inductive and capacitive loads employing power factor correction schemes by retuning the capacitances that terminate the receiver coil for realizing resonant power transfer, while maintaining positive voltage regulation. Retuning the receiver coils using advanced

Table 1 Extracted Coil parameters

Parameters	Source coil	Receiver coil
SelfInductance L_S, L_R	0.916 μH	0.916 μH
Parasitic-Capacitance CPS,C_{PR}	5.08×10^{-11}	5.08×10^{-11}
External Capacitance C_{ES}, C_{ER}	0.02 μF	0.02 μF
Resistance R_T, R_R	0.1 Ω	0.1 Ω
Quality factors Q_S, Q_R	67.59	67.59
Resonant frequency f_S, f_R	1.1750×10^6Hz	1.1750×10^6Hz

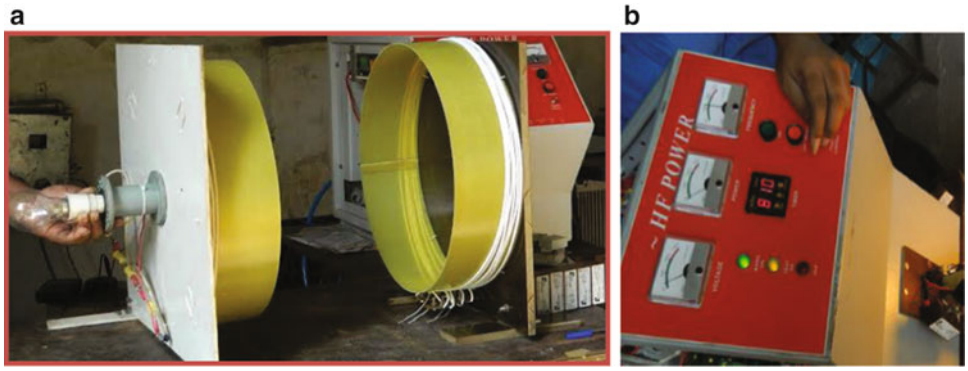


Fig. 8 (**a**) The experimental system with the drive circuit, coaxial source and receiver coils and the 60 W bulb attached to the receiver forming the load (**b**) The high frequency high power driver circuitry to the source coil

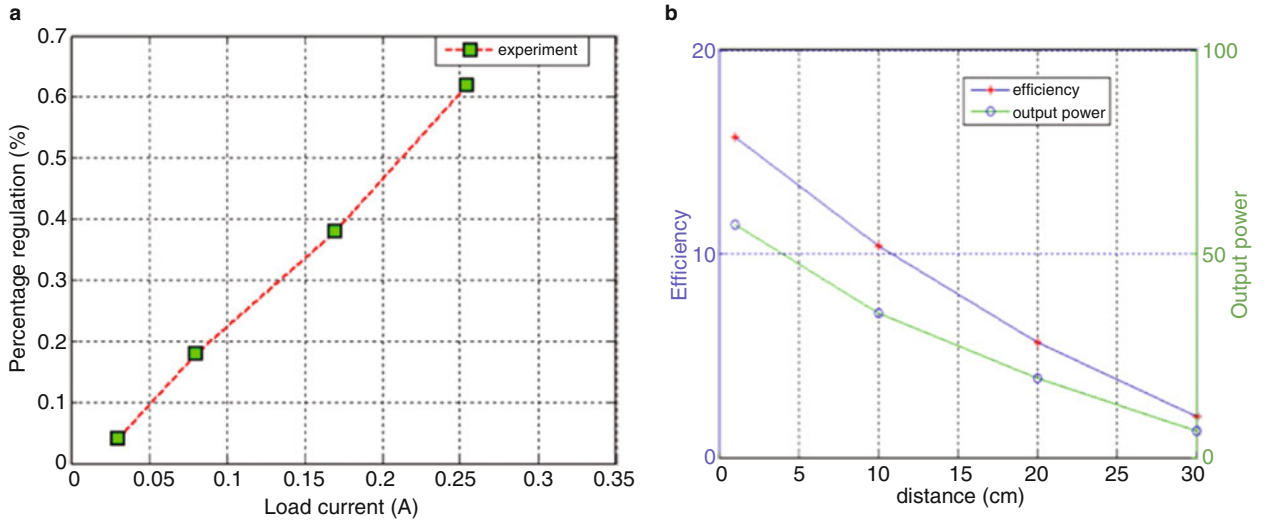


Fig. 9 (a) Variation of efficiency and output power across the load with distance. (b) Variation of percentage regulation with load current

control circuits is a scope of further work. The above study can also be performed using varying coil geometries and inclined coil orientations using specialized conductors like hollow copper tubes with reduced skin effect.

Nomenclature

E_s, E_R	Voltage applied across the source coil and receiver coil (V)
R_i	Resistance of the powering circuit (Ω)
L_s, L_R	Self- Inductance of the source and receiver coils (H)
R_s, R_R	Resistance of the source and receiver coils (Ω)
C_s, C_R	Total capacitance of the source and receiver coils(F)
C_{PS}, C_{PR}	Parasitic capacitance of the source and receiver coils(Ω)
C_{ES}, C_{ER}	External capacitance inserted in the source and receiver coils(Ω)
M	Mutual Inductance between source and receiver coils(H)
Q_T, Q_R	Quality factor of the source and receiver coils
Z_M	Coupling impedance connected between the source and receiver coils(Ω)
L_{SR}, R_{SR}, C_{SR}	Inductance, Resistance and Capacitance inserted between the source and receiver coils (Ω).

Acknowledgment This work was supported by the grants from the Kerala State Council for Science, Technology and Environment, Kerala, India. The second author acknowledges Speed-IT fellowship of the IT Department of the Government of Kerala, India.

References

[1] A. Kurs; A. Karalis; R. Moffat; J.D. Joannopoulos; P. Fischer and M. Soljacic, "Wireless power transmission via strongly coupled magnetic resonances," *Science*, vol. 317, pp. 83–86, 2007.
[2] Alanson P. Sample; David A. Meyer and Joshua R.Smith, "Analysis, experimental results and range adaptation of magnetically coupled resonators for wireless power transmission," *IEEE Trans. Ind. Electron.*, pp.1–11, May.2010.
[3] Qiaowei Yuan; QiangChen, Long Li and KunioSawaya, "Numerical analysis on transmission efficiency of evanescent resonant coupling wireless power transmission system," *IEEE Antennas and Propag; vol.58*, no.5, pp.1751–1758, 2010.
[4] Hema Ramachandran and Bindu G.R; 2013, "General public exposure radiation measurements in the vicinity of a wireless power transfer prototype system," *International Journal on Communicatons, Antenna and Propagation, (IRECAP)*, vol. 3 no. 3, pp. 168–75, June 2013.
[5] Constantine A. Balanis ; 2005, "Antenna Theory: Analysis and Design, " 3rd edition Ch. 2 p. 34.
[6] C.J. Chen, T.H. Chu, C.A.Lin and Z.C. Jou; 2010, "A study of loosely coupled coils for wireless power transfer," *IEEE Trans. Power Electron;* vol.24, n0.7, pp.1819–1825.
[7] Thuc Phi Duong and Jong-Wook Lee, "Experimental results of high-efficiency resonant coupling wireless power transmission using a variable coupling method," IEEE *Microwave Wireless Comp. Lett.*; vol.21, no.8, 2011.
[8] E.R Giler et al ; 2010, "Wireless energy transfer using repeater resonators", U.S. Patent 2010/0259108.
[9] H. Nagaoka, "Inductance coefficients of solenoids," *Jour. Coll. Sci. Tokyo*, vol. 27, art.6, 1909.
[10] David W. Knight, "Self -resonance and self-capacitance of solenoid coils," ver. 0.01, May 2010.
[11] S. Ramo, J.R Whinnery and T. Vanduzer, " Fields and Waves in Communication Electronics", 3rd ed. New York, Wiley,1994.

Security Breach Possibility with RSS-Based Localization of Smart Meters Incorporating Maximum Likelihood Estimator

Mahdi Jamei, Arif I. Sarwat, S. S. Iyengar, and Faisal Kaleem

1 Introduction

Recently, there has been a trend towards the Smart Grid (SG) to have secure and reliable electricity [1]. The SG is two-way data transfer in which the information plays a central role in energy dispatching [2]. Smart meter is one of the key components which enable the SG to involve the consumer engagement and demand response concepts.

Increasing rate of the electricity consumers necessitates a revolution in the conventional energy consumption metering and current methods of billing. In addition, human errors and operating costs and the need of improvement in the metering efficiency have encouraged utilities to employ Advanced Metering Infrastructure (AMI) systems [3]. In a report of the U.S. Energy Information Administration, about 37,290,374 AMI were installed by 493 U.S. electric utilities in 2011 [4].

AMI systems collect the energy usage data remotely to remove the inaccuracy and the reading cost. They can also provide both utilities and consumers with real-time consumption data to enable the grid for a better response to demand changes [5]. AMI transmitting and receiving methods have evolved gradually. Early model of AMI used telephone lines to send and receive the data. Power Line Communication (PLC), low power Radio Frequency (RF) and satellite-based communication are the next generations of the AMI systems.

However, smart meters have raised concerns over security and privacy issues [6]. Data collected and transmitted by the AMI systems can endanger the privacy of consumers. For example, robbers can receive and demodulate the sent data by smart meters to localize the unoccupied houses. People's daily routines can also be identified from the time and the amount of the energy consumption [7]. On the other hand, it is possible that home-owners hack the system and falsify the reported data through breaking into the communication channel or feeding the system by counterfeit consumption data [8].

To show the significance of the related security issues, this paper presents a method of localizing a smart meter using the Received Signal Strength (RSS) of the RF transmitted signals. This method shows the vulnerability of the AMI system's security since it allows hackers to identify the location of an intended owner via the energy usage data. It is important to recognize and study the attack scenarios and their effectiveness to find efficient anti-attack remedies in order to increase the impenetrability of the system.

RSS-based localization are mainly categorized as range-based and range-free methods. RSS is used as a distance reference in range-based techniques while range-free methods do not use distances in the localization procedure. In the range-based system, the location of the emitter can be identified by distances from a set of sensors with known positions [9]. For this purpose, at least three sensors are required to make the localization possible.

RSS-based localization problem requires estimation of an unknown parameter i.e. the coordination of the emitter, from a collection of observation data, x[n], by sensors in which additive noise, sensor inaccuracies, shadowing, multipath and path loss exponent have been included. The Maximum Likelihood (ML) estimator is an approach in estimating a parameter when the Probability Density Function (PDF) is known. With MLE, the unknown parameter is estimated by maximizing the PDF [10]. In this paper, RSS-based localization of a smart meter from the sent data by AMI system under the assumption of a log-normal path loss model and additive Gaussian noise has been proposed. ML estimator is employed to estimate the coordination and the reference power of the smart meter using the received signals by the sensors located at the

M. Jamei (✉) • A.I. Sarwat • F. Kaleem
Electrical and Computer Engineering Department, Florida International University, Miami, Florida, USA
e-mail: mjame044@fiu.edu; asarwat@fiu.edu; kaleemf@fiu.edu

S.S. Iyengar
School of Computing and Information Sciences, Florida International University, Miami, Florida, USA
e-mail: iyengar@cis.fiu.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_20,

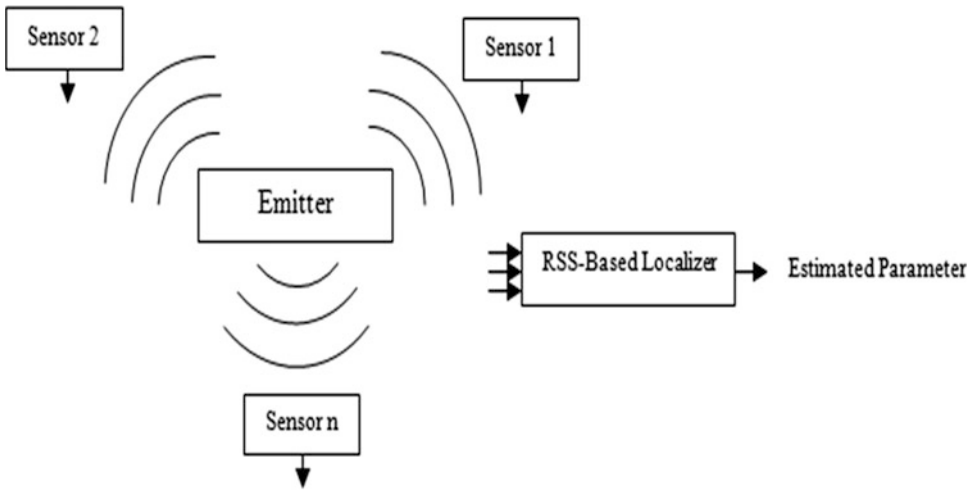


Fig. 1 General Schematic of the Localizer

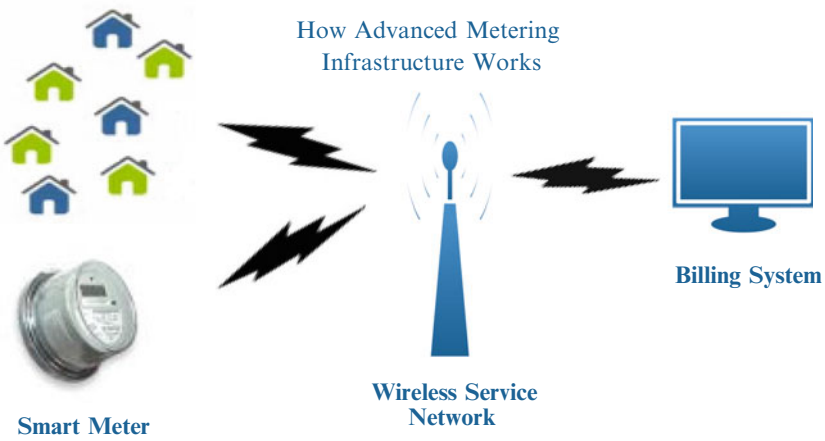


Fig. 2 Outline of the AMI System

known positions. The general schematic of the paper is shown in Fig. 1. The effectiveness of the proposed method is investigated through MATLAB simulation considering the FSK modulation and demodulation. PSO has been implemented to maximize the likelihood function in the ML estimator. Finally, the effect of the variance, the number of the sensors and the path loss exponent has been studied on the average Miss Distance Error (MDE).

The rest of this paper is organized as follows. Section 2 introduces the AMI main framework architecture. Section 3 presents the ML estimator concepts and the proposed method of the RSS-based localization problem. Simulation data and results are presented and discussed in Section 4. Finally, the conclusion is given in Section 5.

2 AMI Architecture

AMI system collects the energy usage data and transmits to the central collector for billing and analysis purposes. Most of the new AMI systems have been equipped with the RF or satellite communication. RF technology is broadly in use since it is more cost effective and easier to be implemented. The RF method is considered to be the communication infrastructure in this paper. Fig. 2 illustrates the outline of the AMI system.

AMIs are mainly classified as AMI meters that collect the energy usage data and AMI readers which receive and send it to the central processing offices [5].

2.1 AMI Meter

AMI meters gather the electricity, gas and water consumption. RF-based meters include Encoder, Receiver and Transmitter (ERT) which has a microprocessor as well as a low power transmitter [5]. The microprocessor processes the meter reading periodically and feed it into the transmitter to be sent along with the information such as the meter ID.

2.2 AMI Reader

AMI readers are used to receive the data from AMI meter and send it to the utilities main data collectors. There are three different types of readers in the current electricity grid [5], [11].

i. Handheld instruments for walk-by data collection
ii. Mobile data collection for drive-by data reading
iii. Permanent infrastructure for real-time data transfer

The first two classes require personnel to gather data so the information can be updated periodically but the permanent infrastructure can provide real-time energy data. On the other hand, the initial start-up cost of installing the fixed data transfer system is considerable so benefit-cost analysis must be conducted before choosing one of these two mentioned categories.

2.3 Communication Protocol

The communication module provides a robust connectivity in the Neighborhood Area Network (NAN). The data rates for the NAN communication is 100 kbps with the transmitter output of 27-30 dBm (500 mW - 1 W). It supports the frequency range from 902-928 MHz with 915 MHz ISM band. The module also include a 2.4 GHz radio for the Home Area Network (HAN) which supports the ZigBee protocol to communicate with the smart devices inside the house. The transmitter output of the HAN communication is 20 to 23 dBm (100 - 200 mW). On-Off Keying (OOK) and FSK are usual types of modulation used in the transmitter and the receiver.

3 Problem Formulation

In this section, the concept of the ML estimator has been introduced first. The localization problem has also been formulated and related equations are given to be implemented in the Smart Meter security issue.

3.1 ML Estimator

The ML estimator is overwhelmingly the most famous approach to obtain practical estimators. In most of the cases, by large enough data observation, the ML estimator would be the optimal one. It can be said that the ML estimator is approximately Minimum Variance Unbiased Estimator (MVUE) [10].

Generally, the ML estimator defined as the value of θ which maximizes the likelihood function. The MLE is said to be asymptotically optimal which means that the properties of unbiasedness and achieving the Cramer-Rao Lower Bound (CRLB) can only be obtained by recording large enough data. The ML estimator of a vector parameter θ is the value maximizing the likelihood function which is already a function of the components of θ.

The ML estimator formulation is given as follows [10]:

$$\hat{\theta}_{MLE} = \underset{\theta}{argmax}\, P(X;\theta) \tag{1}$$

Where $\hat{\theta}$ is the estimated vector parameter of θ, X is the observation data vector, $P(X; \theta)$ is the PDF of X depending on θ which has been parameterized on it.

3.2 Proposed Method

Suppose that the smart meter is located at an unknown position (x, y) and there are n receiver sensors located at known coordination (x_l, y_l), $1 \leq l \leq n$. The RSS localization method use the distance between the smart meter and sensors as the main measurements to obtain the position of the emitter. Friis Transmission formula is the most common one used in this model [12] as shown in (2).

$$\Omega_l = P_t + G_t + G_r + 10\alpha \log_{10}\left(\frac{\lambda}{4\pi d_l}\right) \tag{2}$$

where Ω_l is the RSS in (dBm) measured by the l-th sensor, P_t denotes the transmission power, G_t is the transmit gain, G_r is the received gain, α is the signal carrier frequency, called path loss exponent and d_l is the propagation distance. (2) can be rewritten as (3) if the received power is known at a close distance to the smart meter, called the log-distance path loss model [9], [12].

$$\Omega_l = C - 10\alpha \log_{10}\left(\frac{d_l}{d_0}\right) + n_l \tag{3}$$

where C is the received power at distance d_0, and d_l is the Euclidean distance between the smart meter and the receiver at (x_l, y_l), i.e.,

$$d_l = \sqrt{(x - x_l)^2 + (y - y_l)^2} \tag{4}$$

The effects of the shadowing is included in (3) denoted by n_l and considered to be a zero-mean Gaussian random variable with known covariance matrix. Let define the column vector θ and Ω as follow:

$$\theta = [x \quad y \quad c]^T \Omega = [\Omega_1 \quad \Omega_2 \quad . \quad . \quad . \quad \Omega_n]^T \tag{5}$$

Assume that $f(\theta|\Omega)$ is representative for the likelihood function of θ based on the RSS measurements, Ω_l, $1 \leq l \leq n$. Using the path loss model in (3), and considering a diagonal covariance matrix with standard deviation σ_l, $1 \leq l \leq n$, $f(\theta|\Omega)$ can be obtained in the following form [12]:

$$f(\theta|\Omega) = c_p \exp\left\{ -\sum_{l=1}^{n} \frac{\left\{ \Omega_l - C + 10\alpha \log_{10}\left(\frac{d_l}{d_0}\right) \right\}^2}{2{\sigma_l}^2} \right\} \tag{6}$$

where c_p is a positive constant and independent from the vector parameter θ. $\hat{\theta}$ which represents the ML estimator of θ can be obtained by solving the optimization problem as given below [9]:

$$\begin{aligned} \hat{\theta}_{MLE} &= \arg\max_{\theta} f(\theta|\Omega) \\ &= \text{argmin} \left\{ \sum_{l=1}^{n} \frac{\left\{ \Omega_l - C + 10\alpha \log_{10}\left(\frac{d_l}{d_0}\right) \right\}^2}{{\sigma_l}^2} \right\} \end{aligned} \tag{7}$$

The PSO algorithm is implemented to solve this non-linear optimization problem. The ML estimator returns the estimated position and the reference transmission power of the smart meter, i.e,

$$\left[\hat{\theta}_1 \quad \hat{\theta}_2 \quad \hat{\theta}_3 \right] = \left[\hat{x}_{rss} \quad \hat{y}_{rss} \quad \hat{c}_{rss} \right] \tag{8}$$

Without any prior knowledge about the position and the reference power, optimization problem must be done through Euclidean R^3 but most of the practical problems contain some prior information about the vector parameter. In this case, Bayesian philosophy can be employed to find an estimation of the unknown vector parameter.

3.3 Particle Swarm Optimization (PSO)

The PSO algorithm, first proposed by Kennedy and Eberhart [13], considered to be a method for optimization based on the social behavior of flocks of birds or schools of fish. The standard PSO algorithm first will start by generating random positions for the particles, within an initialization region. Initializing the velocities can be done through using values within that region or they can be selected zero or small random values to prevent particles from leaving the search space during the first iterations. During the main loop of the algorithm, the velocities and positions of the particles are iteratively updated until a stopping criterion is met. The update rules are completely described in [13].

4 Simulation Results

This section investigates the effectiveness of the proposed method through MATLAB simulations. Three scenarios are considered to scrutinize the effect of the variance, the number of the sensors, and the path loss exponent on the localization accuracy and the average MDE. Noise is assumed to be Additive White Gaussian Noise (AWGN). MDE is defined in the following form:

$$MDE = \sqrt{\left(\hat{x}_{rss} - x\right)^2 + \left(\hat{y}_{rss} - y\right)^2} \tag{9}$$

PSO parameters are, iteration number =100, population size = 30 and $C_1, C_2 = 2.05$.

4.1 Scenario1

This scenario presents the effect of the variance on the localization problem. 6 sensors are located at vertices of a regular hexagon with 20 meters diameter and (0,0) center. The smart meter position is at (1,-1) and the AWGN has the $N(0, \sigma^2)$ distribution. From Fig. 3, it can be inferred that as the variance increases, the average MDE will go up. It is rational since higher variances for the noise means that the observation data deviates from the real value in a way that estimating the location of the emitter will contain more errors. Table 1 illustrates the estimated values of the vector parameter θ w.r.t the variance.

4.2 Scenario 2

The effect of the number of the sensors has been inspected in this part. The smart meter position is at (1,-1) and the AWGN has the $N(0, 4)$ PDF distribution. Sensors are located around the smart meter at known coordination on the vertices of a square, hexagon, octagon and ten-sided, respectively. Fig. 4 depicts the variation of the average MDE w.r.t the number of the sensors. As it can be seen, increasing the number of the sensors will decrease the MDE error since the number of the observations will increase. As a result, the estimated vector parameter will be more accurate and will return the results close enough to the real values. Table 2 shows the estimated values of the parameter and indicates the direct proportional relationship between the sensors and the average MDE.

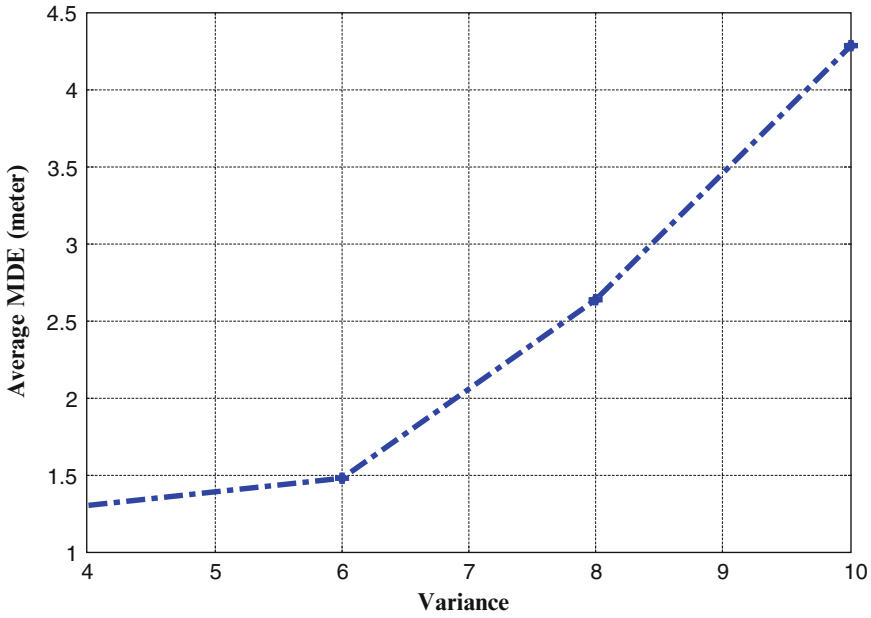


Fig. 3 Average MDE w.r.t the Variance

Table 1 Estimated parameters w.r.t the variance

Variance	$\hat{x}_{rss}$ (m)	$\hat{y}_{rss}$ (m)	$\hat{c}_{rss}$ (dBm)	MDE(m)
4	1.2777	-2.2742	27.0011	1.3041
6	0.0917	-2.1638	27.0599	1.4775
8	2.9358	-2.7934	26.9991	2.6381
10	1.8748	-3.1882	27.0037	4.2786

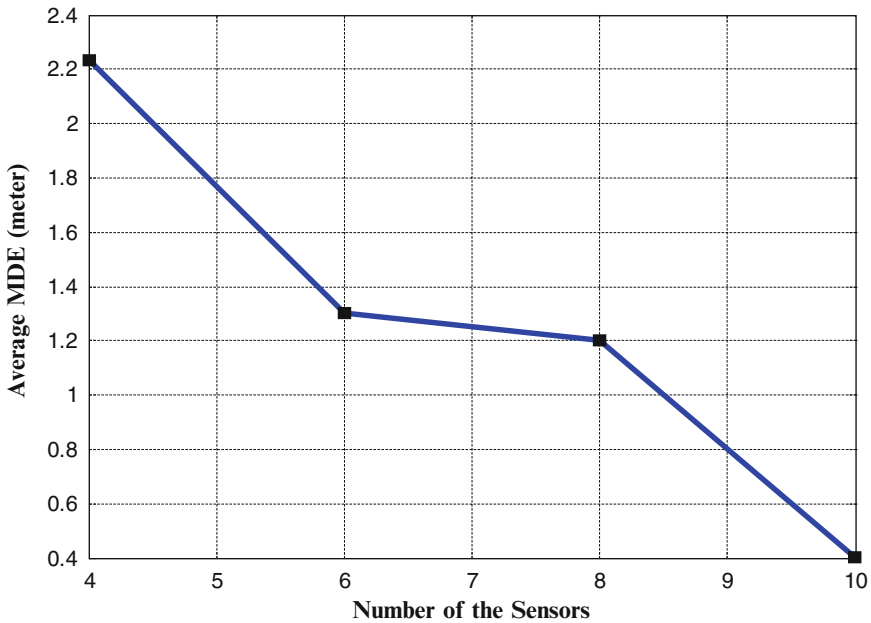


Fig. 4 Average MDE w.r.t the Number of the Sensors

4.3 Scenario 3

The results for the impact of the path loss exponent on the localization problem and the average MDE are given in this part. 6 sensors are located around the smart meter and the AWGN has the $N(0, 4)$ PDF distribution. Fig. 5. illustrates the variation of the average MDE w.r.t the path loss exponent. In the practical applications, the value of the path loss exponent ranges from 2 to 6.

It can be seen that the average MDE is inversely proportional to the path loss exponent. This is because as the path loss exponent increases, the PDF will get sharper so the estimation will be more accurate and probable. Table 3. represents the estimated values of the parameter θ in this scenario.

Table 2 Estimated parameters w.r.t the Number of Sensors

Sensors	$\hat{x}_{rss}$ (m)	$\hat{y}_{rss}$ (m)	$\hat{c}_{rss}$ (dBm)	MDE(m)
4	2.3278	0.7945	26.9971	2.2323
6	1.2777	-2.2742	27.0011	1.3041
8	0.5313	0.1052	27.0048	1.2005
10	1.3774	-0.8614	26.9988	0.4021

Table 3 Estimated parameters w.r.t the path loss exponent

Path Loss Exponent	$\hat{x}_{rss}$ (m)	θ(m)	MDE(m)
2	1.2777	-2.2742	1.3041
3	1.9293	-0.7430	0.9641
4	1.3646	-0.3421	0.7521
5	1.0373	-1.5144	0.5158

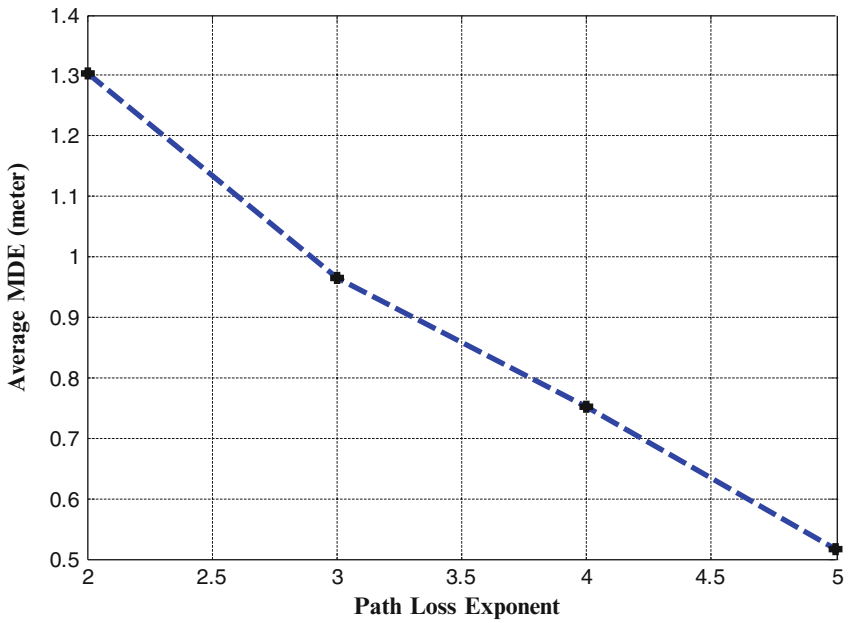


Fig. 5 Average MDE w.r.t the Path Loss Exponent

5 Conclusion

In this paper, the RSS-based localization of a smart meter is investigated as one of the AMI system's security aspects. The architecture of the studied AMI system is introduced and the included components are explained. Under the assumption of a log-normal path loss model and the AWGN shadowing, the ML estimator is implemented to estimate the position and the reference transmission power of the smart meter. Proposed method is verified through MATLAB simulations and the effect of the variance, number of the sensors and the path loss exponent are also inspected. Results show that the average MDE increases as the variance goes up and reduces as the number of the sensors increase. Additionally, it can be deduced that the average MDE is inversely proportional to the path loss exponent.

References

1. I. H. Cavdar. A solution to remote detection of illegal electricity usage via power line communications. IEEE Transactions on Power Delivery, 19(4):1663–1667, Oct. 2004.
2. F. Cleveland. Cyber security issues for advanced metering infrastructure (AMI). In Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, 2008.
3. Aravinthan, Visvakumar, et al. "Wireless AMI application and security for controlled home area networks." *Power and Energy Society General Meeting, 2011 IEEE*.
4. E. I. A, U.S. Energy Information Administration, Independent Statistics & Analysis, Available Online: http://www.eia.gov/.
5. Rouf, Ishtiaq, et al. "Neighborhood watch: Security and privacy analysis of automatic meter reading systems." Proceedings of the 2012 ACM conference on Computer and communications security. ACM, 2012.
6. Ishtiaq Roufa, et al. "A Practical Study of Security and Privacy Issues in Automatic Meter Reading System." *IEEE Spectrum*, October 2010.
7. M. Lisovich and S. Wicker, "Privacy concerns in upcoming residential and commercial demand-response systems," in *2008 Clemson* University Power Systems Conference. Clemson University, 2008.
8. P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," *IEEE Security and Privacy*, no. 3, pp. 75–77, 2009.
9. Taylor, R. C. (2013). "Received Signal Strength-Based Localization of Non-Collaborative Emitters in the Presence of Correlated Shadowing" (Doctoral dissertation, Virginia Polytechnic Institute and State University).
10. Kay, Steven M. "Fundamentals of Statistical signal processing," Vol 2: Detection theory. Prentice Hall PTR, 1998.
11. Sargolzaei, Arman, Kang K. Yen, and M.N. Abdelghani. "Time-Delay Switch Attack on Load Frequency Control in Smart Grid." Advances in Communication Technology, Vol.5 (2013), 55–64.
12. Sichun Wang, Robert Inkol, and Brad R. Jackson. "Relationship between the maximum likelihood emitter location estimators based on received signal strength (rss) and received signal strength diference (rssd)". In Communications (QBSC), 2012 26th Biennial Symposium on, pages 64–69.
13. J. Kennedy and R.C. Eberhart (1995), Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks, Perth, Australia, IEEE Service Center, Piscataway, NJ, 4, pp. 1942–1948.

Active/Reactive Power Control of Three Phase Grid Connected Current Source Boost Inverter Using Particle Swarm Optimization

Arman Sargolzaei, Mahdi Jamei, Kang Yen, Arif I. Sarwat, and Mohamed Abdelghani

1 Introduction

The future is for sustainable energy (SE) sources [1, 2]. However, to utilize SE sources efficiently one needs to integrate them into the AC Power-Grid. To do so, interface electronic circuits are needed to connect SE sources to local loads or directly to the utility grid. These interfaces are known as the SE-Grid inverters. SE-Grid inverters convert direct current (DC) to alternating current (AC). They are designed with the ability to synchronize their outputs with the utility line and have as little harmonic content as possible.

Inverters come in several different topologies: voltage source inverters (VSIs) [3], current source boost inverters (CSBIs) [4–6], multi-level inverters [7, 8], and matrix converters [9, 10]. VSIs have replaced CSBIs in many industrial applications [11]. However, unlike other inverters CSBIs are able to invert and boost current in a single stage and transferring power from low a DC source, for example, photovoltaic and fuel cells, to a larger AC voltage [12]. This makes CSBI one of the best choices for SE-Grid conversion systems.

Here we will consider the control of active/reactive power in CSBIs with the switching pattern and control circuit which has been proposed in [13–15]. The proposed CSBI provides more robust inverter to control the voltage and the injected power. Different approaches to control active and reactive power were proposed: sliding mode control [16], fuzzy logic-based method [17] and predictive control [18]. All these control methods minimize the instantaneous errors in active and reactive power by controlling the input voltage. However, these methods are susceptible to variation of power line inductance and suffer from changing switching frequencies. In [19] and [20], direct power control technique was improved to have a constant switching frequency. In [21], the space vector modulation has been applied to a direct power controlled inverter for the purpose of having a simpler control system with fewer harmonics. One of the application to control active power in the future power system is Plug-In Electric Vehicle (PIEV) which can serve as flexible load [29].

PID control is commonly used in industrial control systems (e.g. SCADA systems and RTU's) [22, 23]. PID minimizes the error which is the difference between a measured and a desired signal, by adjusting the process control inputs [24]. The PID controller involves three parameters: a term proportional to the error, another to the integral of error and the third to the derivative of error. The parameters of the PID controller have to be found through a design process involving the optimization of a set of desired requirements like settling time, rise time etc.

Particle swarm optimization (PSO) has been developed by Kennedy and Eberhart [25]. PSO is an optimization algorithm that finds optimal solution to a problem by iteratively improving candidate solutions based on a measure of quality. It has many different applications [26–28]. PSO starts with a population of particles (candidate solutions) and by changing the particles 'locations' and 'momenta' in the search-space it explores better solutions of the problem. The particles' locations and momenta are updated based on its local relations to other particles and by the objective function that determines the quality of solution. The process of updating the particles locations and momenta moves the swarms of particles to the optimal solution. Particles Swarm Optimization (PSO) technique is a powerful and simple method that can be used to find the PID controller parameters as to optimize design requirements. PSO-PID has been used successfully in control of automatic voltage regulators [28] and control of the linear brushless DC motor [27].

A. Sargolzaei (✉) • M. Jamei • K. Yen • A.I. Sarwat
Department of Electrical and Computer Engineering, Florida International University, Florida, USA

M. Abdelghani
Department of Mathematical and Statistical Science, University of Alberta, Alberta, Canada

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_21,

In this paper we are going to use the PSO-PI method to control of three-phase CSBIs with state vector pulse-width modulation (SVPWM) switching pattern. The paper is organized as follows; first we introduce three phase inverter circuit topology and switching mechanism. Then we introduce the PSO-PI method of control for CSBI inverters and present simulation results. Finally we discuss the pros and cons of this method and future work.

2 The Circuit Topology and Review of Switching Pattern

This section briefly describes the power topology of a three-phase CSBI and the switching pattern which has been presented for the first time in [13]. The power topology and switching arrangement has been offered for the purpose of continuity. Fig. 1 shows the schematic of the single-stage CSBI. In this figure, V_i is provided from the PV source, L_i is an inductor which has been used as a dc-link, AC side film capacitors are described by C_o and finally L_o described the line inductors. Like all three phase inverters, the main goal of this switching pattern and circuit topology is to produce sinusoidal currents to loads and also the grid. They present six sectors in each voltage cycle to inject in-phase sinusoidal currents into the three phases of the grid [14].

The switching pattern of the three stages is [14]:

1. During the charging time period (1st stage) two switches are on in leg "1", the dc inductor will be charged and subsequently the current increases.
2. During period of the first discharging time (2nd stage) V_{12} will be applied to output of the inverter; the dc current will be injected into first phase, and will be drawn from second phase. Hence, the first leg higher switch of and the second leg bottom switch are on.
3. During the second discharging time period (3rd stage) V_{13} will be applied to output of the inverter; the dc current will be injected into first phase, and is drawn from third phase. In this case, both the first leg upper switch and the third leg bottom switch are on.

The switching pattern on all sectors can be found in Table 1.

Here t_c is charging time, t_{d1} and t_{d2} are time intervals of discharging. The relationship of these time intervals gives

$$T_s = t_c + t_{d1} + t_{d2} \tag{1}$$

Based on [15] the following expression will be resulted for first sector

$$V_{dc}T_s = v_{ab}t_{d1} + v_{ac}t_{d2} \tag{2}$$

Above equation can be written in expressions of duty cycle

$$V_{dc} = v_{ab}d_1 + v_{ac}d_2 \tag{3}$$

where d_1 and d_2 can be computed by

$$d_1 = \frac{2V_{dc}}{3V_m}\sin\left(\frac{\pi}{3} - \omega t\right) \tag{4}$$

Table 1 The switching pattern [14]

Sector	I	II	III	IV	V	VI
	ν_{12}	ν_{13}	ν_{23}	ν_{21}	ν_{31}	ν_{32}
	ν_{13}	ν_{23}	ν_{21}	ν_{31}	ν_{32}	ν_{12}
S_{11}	T_s	t_{d1}	0	t_c	0	t_{d2}
S_{12}	t_c	0	t_{d2}	T_s	t_{d1}	0
S_{21}	0	t_{d2}	T_s	t_{d1}	0	t_c
S_{22}	t_{d1}	0	t_c	0	t_{d2}	T_s
S_{31}	0	t_c	0	t_{d2}	T_s	t_{d1}
S_{32}	t_{d2}	T_s	t_{d1}	0	t_c	0

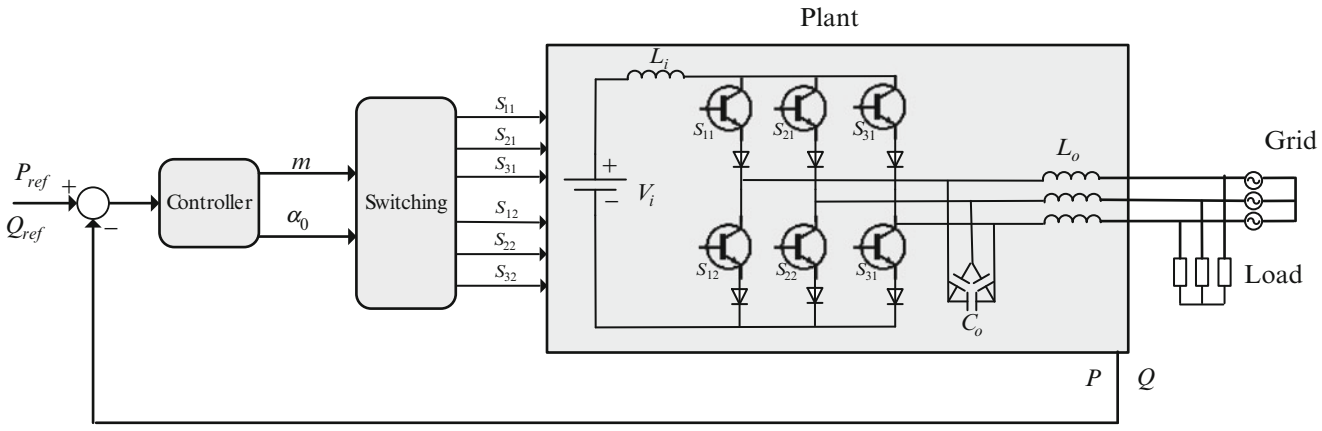


Fig. 1 The Proposed Circuit Schematic

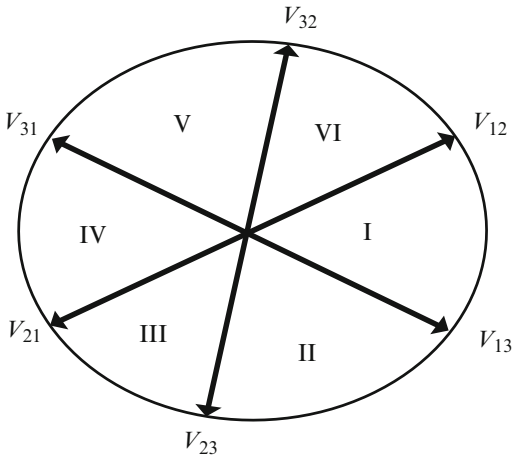


Fig. 2 Sectors of the Proposed Switching

$$d_2 = \frac{2V_{dc}}{3V_m}\sin(\omega t) \tag{5}$$

where the grid neutral voltages peak value is denoted by V_m.

For the case of grid connected with the reference signal measured as $v_{ab} = \sqrt{3}V_m\cos(\omega t)$, The duty cycles (4) and (5) can be computed by

$$d_1 = \frac{2}{3}\frac{V_{dc}}{V_m\cos\left(\alpha_0 + \frac{\pi}{6}\right)}\cos(\omega t - \alpha_0) \tag{6}$$

$$d_2 = \frac{2}{3}\frac{V_{dc}}{V_m\cos\left(\alpha_0 + \frac{\pi}{6}\right)}\cos\left(\omega t - \alpha_0 - \frac{2\pi}{3}\right) \tag{7}$$

For a full power control the charging and discharging time intervals will be described as

$$d_1 = m\cos(\omega t - \alpha_0) \tag{8}$$

$$d_2 = m\cos\left(\omega t - \alpha_0 - \frac{2\pi}{3}\right) \tag{9}$$

$$d_c = 1 - d_1 - d_2 \tag{10}$$

where α_0 is denoted the phase shift according to line-line reference voltage, and the modulation index is shown as m which should be between zero and one.

3 Control Formulation

In this section, we briefly study the relationship between the input current with the modulation index, m, and the angle, α_0 between the grid voltage and the desired inverter output current to control the active and reactive power. The modulation index and phase angle are our circuit parameters and has been presented in equations (8) and (9) respectively. The reactive and active power can be controlled using the angle α_0 and the coefficient m.

Based on the switching pattern aforesaid [13], the inverter output current of phase-A can be described as

$$i_{a1}(t) = A(m)\cos(\omega t - \Psi(\alpha_0)) \tag{11}$$

The active power for three-phase case can be expressed by

$$P = 0.5\sqrt{3}V_m A(m)\cos(\Psi(\alpha_0)) \tag{12}$$

And the reactive power

$$Q = Q_s + 0.5\sqrt{3}V_m A(m)\sin(\Psi(\alpha_0)) \tag{13}$$

where, Q_s is the reactive power generated by AC capacitors shown in Fig.1. This reactive power can be approximated by [15]

$$Q_S = 18\pi f C_s\left(\frac{V_m}{\sqrt{2}}\right)^2 \tag{15}$$

Two optimal PI controllers have been designed using PSO to control above parameters.

$$\begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} = \begin{bmatrix} K_{P1}e_1(t) + K_{I1}\int e_1(t)dt \\ K_{P2}e_2(t) + K_{I2}\int e_2(t)dt \end{bmatrix} \tag{16}$$

where $u_i(t)$ are the control inputs, K_{Pi} and K_{Ii} are proportional and integral coefficient of PI controllers and finally e_i are the difference between the actual and the desired active and reactive power.

Based on the details in [15], we can calculate the input power using the modulation index and other parameters of the circuit,

$$P_{in} = V_{dc}I_{dc} = \frac{V_{dc}^2 - (V_{SW} + g(m, V_m))V_{dc}}{R_{dc} + R_{SW}} \tag{17}$$

where g is a monotone function of the modulation index and peak value of the output line-neutral voltage, $R_{SW} = 2R_{Diode} + 2R_{IGBT}$ and $V_{SW} = 2V_{Diode} + 2V_{IGBT}$ can be found. It clearly shows that when the modulation index decreases, the power loss increases.

4 Particle swarm optimization (PSO)

Particle swarm optimization (PSO) is a method of solving continuous and discrete optimization problems with a population-based stochastic approach. Kennedy and Eberhart proposed this algorithm in 1995 [25] as a novel

heuristic and computational technique. The general PSO algorithm will first generate particles with randomly assigned positions, within an initialization area. Initialization of the velocities is usually done within that region. Particles can be initialized to zero or to small random values in order to avoid them from leaving the search space during the initial iterations. During the main loop, the algorithm will continue to update the velocities and positions of the particles until a stopping criterion is met. The update rules can be calculated as follows:

$$V_i^{k+1} = wV_i^k + C_1 rand_1 + \left(Pbest_i^k - X_i^k\right) + C_2 rand_2 \times \left(Gbest_i^k - X_i^k\right) \quad (18)$$

$$X_i^{k+1} = X_i^k + V_i^{k+1} \quad (19)$$

where $X_i = [X_{i1}, X_{i2}, \ldots, X_{id}]$ and $V_i = [V_{i1}, V_{i2}, \ldots, V_{id}]$ are the position and velocity of the particle i. Let $Pbest_i = [X_{i1\ Pbest}, X_{i2\ Pbest}, \ldots, X_{id\ Pbest}]$ and $Gbest_i = [X_{1\ Gbest}, X_{2\ Gbest}, \ldots, X_{d\ Gbest}]$ be the best position of particle i and its neighbors best position so far, respectively.

The value of the parameters such as w, C_1, C_2 should be determined in advance in velocity updating process. As the iteration proceeds, the inertia weight w is linearly decreasing.

$$w = w_{\max} - \frac{(w_{\max} - w_{\min})iter}{iter_{\max}} \quad (20)$$

5 Proposed Method

Two PI controllers have been employed to control the active and reactive power independently but simultaneously. For this purpose, PSO has been used to find the optimum values for the parameters of the PI controllers. The multi-objective cost function is to minimize the summation of the Integral Time Absolute Error (ITAE) of the active and reactive power. Sub-objective functions should be normalized to be comparable within the range of [0, 1] and the weighted based importance and priorities. Finally, the cost function can be represented as follows:

$$\min f = \omega_1 f_1 + \omega_2 f_2 \quad (21)$$

where ω_i $(i = 1, 2)$ are the weighting factors which determine the contribution of each term in the cost function and f_i $(i = 1, 2)$ are the normalized ITAE of the active and reactive power, respectively. Weighted sum method is used to select the involvement factors.

Each particle, X_i, in the PSO algorithm has four dimensions, $d = 4$, which are corresponding to the proportional gain, and integral gain of the active power controller as well as the proportional gain, and integral gain of the reactive power controller, respectively. The PSO algorithm will first randomly assign values to the parameters of the two controllers. Then, the actual active and reactive power injected to the grid will be compared with the desired values to find the errors. As a result, the cost function can be computed and particles will be updated in each iteration to minimize the ITAE of the active and reactive powers. The algorithm will continue until the stopping criterion, the iteration number in this paper, is met.

6 Simulation and Results

The simulation results are presented in this section to show the effectiveness of the proposed method of designing the PI controllers based on the PSO algorithm. The circuit parameters in this simulation have been selected as: $V_{Pv}(MPP) = 70v$, $C_o = 16.8\mu F$, $L_i = 6mH$, $L_o = 1.2mH$, the PWM switching frequency is set to be 8 kHz and the line-to-line rms voltage value is 207.846 v. The simulations are carried out using MATLAB Simulink. PSO parameters are, iteration number = 100, population size = 30, and $C_1 = C_2 = 2.05$.

The behavior of the cost function is shown in Fig. 3. As it can be seen, the algorithm has successfully minimized the cost function to find the optimum values of the controllers. The dynamic response of the injected active and reactive power are plotted over 6 seconds in Fig. 4 (a) and Fig. 4 (b), respectively. To evaluate the efficacy of the controllers in tracking the reference command, the active power, P_{ref}, will instantly change at $t = 0.75$ sec from 250 to 350 watts. Since the PV farm can be considered as a reactive power source for the power system, the controller should also tolerate sudden changes in the reactive power. For this purpose, the reactive power, Q_{ref}, will change at $t = 3$ sec

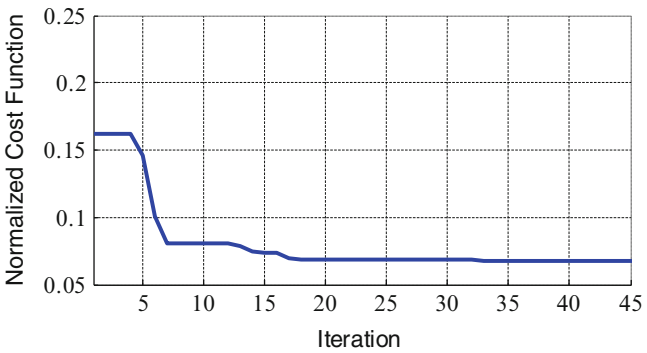


Fig. 3 Normalized Cost Function w.r.t the Number of Iterations

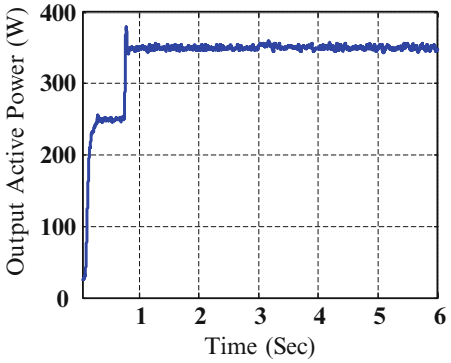


Fig. 4 (a) Active Power Output (W)

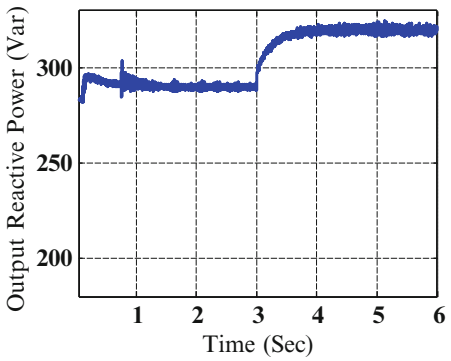


Fig. 4 (b) Reactive Power Output (Var)

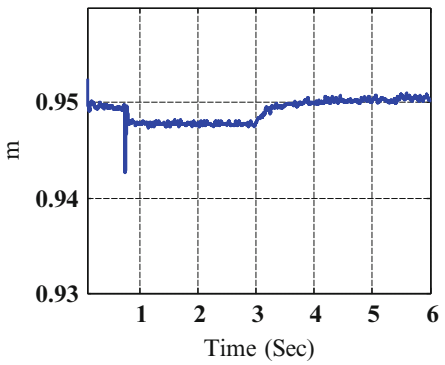


Fig. 4 (c) The Coefficient m

from 290 to 320 vars. Variations of m and α_0 have been illustrated in Fig. 4 (c) and Fig. 4 (d).

As it can be observed, the controllers can effectively trace the active and reactive power input without harming the stability of the system. The dynamic behavior of the P and Q are fast enough for a grid-connected PV farm and also keeping an acceptable range of overshoot.

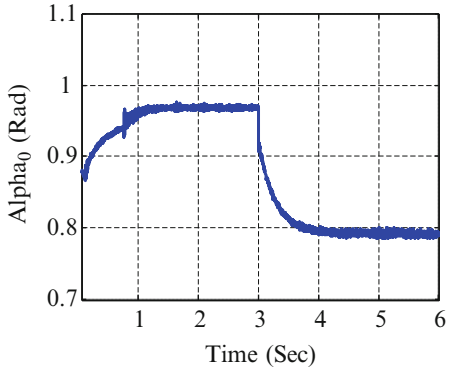


Fig. 4 (d) The Angle Alpha0

7 Conclusion

Here, we have used particle swarm optimization (PSO) method to find the optimal parameters of a PID controller. We applied this method to design an optimal PID controller for active and reactive power in a three phase grid connected current source boot inverter (CSBI). Simulation results show that our PSO-PI parameters selection method leads to better performance. With PSO-PID we were able to control active and reactive powers of this CSBI simultaneously. We also achieved lower power overshoot in comparison to PID control without the use of the PSO.

References

1. Gholami, A., Jamei, M., Ansari, J., & Sarwat, A. I. (2014). Combined Economic and Emission Dispatch Incorporating Renewable Energy Sources and Plug-In Hybrid Electric Vehicles. International Journal of Energy Science, 4(2).
2. M.H. Amini, B. Nabi and M-R. Haghifam, "Load management using multi-agent systems in smart distribution network," *IEEE PES General Meeting*, Vancouver, BC, Canada, 2013.
3. Wang, F., & Shen, W. (2009). Voltage source inverter. *IEEE Industry Applications Magazine, 15*(2).
4. Maeda, T., Matsuda, Y., & Matsumura, T. (1979). Current source inverter: Google Patents.
5. Sato, S. (2010). Current source inverter: Google Patents.
6. Shen, D., & Lehn, P. (2002). Modeling, analysis, and control of a current source inverter-based STATCOM. *Power Delivery, IEEE Transactions on, 17*(1), 248-253.
7. Malinowski, M., Gopakumar, K., Rodriguez, J., & Perez, M. A. (2010). A survey on cascaded multilevel inverters. *Industrial Electronics, IEEE Transactions on, 57*(7), 2197-2206.
8. Rodriguez, J., Lai, J.-S., & Peng, F. Z. (2002). Multilevel inverters: a survey of topologies, controls, and applications. *Industrial Electronics, IEEE Transactions on, 49*(4), 724-738.
9. Kandasamy, K., & Sahoo, S. K. (2013). A Review of Matrix Converter and Novel Control Method of DC-AC Matrix Converter. *ijm, 1*(3), 1.

10. Sahoo, A. K., Basu, K., & Mohan, N. (2013). *Comparison of filter components of back-to-back and matrix converter by analytical estimation of ripple quantities*. Paper presented at the Industrial Electronics Society, IECON 2013-39th Annual Conference of the IEEE.
11. Vander Meulen, A., & Maurin, J. (2010). Application Engineers. *Current source inverter vs. Voltage source inverter topology*.
12. Klumpner, C. (2007). A new single-stage current source inverter for photovoltaic and fuel cell applications using reverse blocking IGBTs. Paper presented at the Power Electronics Specialists Conference, 2007. PESC 2007. IEEE.
13. Mirafzal, B., Saghaleini, M., & Kaviani, A. K. (2011). An SVPWM-based switching pattern for stand-alone and grid-connected three-phase single-stage boost inverters. *Power Electronics, IEEE Transactions on, 26*(4), 1102-1111.
14. Saghaleini, M. (2012). Switching Patterns and Steady-State Analysis of Grid-Connected and Stand-Alone Single-Stage Boost-Inverters for PV Applications.
15. Saghaleini, M., & Mirafzal, B. (2011). *Power control in three-phase grid-connected current-source boost inverter*. Paper presented at the Energy Conversion Congress and Exposition (ECCE), 2011 IEEE.
16. Hu, J., Shang, L., He, Y., & Zhu, Z. (2011). Direct active and reactive power regulation of grid-connected DC/AC converters using sliding mode control approach. *Power Electronics, IEEE Transactions on, 26*(1), 210-222.
17. Bouafia, A., Krim, F., & Gaubert, J.-P. (2009). Fuzzy-logic-based switching state selection for direct power control of three-phase PWM rectifier. *Industrial Electronics, IEEE Transactions on, 56* (6), 1984-1992.
18. Cortés, P., Rodríguez, J., Antoniewicz, P., & Kazmierkowski, M. (2008). Direct power control of an AFE using predictive control. *Power Electronics, IEEE Transactions on, 23*(5), 2516-2523.
19. Hu, J., Shang, L., He, Y., & Zhu, Z. (2011). Direct active and reactive power regulation of grid-connected DC/AC converters using sliding mode control approach. *Power Electronics, IEEE Transactions on, 26*(1), 210-222.
20. Zhi, D., Xu, L., & Williams, B. W. (2009). Improved direct power control of grid-connected DC/AC converters. *Power Electronics, IEEE Transactions on, 24*(5), 1280-1292.
21. Restrepo, J. A., Aller, J. M., Viola, J. C., Bueno, A., & Habetler, T. G. (2009). Optimum space vector computation technique for direct power control. *Power Electronics, IEEE Transactions on, 24*(6), 1637-1645.
22. Sargolzaei, A., Yen, K. K., & Abdelghani, M. N. (2014). Control of Nonlinear Heartbeat Models under Time-Delay-Switched Feedback Using Emotional Learning Control. International Journal on Recent Trends in Engineering & Technology, 10(2).
23. Sargolzaei, A., Yen, K. K., & Abdelghani, M. N. (2013). Time-Delay Switch Attack on Load Frequency Control in Smart Grid. *Advances in Communication Technology*, *5*, 55-64.
24. Åström, K. J., & Hägglund, T. (2006). *Advanced PID control*: ISA-The Instrumentation, Systems, and Automation Society; Research Triangle Park, NC 27709.
25. Kennedy, J., & Eberhart, R. (1995). *Particle swarm optimization*. Paper presented at the Proceedings of IEEE international conference on neural networks.
26. Sargolzaei, A., Faez, K., & Sargolzaei, S. (2011, May). A new method for Foetal Electrocardiogram extraction using Adaptive Nero-Fuzzy Interference System trained with PSO algorithm. In *Electro/Information Technology (EIT), 2011 IEEE International Conference on* (pp. 1-5). IEEE.
27. Nasri, M., Nezamabadi-Pour, H., & Maghfoori, M. (2007). A PSO-based optimum design of PID controller for a linear brushless DC motor. *World Academy of Science, Engineering and Technology, 26* (40), 211-215.
28. Gaing, Z.-L. (2004). A particle swarm optimization approach for optimum design of PID controller in AVR system. *Energy Conversion, IEEE Transactions on, 19*(2), 384-391.
29. Mohammadhadi Amini, Arif I. Sarwat. Optimal Reliability-based Placement of Plug-In Electric Vehicles in Smart Distribution Network. International Journal of Energy Science, 2014, 4(2), 43-49. doi: 10.14355/ijes.2014.0402.02

Anti-Islanding Test Results for Multiple PV Inverter Operations

Byunggyu Yu and Youngseok Jung

1 Introduction

Distributed generation (DG) system like photovoltaic (PV) and wind is growing larger and more complicated in order to avoid environmental problem and fossil fuel shortage crisis. Especially, PV is known as a method for generating electric power by using solar cells to convert energy from the sunlight into electricity. Since the input energy source, sunlight, is free, unlimited, nonpolluting, and available all over the vast surface of the earth, it has been believed that photovoltaic generation can be the solution to limited fossil fuel problem and its environmental pollution problem [1].

Under these circumstances, photovoltaic has been regarded as a key technology to be a possible solution. Especially, regardless for what reasons and how fast the oil price and energy prices will increase in the future, photovoltaic is the one to offer a reduction of price rather than an increase in the future [2]. With this potential cost competitiveness feature, the photovoltaic market continues to be driven by government incentives. However, there have been concerns about the rapid dissemination of utility interactive PV system on the safety issue, power quality, reliability, stability and so on [3]-[5]. This is because the conventional electric power systems were not designed to accommodate active power generation from distributed generation like photovoltaic [5].

The most issued safety problem by PV generation is islanding phenomenon which PV has an independent powering to a portion of the utility system even though the portion has been disconnected from the remainder of the utility source. As a safety problem, islanding is undesirable for several reasons. First, it can cause safety hazard to utility service personnel who may be unaware that the isolated utility section is being energized by an unknown PV generation. Second, while the utility section connected to photovoltaic is isolated from the remainder of the utility section, a phase difference between the utility source voltage and PV output voltage can arise. If the utility source were attempt to reconnect to the isolated utility section, the large overcurrent could be generated by the unsynchronized reconnection between PV output voltage and the utility source and it could damage the PV system or customer load within the isolated section [1]. Out of phase resclosing, if occurs at a voltage peak, will generate a very severe capacitive switching transient and in a lightly damped system, the crest over-voltage can approach three times rated voltage [5]. Consequently, utility companies and PV system owners require that the grid-connected PV systems include the non-islanding inverters [4, 5].

Until now, various anti-islanding methods have been proposed [6]. However, these are mostly focus on the islanding prevention method for single PV system. Most of them are focusing on the anti-islanding performance of single PV system according to the related international and domestic standard test procedures [4–6]. A few studies on the islanding phenomenon for multiple photovoltaic operations in parallel were made [7]. This paper presents performance analysis of anti-islanding functions for multiple grid-connected PV systems by using commercial products.

2 Test Procedure for Anti-Islanding

Since the islanding phenomenon of DG including PV can cause the safety problems to the utility engineer or utility facilities, the DG is required to meet the anti-islanding standards for evaluating anti-islanding capability. In other words, these standards can be important guidelines to evaluate the developed anti-islanding methods. In particular, IEEE standard 1547, IEEE standard 929, the IEC standard

B. Yu (✉)
Kongju National University, Chungnam, Repuplic of Korea
e-mail: bgyuyu@kongju.ac.kr

Y. Jung
Korea Institute of Energy Research, Daejeon, Repuplic of Korea
e-mail: Jung96@kier.re.kr

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_22,

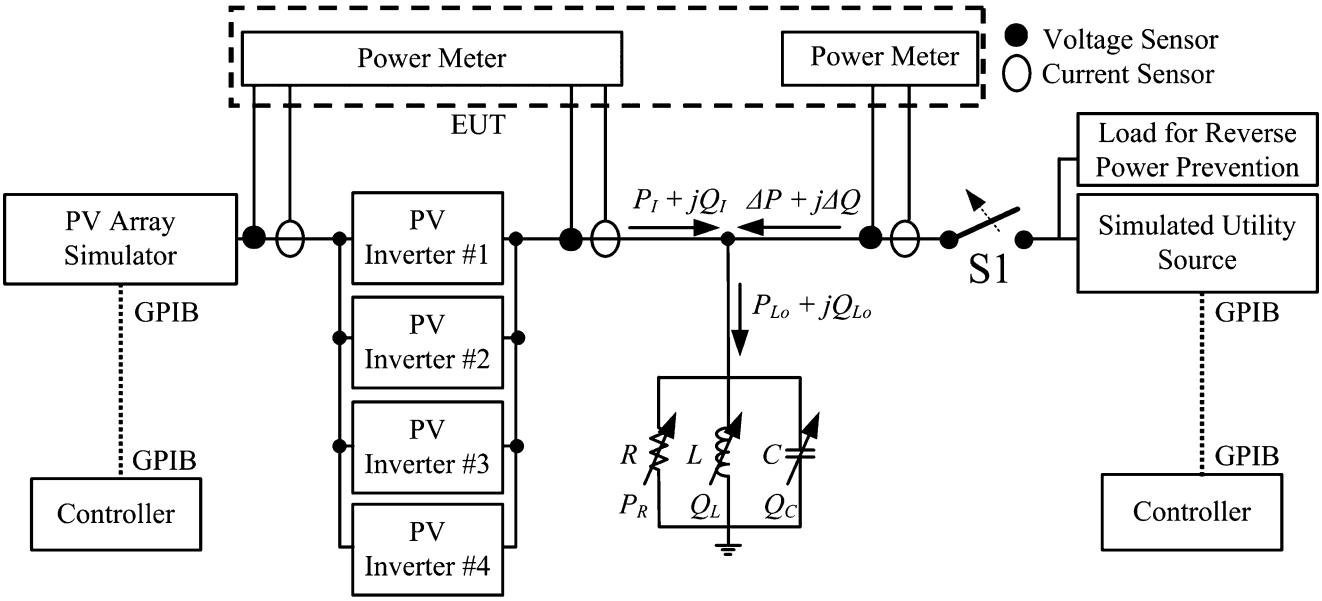


Fig. 1 Test circuit for islanding detection in a PV inverter

62116, Japanese Standard, and Korean Standard are worth considering.

According to these standards, Fig. 1 shows the typical test circuit for islanding detection capability in a PV inverter. In the Fig. 1, the quality factor Q_f, which is very critical parameter of islanding test condition, is defined as the strength of resonance of the islanding test load defined as (1).

$$Q_f = R\sqrt{\frac{C}{L}} = \sqrt{\frac{Q_L \times Q_C}{P_I}} \tag{1}$$

where R, L, C are the local resistive/inductive, capacitive loads in Fig. 1, P_I is the active output power, Q_L is the inductive reactive power, and Q_C is the capacitive reactive power.

For IEC standard 62116, which is similar with IEEE Std. 1547, a condition value of quality factor $Q_f = 1$ was chosen based on the calculated results of the ratio of the contract demand [kW] at the 723 measurement points in Japan to the installed shunt capacitor [kVar] needed to make the power factor 1 at that point [8]. In this paper, the IEEE Std. 1547 is used to evaluate the anti-islanding test results for multiple commercial PV inverters because it is widely used for certification throughout the world.

In order to analyze the anti-islanding performance for multiple commercial PV inverters, two different popular PV inverters in the recent commercial PV inverter market has been chosen and tested, as shown in Table 1. Fig. 2 shows the front-view of islanding testing facility for the evaluation.

Table 1 Commercial PV inverter products to be tested for multiple operations

Manufacturer	Model A	Model B
Power rating	10 kW	16.5 kW
Isolation	60 Hz isolation transformer	No isolation
Output voltage	380 V, 60Hz	380 V, 60Hz
Phase	3 phase, 4 wire	3 phase, 4 wire

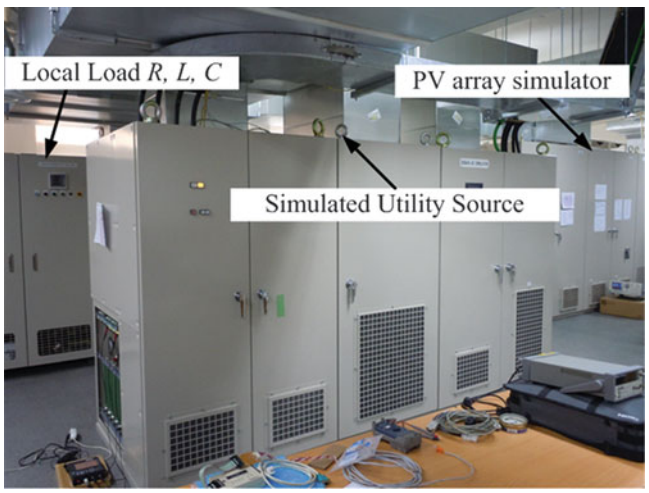


Fig. 2 Front-view of islanding testing facility

By using the islanding testing facility, these PV inverters are tested according to IEEE Std. 1547. Firstly, the local loads R, L, C in Fig. 1 are tuned to make zero power to the utility like both $\Delta P = 0$ and $\Delta Q = 0$ under different Cases in Table 2.

Table 2 Testing scenarios

	Tested PV inverters	Testing local load
Case 1	Single "model A" PV inverter	$P_R = 10$ kW $Q_L = 10$kVA $Q_C = 10$kVA
Case 2	Single "model B" PV inverter	$P_R = 16.5$ kW $Q_L = 16.5$kVA $Q_C = 16.5$kVA
Case 3	Two "model A" PV inverters	$P_R = 20$ kW $Q_L = 20$kVA $Q_C = 20$kVA
Case 4	Two "model B" PV inverters	$P_R = 33$ kW $Q_L = 33$kVA $Q_C = 33$kVA
Case 5	Two "model A" PV inverters together with single "model B" PV inverter	$P_R = 36.5$ kW $Q_L = 36.5$kVA $Q_C = 36.5$kVA
Case 6	Two "model B" PV inverters together with single "model A" PV inverter	$P_R = 43$ kW $Q_L = 43$kVA $Q_C = 43$kVA
Case 7	Two "model B" PV inverters together with two "model A" PV inverters	$P_R = 53$ kW $Q_L = 53$kVA $Q_C = 53$kVA

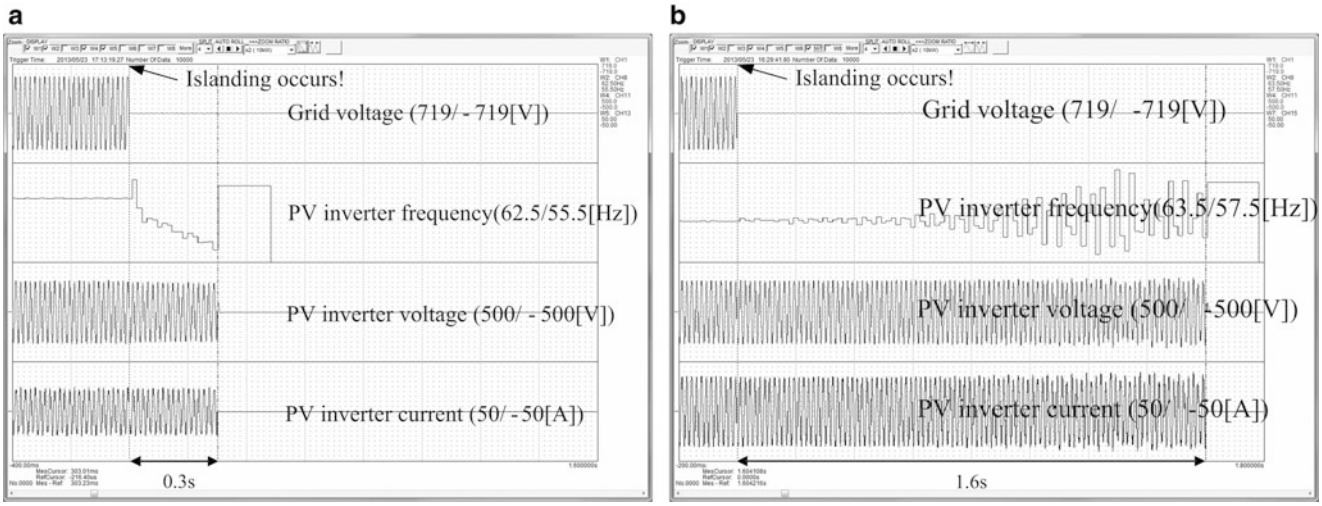


Fig. 3 Islanding test result for single PV inverter. (**a**) Under case 1, (**b**) Under case 2

3 Experimental Results for Multiple PV Operations

Islanding test results are conducted with several testing scenarios as shown in Table 2 according to IEEE Std. 1547. First, a single commercial PV inverter is tested for anti-islanding. Then, two PV inverters with a few combinations are tested in addition to the other single PV inverter model. Finally, totally four PV inverters with two different models are evaluated for islanding detection capability.

From Fig. 3 to Fig. 7, these show the main operational waveforms for islanding performance test according to IEEE Std. 1547 like grid voltage, frequency of PV inverter output voltage, PV inverter output voltage, and PV inverter output current.

Fig. 3 shows the islanding detection test performance for single PV inverter under case 1 and case 2. Single model A PV inverter can detect islanding within 0.3 s by drifting the PV inverter frequency to under frequency relay, as shown in Fig. 3 (a). In the same way, Single model B PV inverter can also detect islanding within 1.6 s by varying the PV inverter frequency periodically with positive feedback.

According to IEEE Std. 1547, islanding should be prevented within 2 s. Thus, these PV inverters under both case 1 and case 2 can meet the requirement for islanding detection time. Fig. 4 shows the islanding detection evaluation results while two PV inverters are operated in parallel.

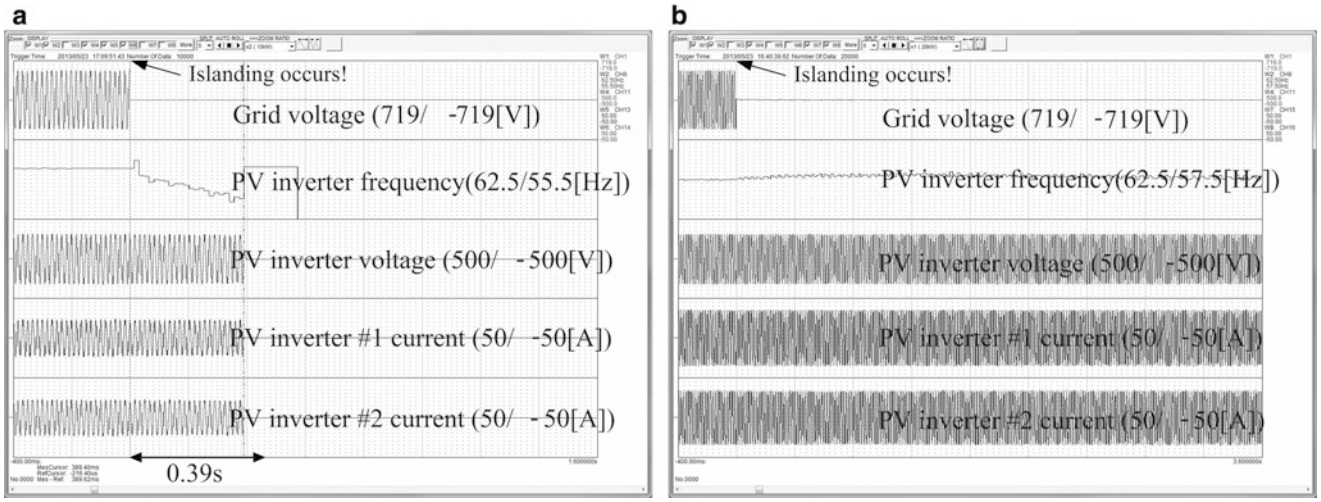


Fig. 4 Islanding test result for two PV inverters. (**a**) Under case 3, (**b**) Under case 4

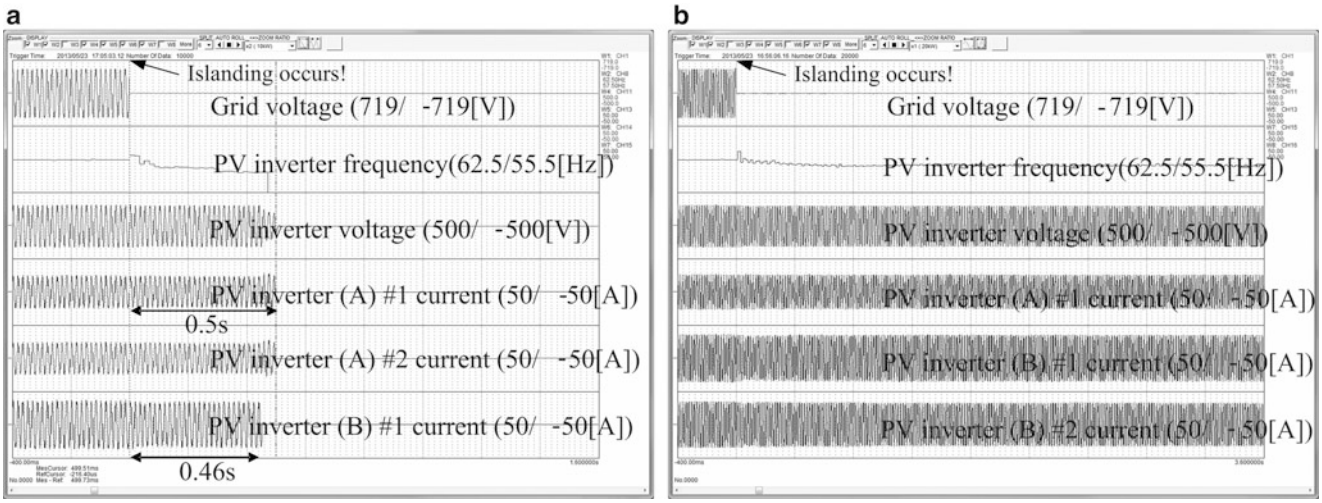


Fig. 5 Islanding test result for three PV inverters. (**a**) Under case 5, (**b**) Under case 6

Under case 3 as shown in Fig. 4 (a), two PV inverters using model A can prevent islanding within 0.39 s, which meet the requirement for islanding detection. However, two PV inverters using model B under case 4 can't detect islanding at all, as shown in Fig. 4 (b). Thus, it can be concluded that islanding can happen due to the cancellation problem between two PV inverters when these are operated even though single one can detect islanding in the laboratory. This is why the frequency variation in Fig. 4 (b) is remarkably reduced due the cancellation rather than in Fig. 3 (b).

Fig. 5 shows the islanding experimental results using three different PV inverters with different manufacturers. Under both case 5 and case 6, islanding detection capabilities have been degraded due to the interaction among three PV inverters. Especially, under case 6, the islanding can't be prevented at all even though model A PV inverter can detect islanding under case 3. As shown in Fig. 6, four PV inverters are evaluated under case 7. In the

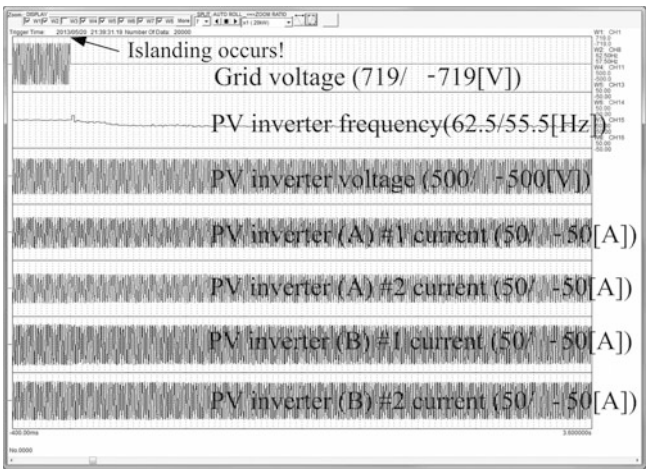


Fig. 6 Islanding test result for four PV inverters under case 7

same way, it is concluded that islanding can't be detected by interaction among the other PV inverters. Thus, it is important that the anti-islanding research and the related testing standard should be considered under multiple PV inverter operations.

4 Conclusion

Various anti-islanding methods for PV power generation have been proposed. However, these are mostly focus on the islanding prevention method for single PV system. Most of them are focusing on the anti-islanding performance of single PV system according to the related international and domestic standard test procedures. This paper presents performance analysis of anti-islanding functions for multiple grid-connected PV systems by using commercial products. From the experimental result, it can be concluded that islanding can happen due to the cancellation problem between two PV inverters when these are operated even though single one can detect islanding in the laboratory. . Thus, it is important that the anti-islanding research and the related testing standard should be considered under multiple PV inverter operations.

Acknowledgments This work was supported by the New & Renewable Energy of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea government Ministry of Knowledge Economy. (No. 20123010010060) and by the Human Resources Program In Energy Technology of the Korea Institute of Energy Technology Evaluation and Planning(KETEP) granted financial resource from the Ministry of Trade, Industry & Energy, Republic of Korea (NO. 20134010200540).

References

1. Du, P., Ye, Z., Aponte, E., Nelson, J., and Fan, L.: Positive-feedback-based active anti-islanding schemes for inverter-based distributed generators. IEEE Trans. Power Electron., vol. 25, no. 12, pp. 2941–2948. (2010)
2. Arnulf, J.: PV status report 2008. Ispra, Italy. Joint Research Center (2008)
3. Greenwood, A.: Electrical Transients in Power Systems, New York: Wiley, pp. 83 (1971)
4. IEEE Standard Conformance Test Procedures for Equipment Interconnecting Distributed Resources With Electric Power Systems, IEEE Std. 1547.1-2005 (2005)
5. IEEE Recommended Practice for Utility Interface of Photovoltaic Systems, IEEE Std. 929-2000 (2000)
6. Yu, B., Matsui, M., Yu, G.:A review of current anti-islanding methods for photovoltaic power system, Solar Energy, vol. 84, no. 5, pp. 745-754 (2010)
7. Zeineldin, H., Conti, S.: Sandia frequency shift parameter selection for multi-inverter systems to eliminate non-detection zone, IET Renewable Power Generation, vol. 5, no. 2, pp. 175–183(2011)
8. Test procedure of islanding prevention measures for utility-interconnected photovoltaic inverters, IEC Std. 62116 Ed. 1.0 (2008)

Presentation Of A Fuzzy Control Training And Test System

K.-D. Kramer, S. Braune, A. Söchting, T. Stolze, and C. Blankenberg

1 Introduction

To control processes with nonlinear characteristics, models are needed, which are not or difficult describable with classical methods. Computational Intelligence based model approaches (CI), realized by Fuzzy Control models (FC) or models with Artificial Neural Networks (NN), are used as alternative concepts to classical approaches.

Technical realizations, depending on special demands, reach form simple Look-Up-Tables and Microcontroller- or DSP-based CI Hard- und Software to PPS-based CI systems.

In university education or further training are on the one hand models necessary which represent the technical system transparent and easy cognizable and on the other hand a programming tool is required that supports an easy development process. That includes tools to verify the results and tuning the system with graphic functions under real time conditions. The methodical and didactical objective in the utilization of these teaching models is to develop solution strategies in CI applications, for example Fuzzy Controller, special to analyse different algorithms of inference or defuzzyfication and to verify and tune those systems effective.

K.-D. Kramer (✉)
Department of Automation and Computer Science, Harz University, Friedrichstrasse 57-59, 38855 Wernigerode, Germany

Institute of Automation and Informatics (IAI), Dornbergsweg 2, 38855 Wernigerode, Germany
e-mail: kkramer@hs-harz.de; k.kramer@iai-wr.de

S. Braune • A. Söchting
Institute of Automation and Informatics (IAI), Dornbergsweg 2, 38855 Wernigerode, Germany
e-mail: s.braune@iai-wr.de; a.soechting@iai-wr.de

T. Stolze • C. Blankenberg
Department of Automation and Computer Science, Harz University, Friedrichstrasse 57-59, 38855 Wernigerode, Germany
e-mail: tstolze@hs-harz.de; cblankenberg@hs-harz.de

Based on know-how in the field of developing such technical models and equivalent design tools in Microcontroller applications at the Harz University graded levels of microcontroller based teaching models in CI applications are developed. The Fuzzy design tool "FHFCE-Shell"®, developed at the Harz University, is optimized to target Microcontroller Z8ENCORE® of ZILOG Inc. [1], [3], [4, 5]. This 8 bit microcontroller possesses the essential hardware features (e.g. ADC, Timer/Counter, PWM, etc.) and, especially useful, a register-to-register architecture with 4096 accumulators, which is an excellent condition to use them to store the fuzzy variables [2].

2 Fuzzy Control

The FC processes differs by MAMDANI from the subprocesses *fuzzification, inference with rule base* and *defuzzification*. The conversion of crisp real input values to fuzzy values is executed in the fuzzification process. This process corresponds to a normalization on the uniformity interval [0...1]. All crisp inputs are assigned to accessory parts of the function terms (linguistic terms), called as membership degree [MD] or membership function [MF].

The result of the *fuzzification* is a set of variables with several, partial overlapping terms of values or linguistic variables. Some algorithms will be described to calculate the linguistic values quickly. Fast fuzzyfication algorithms in Low Cost applications use those trapezoidal shapes (e.g. modified equation of straight line) exclusively.

According to the rules of the rule-base the calculation of the normalized inputs by different operators is performed in the next step. With these operators some treatments will be carried out with the inputs in these three steps: aggregation, implication and accumulation.

The third sub- process, called *defuzzification*, calculates the crisp overall results of the fuzzy algorithm as the weighted sum of all rules consequents by their premise degree of activation. Different mathematical methods are

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_23,

well-known. The mathematical expense is invert to the precision of the results in these methods, that's why in Low-Cost-Microcontrollers only non expensive methods are used. [6, 7]

Therefore methods are developed, which are on the one hand easy to realize, but imprecise and on the other hand an algorithm (e.g. Center of Area (CAO)), which uses a special formula of Newton-Cotes realized by KRAMER/SCHÜTT [3], to give exact results with a limited computational expense.

3 FC Design Process

3.1 Goals of FC Development Process

The FHFCE-Tool has been developed to realize the following approaches:

- target-optimized generation of fuzzy- and NON-fuzzy algorithms for the microcontroller Z8E
- system design by using a uniform graphical platform
- verification of the results (fuzzy debugging, fuzzy monitoring, fuzzy tracing)

3.2 Project Editing of FC's

The structure of the design process of a fuzzy system is realized analogously to the fuzzy control process shown in Fig. 1. All the definitions of the structure and the algorithms of the fuzzy process must be determined, including all data conversions and data manipulations.

All icons include software macros, written in assembler, to generate an optimized code and down- load them on the target controller. The structure of the FC is defined by the red lines (Fig. 1). In the code generation process, the machine code of the MC will be generated depending on this structure.

3.3 Fuzzy Control Debug

In addition to the design process a verification and tuning phase is necessary. This verification and tuning phase can be realized directly (On-line or Off-line) or indirectly (Off-line or with a simulator). The modification of parameters or rules has to be carried out in this phase. Changing functions of the system is an initial requirement.

There are different goals:

- improving the dynamic of the system
- improving the exactness of the system
- optimizing the system in general

All these processes, described as tuning and optimizing processes, require efficient system support:

- On-line debugging (realized by the system debugger)
- On-line debugging with management of the INPUT/OUTPUT data (generally realized by a real-time system)
- Complex analysis of performance (measuring methods, monitoring, tracing, Fuzzy benchmarks)

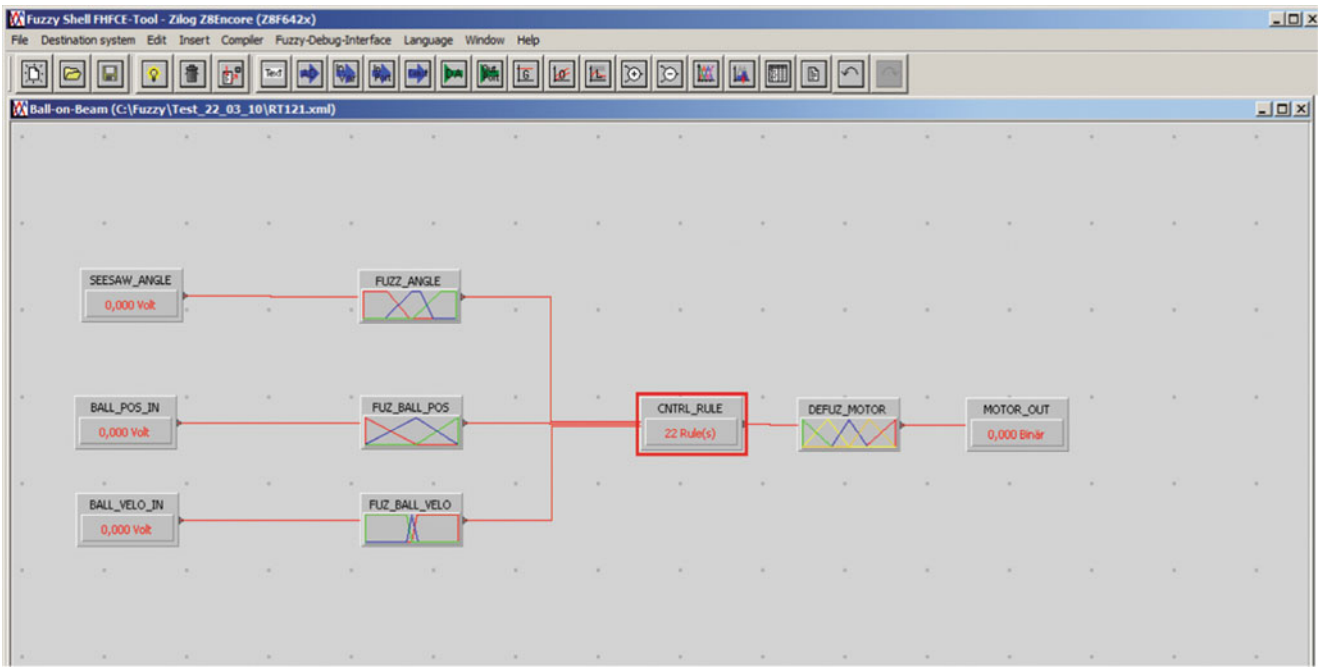


Fig. 1 Graphical Editing Window of FHFCE tool

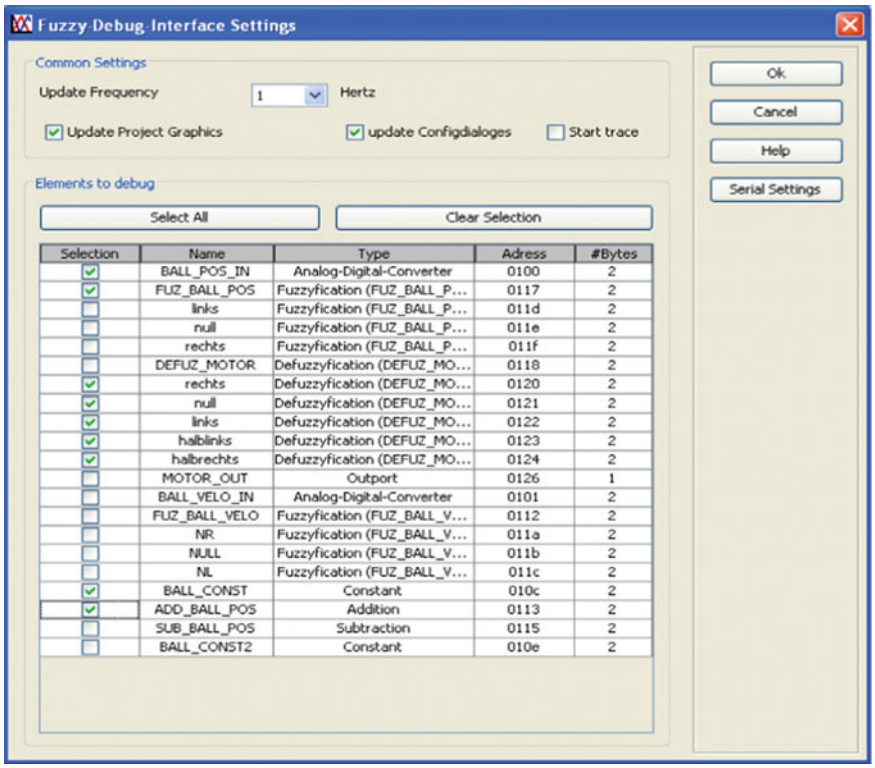


Fig. 2 Fuzzy debugging – selection of relevant values

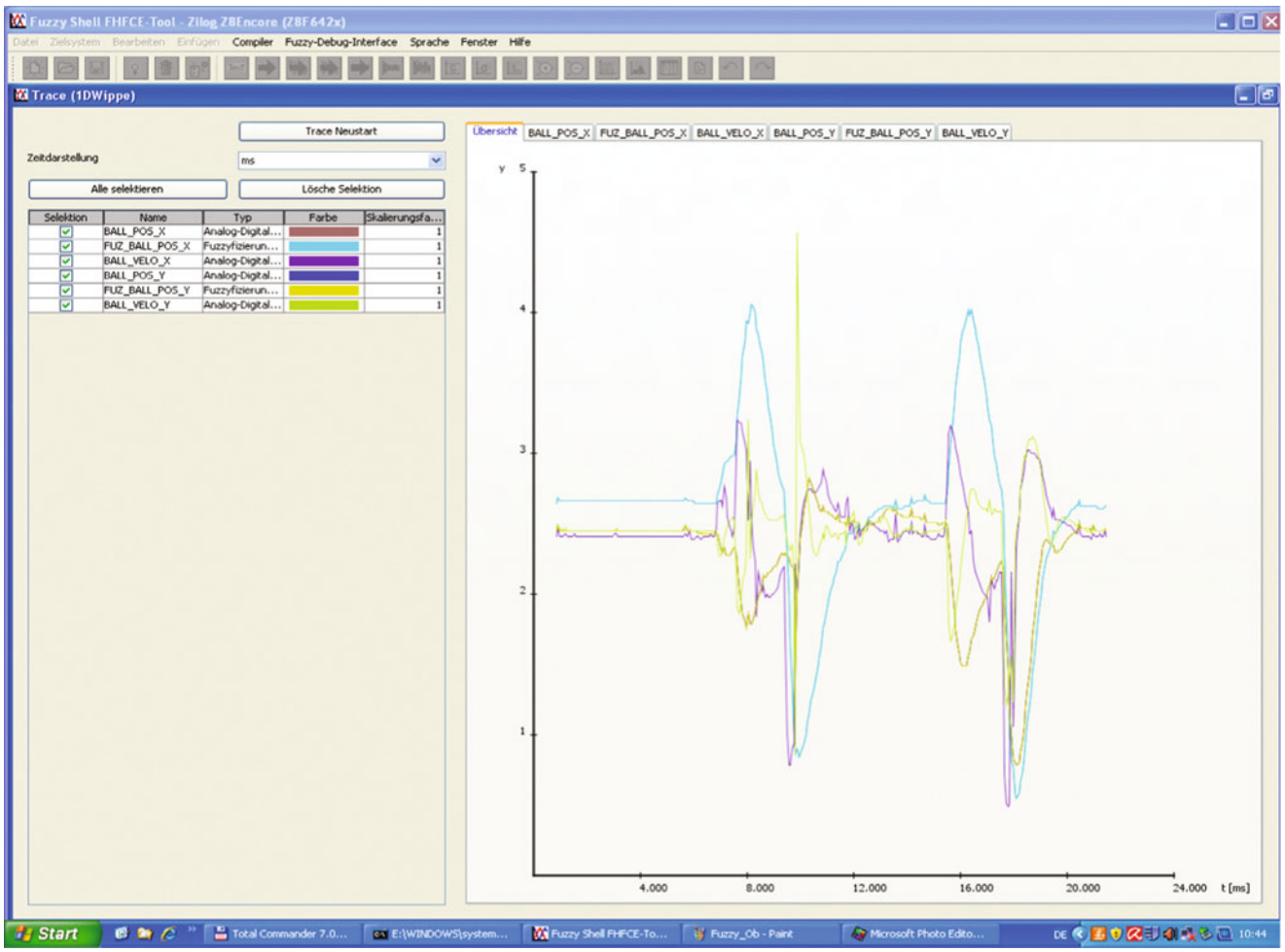


Fig. 3 Fuzzy Control tracing

4 FC Teaching Models

4.1 Overview

At the Harz University 4 teaching models with different levels were developed:

- 1-dimensional seesaw (ball seesaw model)
- inverse pendulum (propeller model)
- 2-dimensional seesaw (ball plane model)
- stick balance car model

4.2 Aspects of using the FC Teaching models

At all 4 models the technical problem is easy cognizable. There are 2 or 3 process inputs and 1 or 2 process outputs. So it is easy to define the rule base and to test the system. Differences between the models exist at the real time demands, the structure of the Fuzzy Controllers and the different FC processes (different inference processes and different defuzzyfication functions).

4.3 Examples

The first model is a 1-dimensional seesaw (ball seesaw model) shown in Fig. 4. Based on three process inputs (position of the ball, speed of the ball and position of the seesaw), the ball must be balanced on the middle of the seesaw. The special topic of this experiment is to learn something about essential FC –functions, the handling of FHFCE-Shell and about the online debugging functions.

A classic case of using Fuzzy Control is the inverse pendulum (Fig. 5). At this model the position of the pendulum will be manipulated with two opposite rotating air twisters. The process inputs are:

Fig. 4 1-dimensional seesaw model

position of the pendulum, speed of the pendulum and, if necessary, the position of a disturbance load.

Two process outputs (air twister I and II) will be controlled, to bring the pendulum in a vertical position. The special topic at this model is the realization of two independent Fuzzy Controllers where each of it has its own independent rule base and special demands on tuning and verification.

The third model is a 2-dimensional seesaw (ball plane model) (Fig. 6). The topic at this model is to realize two total independent Fuzzy Controllers with each separated inputs (ball position and ball speed in x and y direction and the table position) and outputs (motor control in x and y direction).

The main problem at this model is to balance the ball in the middle of the table with the two independent cooperating Fuzzy Controllers. There is a possibility that the task is to modify the target place of the ball or to learn a complex motion in teach-in-mode.

The fourth model is the stick balance car model (Fig. 7), a special version of the inverse pendulum. With two or three process inputs (position of the stick, angle of the stick, position of the car), the pendulum must be held in the vertical position. The extremely high real time conditions require a special system design and the definition of special tuning strategies.

Fig. 5 inverse pendulum (propeller model)

Fig. 6 2-dimensional seesaw

Fig. 7 stick balance car model

5 Conclusions

The teaching models represent different examples of using FC applications. The research of varying FC algorithms with reference to exactness, speed and ability to run in real time processes with different algorithms is an important aim in teaching processes. In addition to this, the FHFCE-tool enables the users to edit algorithms or FC functions. So users can develop and test new approaches at real technical models.

The models are offered by the GUNT-Company in Hamburg (Germany) (www.gunt.de).

In further developments the system will extended to Neuro-Fuzzy methods to optimize the FC (optimization of fuzzyfication and/or of rule base) and test it by the FHFCE-Debug system.

A second step is to use Neural Networks with the models. So it is possible to learn special functions in teach-in-modes, e.g. with the ball plane model, and control this with a Neural Network programmed in a microcontroller.

References

1. Becker, C: Entwicklung einer Programmieroberfläche für PMS500IF, Diploma Thesis: Harz University (1996)
2. Blankenberg, C.: Entwicklung einer Fuzzy-Shell für den Mikrocontroller Z8ENCORE, Diploma Thesis: Harz University (2004)
3. Kramer, K.-D., Braune, St.: Fuzzy Design Tool for Low-Cost Microcontrollers, ISIC-2001, Singapore, NTU, Proceedings p.473–475 (2001)
4. nn: RT121 – RT124 Teaching Systems for Fuzzy methods in Automation, IMPRINT, GUNT Gerätebau GmbH, Hamburg (2007)
5. Söchting, A, Stolze, T., Kramer, K-D., Braune, S.: Trace function at the FHFCE-Tool, Harz University, internal Paper (2009)
6. Lee, C.C.: Fuzzy Logic in Control Systems: Fuzzy Logic Controller – Part I, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 20, No. 2 (1990)
7. Driankov, D., Hellendoorn, H., Reinfrank, M.: An Introduction to Fuzzy Control, second edition, Springer-Verlag, Berlin (1996)

Web Service Intrusion Detection Using a Probabilistic Framework

Hassen Sallay, Sami Bourouis, and Nizar Bouguila

1 Introduction

Recent advances in web technologies and the constant growth of the Internet have led to many online service applications. Examples include e-commerce, social networks, online banking, business intelligence, web search engines, etc. An important feature of these web services is that they are based on software applications running at the server-side and generating new web content in an online fashion, which makes them flexible to exchange information on the Internet [32, 10, 29]. The flexibility of web services poses also vulnerabilities which make them the targets for attacks (e.g. code injection attacks, SQL/XML injection, buffer overflow attacks, denial of service, etc.) by cyber-criminals who can collect confidential information from servers or even compromise them [34, 9, 40, 14, 17]. Then, there is an urgent need to protect the servers on which the applications are running [45, 44, 18, 8]. Indeed, intrusion detection systems (IDSs) need to be deployed. An overview of current intrusion detection techniques and related issues was proposed in [42, 38]. Recently, data mining and machine learning approaches have been used in this growing area in order to improve the performance of existing systems [31, 22, 41, 12, 16, 20]. The key idea for these works is to use machine learning techniques (e.g. decision trees, artificial neural networks, support vector machines, mixture models, etc.) to train a classifier and to recognize attacks based on a list of features, which generally reduces the intrusion detection problem to an adversarial learning task [24].

H. Sallay (✉)
Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia
e-mail: hmsallay@imamu.edu.sa

S. Bourouis
Taif University, Taif, Saudi Arabia
e-mail: s.bourouis@tu.edu.sa

N. Bouguila
Concordia University, Montreal, QC, Canada
e-mail: nizar.bouguila@concordia.ca

Based on the analysis methods, IDSs are usually classified into two main categories: misuse (i.e. signature-based) detection and anomaly detection systems [35]. In misuse detection systems, the goal is to detect the occurrence of attacks that have been previously identified as intrusions. For this type of IDS, attacks must be known a priori. Misuse detection can be viewed then as a supervised learning problem. Alternatively, anomaly detection systems detect unknown attacks by observing deviations from normal activities of the system. It is based on the assumption that intrusive activities are noticeably different from normal system activities and hence detectable. Data clustering and unsupervised learning approaches have been widely used to develop anomaly detection systems. Several of recent clustering approaches quantify deviation from normal behavior using thresholds (see, for instance, [33, 45, 44]). Unlike these approaches we consider a robust finite Gaussian mixtures to model normal traffic and then to automatically detect potential intrusions (i.e. anomalous traffic). Our main idea is based on incorporating into the Gaussian mixture an auxiliary outlier component, to which we associate a uniform density, to represent abnormal requests. The resulting model is learned using an expectation-maximization algorithm.

The rest of this paper is organized as follows: the proposed web service intrusion detection model is described in Section 2. Then, obtained results using a data set containing both normal and intrusive requests which were collected from a large real-life web service. are given and analyzed in Section 3. Finally, Section 4 concludes the paper.

2 A finite mixture model with outliers detection

In this section, we present our mixture model for web service intrusion detection and the motivations behind it. Moreover, we propose a detailed approach to learn the parameters of the proposed model.

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_24,

2.1 The model

The mixture model Let us consider a training data set of N vectors $\chi = X_1, \ldots, X_N$, where each $X_i = (X_{i1}, \ldots, X_{iD})$ is a D-dimensional vector of features representing a given request on a web server. Set of vectors generally contains examples that belong to many clusters (i.e. categories of requests) and can be modeled by a finite mixture of distributions [28]

$$p(X_i|\Theta_K) = \sum_{k=1}^{K} p_k p(X_i|\theta_k) \tag{1}$$

where $p_k > 0$, are the mixing proportions, M is the number of mixture components, and $\Theta_K = P = (p_1, \ldots, p_K), \theta = (\theta_1, \ldots, \theta_K)$ is the set of parameters in the mixture model. A critical problem in this case is the choice of the probability density function to represent each component. In this paper, we consider a classic choice namely a Gaussian distribution with parameters μ_k and Σ_k:

$$p(X_i|\theta_k) = \frac{1}{(2\pi)^{D/2}\sqrt{|\Sigma|_k}} \exp\left(-\frac{1}{2}(X_i - \mu_k)^T \Sigma_k^{-1}(X_i - \mu_k)\right) \tag{2}$$

Outliers detection Legitimate requests are generally in higher number than malicious ones. Thus, it is possible to formalize abnormal requests as outliers when considering statistical models. Many outliers detection approaches have been proposed in the past [15, 1, 43]. Indeed, the problem is a fundamental classic task in data mining and is generally related to fraud detection problems in the security domain (e.g. credit-card fraud, fraudulent cellular call detection, etc.) [39]. Here we approach the problem by incorporating an auxiliary outlier component, to which we associate a uniform density [37, 19, 4, 25], into the mixture model:

$$p(X_i|\Theta_K) = \sum_{k=1}^{K} p_k p(X_i|\theta_k) + p_{K+1} U(X_i) \tag{3}$$

where $p_{K+1} = 1 - \sum_{k=1}^{K} p_k$ is the probability that X_i was not generated by the central mixture model and $U(X_i)$ is a uniform distribution common for all data to model isolated vectors which are not in any of the K clusters and which show significantly less differentiation among clusters. Assuming uniform distribution for outliers is a common assumption that has been previously used successfully in [37, 43, 26]. It is noteworthy that when $p_{M+1} = 0$ the outlier component is removed and the previous equation is reduced to Eq. 1.

2.2 Model Learning

The most widely used approach for unknown parameters estimation is maximum likelihood (ML) based on maximizing the log-likelihood function as following

$$\widehat{\Theta} = \arg\max_{\Theta} \left\{ \log p(\chi|\Theta) = \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} p_k p(X_i|\theta_k) + p_{K+1} U(X_i) \right] \right\}$$

ML estimation is generally performed with expectation maximization (EM) algorithm [27] which is well-known in the case of finite mixture models. The main modification, in our case, is related to the E-step in which the posterior probabilities are calculated as following in iteration q:

$$\begin{cases} \gamma_{ik}^{(q)} := P\left(c_k/X_i, \theta_k^{(q)}\right) := \dfrac{p_k^{(q-1)} p\left(X_i|\theta_k^{(q-1)}\right)}{\sum_{k=1}^{K} p_k^{(q-1)} p\left(X_i|\theta_k^{(q-1)}\right) + p_{K+1}^{(q-1)} U(X_i)} & \text{for} \quad k = 1, \ldots, K \\ \gamma_{i(K+1)}^{(q)} := \dfrac{p_{K+1}^{(q-1)} U(X_i)}{\sum_{k=1}^{K} p_k^{(q-1)} p\left(X_i|\theta_k^{(q-1)}\right) + p_{K+1}^{(q-1)} U(X_i)} \end{cases} \tag{4}$$

where $\gamma_{i(K+1)}$ is the probability of affecting the vector X_i to the set of outliers (or abnormal requests). Thus, the complete estimation algorithm is summarized as follows:

1. Initialization-step: initialization using the K-means algorithm

For each iteration q:

2. Expectation-step: Calculate $\gamma_{ik}{}^{(q)}$, $i = 1, \ldots, N$, $k = 1, \ldots, K+1$ using Eq. 4.
3. Maximization-step:

$$\begin{cases} p_k^q := \dfrac{\sum_{i=1}^{N} \gamma_{ik}^{(q)}}{N} \\ \mu_k^q := \dfrac{\sum_{i=1}^{N} \gamma_{ik}^{(q)} p\left(X_i|\theta_k^{(q-1)}\right)}{\sum_{i=1}^{N} \gamma_{ik}^{(q)}} \\ \Sigma_k^q := \dfrac{\sum_{i=1}^{N} \gamma_{ik}^{(q)} \left(X_i - \mu_k^{(q)}\right)\left(X_i - \mu_k^{(q)}\right)^t}{\sum_{i=1}^{N} \gamma_{ik}^{(q)}} \end{cases} \quad (5)$$

The iterations in the previous algorithm are repeated until we reach convergence. Examples of convergence tests include the stabilization of the likelihood function, the parameters, or the posterior probabilities. Concerning the determination of K, the number of mixture components, which can be viewed as the number of request categories, different selection criteria have been proposed in the past and could be considered [28, 6]. In this work, we use the mixture minimum description length (MMDL) criterion developed in [13]:

$$MMDL(K) = -\log(p(\chi|\Theta)) + \frac{1}{2} N_p \log(N) + \frac{c}{2} \sum_{k=1}^{K} \log(p_k) \quad (6)$$

where N_p is the total number of free parameters in the model and c is the number of parameters describing each component. We select the K that yields the minimum value of $MMDL(M)$. Moreover, according to our operational definition of outliers, they should have a uniform distribution, since they do not follow the pattern of the majority of the data. A common approach, to define this uniform distribution, is to suppose that the data follow a single component model averaged over all the observation [37]. Thus, in our case, we choose the following:

$$U(X) = \frac{1}{N} \sum_{i=1}^{N} p\left(X_i|\widehat{\theta}\right) \quad (7)$$

where the parameters $\widehat{\theta}$ is estimated using ML technique. This formulation takes into account the fact that outliers should be sparsely distributed.

3 Experimental Results

The proposed framework is tested using logs collected from a real-life web service (from several Apache servers) in a two weeks time interval. The collected data set contains normal requests, anomalies as well as intrusions. More specifically, our training data is collected at the beginning and is composed of 10000 requests. The majority of these requests are legitimate, but some are attacks (e.g. cross-site scripting, SQL injections, buffer overflows, etc.). After using these data to train our mixture model, by considering 3-gram representation as done in [44], new requests are considered and classified as normal or abnormal (i.e. outlier) using the technique proposed in this paper. These new requests constitute the testing set and their number is equal to 35000. It is noteworthy that these data are used also to update the model using the approach proposed in [36]. Other incremental versions of the EM algorithm such as those in [30, 23] could be used, also. Updating the model's parameters allows to take into account new request categories and new intrusion pattern which didn't appear in initial training data but may emerge in future data. It is noteworthy that this is crucial in practice in order to plan and design new countermeasures.The evaluation of the performance of our approach has been based on the following measures:

- True positive rate which represents the number of correctly detected intrusions over the number of intrusions in the testing set.
- False positive rate which represents the number of normal requests considered as intrusions over the total number of normal requests in the testing set.
- True negative rate which represents the number of correctly classified normal requests over the total number of normal requests in the testing set.
- False negative rate which represents the number of misclassified intrusions over the number of intrusions in the testing set.
- Accuracy which represents the number of correctly classified requests over the total number of requests in the testing set.
- Precision which represents the number of correctly classified intrusions over the number of intrusions

The performance results of our approach are presented in Table 1. According to this table, it is clear that our approach provides excellent detection results and that the different N-gram approaches perform comparably. We compared our algorithm with the SDEM and SDPU approaches in [39] based on Gaussian mixture models and kernel mixtures

Table 1 Performance of the method in detecting web service intrusion when considering 1-gram, 2-gram and 3-gram models to describe features.

	1-gram	2-gram	3-gram
True positive rate	98.02	98.03	98.12
False positive rate	0.97	0.98	0.98
True negative rate	98.60	98.63	98.77
False negative rate	1.04	1.01	1.01
Accuracy	97.89	97.92	97.95
Precision	98.04	98.13	98.21

Table 2 Performance of the SDEM method in detecting web service intrusion.

	1-gram	2-gram	3-gram
True positive rate	97.90	97.97	98.09
False positive rate	1.00	1.01	1.02
True negative rate	98.56	98.60	98.65
False negative rate	1.02	1.02	1.05
Accuracy	97.87	97.92	97.92
Precision	98.02	98.11	98.17

Table 3 Performance of the SDPU method in detecting web service intrusion.

	1-gram	2-gram	3-gram
True positive rate	97.91	97.98	98.08
False positive rate	1.01	1.02	1.03
True negative rate	98.50	98.65	98.66
False negative rate	1.03	1.03	1.05
Accuracy	97.85	97.90	97.91
Precision	98.08	98.13	98.18

Table 4 Performance of K-nearest neighbor method in detecting web service intrusion.

	1-gram	2-gram	3-gram
True positive rate	95.02	95.09	95.15
False positive rate	1.33	2.22	2.13
True negative rate	94.14	94.35	94.41
False negative rate	2.51	2.43	2.39
Accuracy	94.38	94.56	94.77
Precision	94.66	94.73	94.85

Table 5 Performance of the growing hierarchical self organizing maps (GHSOMs) [45] method in detecting web service intrusion.

	1-gram	2-gram	3-gram
True positive rate	98.01	98.01	98.08
False positive rate	1.02	1.02	1.01
True negative rate	98.60	98.60	98.65
False negative rate	1.07	1.04	1.04
Accuracy	97.70	97.76	97.88
Precision	98.08	98.09	98.17

as shown in Tables 2 and 3, respectively. Moreover, we performed comparisons with the well-known nearest-neighbor technique as shown in Table 4 and three recent state of the art approaches, namely GHSOMs [45], diffusion maps [21], and the algorithm proposed in [44] as shown in Tables 4, 5, and 6, respectively. The results shown in all the tables demonstrate that our statistical framework is promising.

Table 6 Performance of diffusion maps method [21] in detecting web service intrusion.

	1-gram	2-gram	3-gram
True positive rate	98.00	98.00	98.03
False positive rate	1.07	1.06	1.06
True negative rate	98.55	98.56	98.63
False negative rate	1.24	1.24	1.15
Accuracy	97.74	97.74	97.77
Precision	98.03	98.08	98.09

Table 7 Performance of the algorithm proposed in [44] in detecting web service intrusion.

	1-gram	2-gram	3-gram
True positive rate	98.04	98.04	98.05
False positive rate	1.03	1.03	1.02
True negative rate	98.65	98.68	98.68
False negative rate	1.06	1.10	1.10
Accuracy	97.79	97.79	97.81
Precision	98.15	98.18	98.19

4 Conclusion

Machine learning techniques have been widely used recently for computer security purposes [7]. Following this interesting trend, we introduce in this paper a new method to detect intrusion attacks on web services using a finite mixture model. The proposed finite Gaussian mixture model is augmented with an auxiliary uniform component to detect suspicious requests which are viewed as outliers. The proposed statistical framework is learned using an EM algorithm. The proposed technique is theoretically reliable and robust and has been validated using real-world data extracted from real web services. There are many avenues for future research. For instance, it is possible to extend the proposed mixture model to the infinite case using Dirichlet processes which allow to model outliers implicitly. Another promising future work is to integrate feature selection within the proposed framework or to consider other mixture models [5, 3, 2, 11]. Other applications such as spam filtering and credit card fraud detection are also possible.

Acknowledgments. The first author would like to thank King Abdulaziz City for Science and Technology (KACST), Kingdom of Saudi Arabia, for their funding support under grant number 11-INF1787-08.

References

1. Barnett, V., Lewis, T.: Outliers in Statistical Data. John Wiley & Sons (1994)
2. Bouguila, N., Ziou, D.: Dirichlet-based probability model applied to human skin detection. In: Proc. of the IEEE International

Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 521–524 (2004)
3. Bouguila, N., Ziou, D.: A powerful finite mixture model based on the generalized Dirichlet distribution: Unsupervised learning and applications. In: Proc. of the 17th International Conference on Pattern Recognition (ICPR). pp. 280–283 (2004)
4. Bouguila, N., Almakadmeh, K., Boutemedjet, S.: A finite mixture model for simultaneous high-dimensional clustering, localized feature selection and outlier rejection. Expert Systems with Applications 39(7), 6641–6656 (2012)
5. Bouguila, N., Ziou, D.: Using unsupervised learning of a finite dirichlet mixture model to improve pattern recognition applications. Pattern Recognition Letters 26(12), 1916–1925 (2005)
6. Bouguila, N., Ziou, D.: Unsupervised selection of a finite dirichlet mixture model: An mml-based approach. IEEE Transactions on Knowledge and Data Engineering 18(8), 993–1009 (2006)
7. Chan, P.K., Lippmann, R.: Machine learning for computer security. Journal of Machine Learning Research 6, 2669–2672 (2006)
8. Corona, I., Giacinto, G.: Detection of server-side web attacks. In: Diethe, T., Cristianini, N., Shawe-Taylor, J. (eds.) WAPA. JMLR Proceedings, vol. 11, pp. 160–166. JMLR.org (2010)
9. Dagdee, N., Thakar, U.: Intrusion attack pattern analysis and signature extraction for web services using honeypots. In: Proc. of the First International Conference on Emerging Trends in Engineering and Technology (ICETET). pp. 1232–1237 (2008)
10. Desmet, L., Jacobs, B., Piessens, F., Joosen, W.: Threat modelling for web services based web applications. In: Chadwick, D., Preneel, B. (eds.) Communications and Multimedia Security, IFIP The International Federation for Information Processing, vol. 175, pp. 131–144. Springer US (2005)
11. Elguebaly, T., Bouguila, N.: Bayesian learning of finite generalized gaussian mixture models on images. Signal Processing 91(4), 801–820 (2011)
12. Fan, W., Bouguila, N., Ziou, D.: Unsupervised anomaly intrusion detection via localized bayesian feature selection. In: Proc. of the EEE International Conference on Data Mining (ICDM). pp. 1032–1037 (2011)
13. Figueiredo, M.A.T., Leitão, J.M.N., Jain, A.K.: On fitting mixture models. In: Hancock, E.R., Pelillo, M. (eds.) EMMCVPR. Lecture Notes in Computer Science, vol. 1654, pp. 54–69. Springer (1999)
14. Gruschka, N., Luttenberger, N.: Protecting web services from dos attacks by soap message validation. In: Fischer-Hebner, S., Rannenberg, K., Yngstrm, L., Lindskog, S. (eds.) Security and Privacy in Dynamic Environments, IFIP International Federation for Information Processing, vol. 201, pp. 171–182. Springer US (2006)
15. Hawkins, D.M.: Identification of Outliers. Chapman and Hall, London (1980)
16. Horng, S.J., Su, M.Y., Chen, Y.H., Kao, T.W., Chen, R.J., Lai, J.L., Perkasa, C.D.: A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert Systems with Applications 38(1), 306 – 313 (2011)
17. Jensen, M., Gruschka, N., Herkenhoner, R., Luttenberger, N.: Soa and web services: New technologies, new standards - new attacks. In: Proc. of the Fifth European Conference on Web Services (ECOWS). pp. 35–44 (2007)
18. Jensen, M., Gruschka, N., Herkenhener, R.: A survey of attacks on web services. Computer Science - Research and Development 24 (4), 185–197 (2009)
19. Ke, Q., Kanade, T.: Robust subspace clustering by combined use of *k*ndd and svd algorithm. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 592–599 (2004)
20. Khan, L., Awad, M., Thuraisingham, B.: A new intrusion detection system using support vector machines and hierarchical clustering. The VLDB Journal 16(4), 507–521 (2007)
21. Kirchner, M.: A framework for detecting anomalies in http traffic using instance-based learning and k-nearest neighbor classification. In: Proc. of the 2nd International Workshop on Security and Communication Networks (IWSCN). pp. 1–8 (May 2010)
22. Laskov, P., Dessel, P., Schefer, C., Rieck, K.: Learning intrusion detection: Supervised or unsupervised? In: Roli, F., Vitulano, S. (eds.) Image Analysis and Processing (ICIAP), Lecture Notes in Computer Science, vol. 3617, pp. 50–57. Springer Berlin Heidelberg (2005)
23. Liang, P., Klein, D.: Online em for unsupervised models. In: Proc. of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 611–619. NAACL '09, Association for Computational Linguistics (2009)
24. Lowd, D., Meek, C.: Adversarial learning. In: Proc. of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). pp. 641–647 (2005)
25. Mashrgy, M.A., Bdiri, T., Bouguila, N.: Robust simultaneous positive data clustering and unsupervised feature selection using generalized inverted dirichlet mixture models. Knowledge-Based Systems 59, 182–195 (2014)
26. Mashrgy, M.A., Bouguila, N., Daoudi, K.: A robust approach for multivariate binary vectors clustering and feature selection. In: Lu, B.L., Zhang, L., Kwok, J.T. (eds.) ICONIP (2). Lecture Notes in Computer Science, vol. 7063, pp. 125–132. Springer (2011)
27. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. New York: Wiley (1997)
28. McLachlan, G., Peel, D.: Finite Mixture Models. New York: Wiley (2000)
29. Mehdi, M., Bouguila, N., Bentahar, J.: Trustworthy web service selection using probabilistic models. In: Proc. of the IEEE 19th International Conference on Web Services (ICWS). pp. 17–24 (2012)
30. Neal, R.M., Hinton, G.E.: A new view of the em algorithm that justifies incremental and other variants. In: Learning in Graphical Models. pp. 355–368. Kluwer Academic Publishers (1993)
31. Patcha, A., Park, J.M.: An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks 51(12), 3448 – 3470 (2007)
32. Pearce, C., Bertok, P., Schyndel, R.: Protecting consumer data in composite web services. In: Sasaki, R., Qing, S., Okamoto, E., Yoshiura, H. (eds.) Security and Privacy in the Age of Ubiquitous Computing, IFIP Advances in Information and Communication Technology, vol. 181, pp. 19–34. Springer US (2005)
33. Pereira, H., Jamhour, E.: A clustering-based method for intrusion detection in web servers. In: Proc. of the 20th International Conference on Telecommunications (ICT). pp. 1–5 (2013)
34. Pinzen, C., Paz, J.F., Zato, C., Perez, J.: Protecting web services against dos attacks: A case-based reasoning approach. In: Romay, M., Corchado, E., Garcia Sebastian, M. (eds.) Hybrid Artificial Intelligence Systems, Lecture Notes in Computer Science, vol. 6076, pp. 229–236. Springer Berlin Heidelberg (2010)
35. S. Northcutt and J. Novak: Network Intrusion Detection: An Analyst's Handbook. New Riders Publishing (2002)
36. Samé, A., Ambroise, C., Govaert, G.: An online classification em algorithm based on the mixture model. Statistics and Computing 17 (3), 209–218 (2007)
37. Titsias, M.K., Williams, C.K.I.: Sequentially fitting mixture models using an outlier component. In: Proc. of the 6th International Workshop on Advances in Scattering and Biomedical Engineering. pp. 386–393 (2003)
38. Tsai, C.F., Hsu, Y.F., Lin, C.Y., Lin, W.Y.: Review: Intrusion detection by machine learning: A review. Expert Systems with Applications 36(10), 11994–12000 (2009)
39. Yamanishi, K., ichi Takeuchi, J., Williams, G.J., Milne, P.: On-line unsupervised outlier detection using finite mixtures with

discounting learning algorithms. Data Mining and Knowledge Discovery 8(3), 275–300 (2004)
40. Yee, C.G., Shin, W.H., Rao, G.S.V.R.K.: An adaptive intrusion detection and prevention (ID/IP) framework for web services. In: Proc. of the International Conference on Convergence Information Technology (ICCIT). pp. 528–534 (2007)
41. Zanero, S., Savaresi, S.M.: Unsupervised learning techniques for an intrusion detection system. In: Proc. of the ACM Symposium on Applied Computing (SAC). pp. 412–419. ACM (2004)
42. Zhou, C.V., Leckie, C., Karunasekera, S.: A survey of coordinated attacks and collaborative intrusion detection. Computers & Security 29(1), 124 – 140 (2010)
43. Zivkovic, Z., Krose, B.: An em-like algorithm for color-histogram-based object tracking. In: Proc. of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). pp. I–798–I–803 Vol.1 (2004)
44. Zolotukhin, M., Hamalainen, T.: Detection of anomalous http requests based on advanced n-gram model and clustering techniques. In: Balandin, S., Andreev, S., Koucheryavy, Y. (eds.) Internet of Things, Smart Spaces, and Next Generation Networking, Lecture Notes in Computer Science, vol. 8121, pp. 371–382. Springer Berlin Heidelberg (2013)
45. Zolotukhin, M., Hamalainen, T., Juvonen, A.: Growing hierarchical self-organizing maps and statistical distribution models for online detection of web attacks. In: Cordeiro, J., Krempels, K.H. (eds.) Web Information Systems and Technologies, Lecture Notes in Business Information Processing, vol. 140, pp. 281–295. Springer Berlin Heidelberg (2013)

Multi-Agent Reinforcement Learning Control for Ramp Metering

Ahmed Fares and Walid Gomaa

1 Introduction

The huge increase in the number and length of traffic jams on freeways has led to the use of several dynamic traffic management measures. The sophistication of traffic network demands, as well as their severity, have also increased recently. Consequently the need for an optimal and reliable traffic control for the freeway networks, has become more vital.

Controlling the number of vehicles entering the freeway from the ramp is called *the rate of ramp metering*, which could be measured by a ramp metering device. The aim of this research is to control the amount of vehicles entering the mainstream freeway from the ramp merging area, i.e., balance the demand and the capacity of the freeway. This keeps the freeway density below the critical one. Consequently, this leads to maximum utilization of the freeway without entering in congestion while maintaining the optimal freeway operation. In order to achieve this an *optimal coordination* of the freeway traffic control *measures* over the network level is highly needed.

Several different techniques have been recently applied including: Intelligent control theory, fuzzy control [1, 2], neural network control [3, 4], and even some hybrid of them [5]. Although there have been some progress with such line of research, these techniques need some expertise to extract the parameters and rules and huge amount of training data. Although, some researches move to use the reinforcement learning (RL) [6–8], they only solve the problem locally. Figuring out a congestion only locally, may lead to that the vehicles run faster into another congestion, whereas still the same amount of vehicles have to pass the bottleneck. As result, the congestion only move from one road segment to another, while the average travel time of the freeway remains the same. So we need to solve this problem in the *network level* by *coordinated freeway traffic control*.

In this paper, we propose a new framework of ramp metering in the network level based on modeling by *collaborative Markov Decision Process* [9, 10] and an associated *cooperative Q-learning* algorithm, which is based on *a payoff propagation* algorithm [11] under the *coordination graph* framework. The proposed system not only avoid the locality of all techniques mentioned above, but also the computational complexity and the risk of being trapped in local optimum of the model predictive control (MPC) [12], In addition, the solid knowledge of the system considered to extract the rules as in *the fuzzy control system*. Our framework was extensively tested in order to assess the proposed model of the joint payoff, as well as the global payoff.

2 Q-Learning

As long as we don't have a *model* of the environment (unknown environment). So we no longer consider the infinite sum of the discount reward as a function of state s only, but as a function of the state action pair (s, a). That is why we used such a Q function [13].

When the agent is in state s and the agent is performing action a and his *long-term reward* is $Q(s, a)$. If each state and action pairs are visited infinitely often; then $Q(s, a)$ converges to $Q^*(s, a)$ [14]. Based on such long-term reward the optimal policy is:

$$\pi^*(s) = \arg\max_{a \in A} Q^*(s, a) \tag{1}$$

Q-learning converges no matter how we choose our actions during the learning as long as every action is selected

A. Fares (✉)
Computer Science and Engineering Department, Egypt-Japan University for Science and Technology (E-JUST), Alexandria, Egypt
e-mail: ahmed.fares@ejust.edu.eg

W. Gomaa
Currently on leave from the faculty of Engineering, Alexandria University, Alexandria, Egypt
e-mail: walid.gomaa@ejust.edu.eg

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_25,

infinitely often (*fair action selection strategy*), but of course the speed of the convergence depends on the action that we used during the learning (*action selection strategy*).

A modification of the traditional Q learning rule is in (2). This new scheme is called a *temporal difference learning* (TD learning) where α is the learning rate.

$$Q^{new}(s,a) = Q^{old}(s,a) + \underbrace{\alpha(s,a)}_{\text{learning rate}} \left[\overbrace{\underbrace{R_{t+1}}_{\text{reward}}(s,a) + \underbrace{\gamma}_{\text{dis. fac.}} \max_{a'} Q^{old}(s',a') - Q^{old}(s,a)}^{\text{Temporal difference(TD)-Error}} \right] \tag{2}$$

For alpha value, the easiest approach is to take a fixed value α, but even better to use a *variable* α. Equation (3) is one possibility where $b(s, a)$ is the number of times the action a has been chosen in the state s.

$$\alpha(s,a) = \frac{1}{b(s,a)} \tag{3}$$

The question now is what is the best action selection strategy. There are two extreme possibilities, one extreme is to always choose the action randomly, this what is called *exploration*, just explore what the environment gives in terms of feedback. In the beginning of our learning process this is a good idea. But after some time when the reinforcement learning based density control agent (RLCA) already learned, we can try other alternative which is to select the best action. This is called *exploitation*. In our case we used a combination of exploitation and exploration by using ε-greedy action selection strategy, that is with a certain probability ε our agent always selects a random action.

RLCA interacts with the traffic network. The agent receives the traffic state s_t from the network detectors. Such state consists of the number of vehicles N_t in the area of interest associated with RLCA. Based on this information the RLCA chooses an action a_t (either green or red). As a consequence of a_t the agent receives a reward r_t as an evaluation of the quality of this immediate transition from s_t to s_{t+1}. The state space, the action space and the reward function of the single control agent were described in [15]. The joint action space and the (joint and global) payoff function of collaborative multi-agent will be described in detail in Sect. 3 and 4.

3 Payoff Propagation

In this section we will discuss the problem of dynamics in MASs and the suggestion that the agents should take the behavior of other agents into account.

3.1 Coordination Graphs

In collaborative multi-agent system, the coordination graph (CG) framework [16] assumes the action of an agent i only depends on a subset of the other agents, $j \in \Gamma(i)$ who may influence its state. The global payoff function $u(\mathbf{a})$ is then broken down into a linear combination of local payoff functions $f_i(\mathbf{a}_i)$. Our design of the local and global payoff will be found in Sect. 4.1 and 4.2 respectively. The global Q function $Q(s, a)$ is then decomposed into a sum of local functions given by:

$$Q(\mathbf{s},\mathbf{a}) = \sum_{i=1}^{n} Q_i(s_i,a_i) \tag{4}$$

Where: $\mathbf{a} = (a_1, \ldots, a_n)$ is the joint action resulting from: each agent i selecting an action a_i from its action set A_i, $\mathbf{s} = (s_1, \ldots, s_n) \in S_1 \times \cdots \times S_n$ is the joint state. The joint Q function $Q_i(s_i, \mathbf{a}_i)$ depends only on $\mathbf{a}_i \subseteq \mathbf{a}$, where $\mathbf{a_i} \in A_i \times \prod_{j \in \Gamma(i)} A_j$, $\Gamma(i)$ is the set of neighborhoods of agent i.

3.2 Max-Plus

The max-plus algorithm approximates the optimal joint action by iteratively sending locally optimized messages between connected (*neighbors*) agents in the coordination graph. This message mainly depends on the decomposition given by (4) and can be defined as follows:

$$\mu_{ij}(a_j) = \max_{a_i} \left\{ f_i(a_i) + f_{ij}(a_i,a_j) + \sum_{k \in \Gamma(i) \setminus j} \mu_{ki}(a_i) \right\} \tag{5}$$

Where: $k \in \Gamma(i) \setminus j$ is the subset of all neighbors connected to i except j, and $\mu_{ij}(a_j)$ is the message from an agent i to agent j after agent j preform the action a_j as an evaluation of the influence of this action on agent i state.

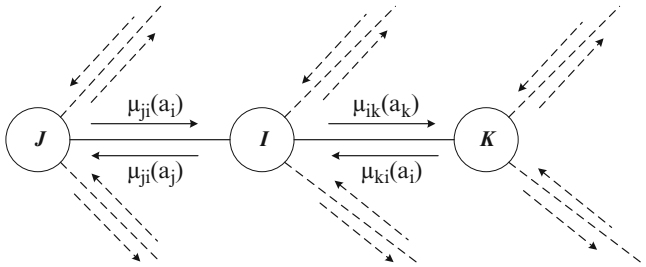


Fig. 1 Representation of the different μ_{ij} sent between neighboring agents

The message approximates the maximum reward agent i can achieve for a committed action of agent j. This message is the sum of the locale payoff $f_i(a_i)$, the joint payoff $f_{ij}(a_i, a_j)$ and all incoming messages agent i received except that from agent j. The design of the joint payoff will be found in detail in Sect. 4.2. Figure 1 shows a CG with three agents and the corresponding messages.

If the graph has no cycle, max-plus always converges after a finite number of steps to a fixed point in which the messages are changed below a certain threshold [17]. At each time step, we can determine the value for an action a_i of agent i as follows:

$$g_i(a_i) = f_i(a_i) + \sum_{j \in \Gamma(i)} \mu_{ji}(a_i) \tag{6}$$

When the convergence is held optimal joint action **a** has the element a_i^* which can be computed as follows:

$$a_i^* = \arg\max_{a_i} g_i(a_i) \tag{7}$$

4 Proposed Framework Based on Multi-Agent and Cooperative Q-Learning

In this section we study the design of different approaches that control the amount of vehicles entering the mainstream freeway from the ramp to keep the freeway density below the critical density. Consequently, this leads to maximum utilization of the freeway without entering in congestion while maintaining the optimal freeway operation.

4.1 Independent Agents (IAs)

The first design to solve the freeway congestion problem is the independent learning. In this design the agents have a greedy behavior, where each control agent chooses an action (either red or green) that maximizes its local reward. The Q function is updated as follows:

$$Q_i(s_i, a_i) = Q_i(s_i, a_i) + \alpha \left[R(s_i, a_i) + \gamma \max_{a'_i} Q(s'_i, a'_i) - Q(s_i, a_i) \right] \tag{8}$$

Where: $R(s_i, a_i) = 1/|\rho(s_i, a_i) - \rho_{cr}|$, with such conception of the objective function the agents try to keep the the freeway density within a small margin of the critical ratio. That ensures the maximum utilization of the freeway without entering in congestion. In this Q function, design the agents are partially observing the environment. An agent observes its state which is defined as the number of vehicles in the areas of interest associated with that agent and chooses the local action either red or green, independent of all other agent actions. This is a cheap design recommended when the distance between ramps is too long.

4.2 Coordinated Reinforcement Learning with Max-Plus

The second design is considered a completely distributed design. This design works on the network level and is based on modeling by *collaborative Markov Decision Process* and an associated *cooperative Q-learning* algorithm which is based on *payoff propagation* algorithm under the *coordination graph* framework. In this design the control agent coordinates its actions with its neighboring agents. The agent updates the cooperative Q function globally as follows:

$$Q_i(s_i, \mathbf{a_i}) = Q_i(s_i, a_i) + \alpha \left[R(\mathbf{s}, \mathbf{a}) + \gamma \max_{a'} Q(\mathbf{s}', \mathbf{a}') - Q(\mathbf{s}, \mathbf{a}) \right] \tag{9}$$

Where: $Q(\mathbf{s}, \mathbf{a})$ can be computed using (4) and $R(\mathbf{s},\mathbf{a})$ can be figured by our conception of the global reward function using the *harmonic mean*. With such design we guarantee the balance between the control agents payoffs, consequently optimal response to the freeway dynamics, as follows:

$$R(\mathbf{s}, \mathbf{a}) = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{R(s_i, a_i)}} \tag{10}$$

The maximum control action $\mathbf{a}'$ in state $\mathbf{s}'$ and its associated estimation of optimal future value $\max_{\mathbf{a}'} Q(\mathbf{s}', \mathbf{a}')$ can be computed using Max-plus algorithm with our design of the joint payoff between two neighboring control agents. With such design we guarantee the balance between connected control agents, as follows:

$$f_{ij}(a_i, a_j) = \frac{2 * f_i(a_i) * f_j(a_j)}{f_i(a_i) + f_j(a_j)} \tag{11}$$

This is a moderate design recommended when the distance between ramps are short and the study network has many ramps.

4.3 Centralized Agent (CA)

The third extreme design is considered a centralized one where the collaborative multi-agent can be considered as one large single agent, in which the joint actions are represented as one single action. The cooperative Q function for the joint actions are updated using a single Q function, as follows:

$$Q(\mathbf{s}, \mathbf{a}) = Q(\mathbf{s}, \mathbf{a}) + \alpha \left[R(\mathbf{s}, \mathbf{a}) + \gamma \max_{\mathbf{a}'} Q(\mathbf{s}', \mathbf{a}') - Q(\mathbf{s}, \mathbf{a}) \right] \tag{12}$$

Nevertheless, this design leads to an optimal solution, it is not scalable as it suffers from the curse of dimensionality. This is a costly design recommended when the distance between ramps is short and the study network has fewer numbers of ramps.

5 The VISSIM Simulation Environment

There exist several types of traffic simulator such as [18, 19]. While VISSIM is considered one of the best technology for advanced simulation and very well suited for our purpose. That is because of its flexibility through *VISSIM COM interface* and *VISSIM DLL API* which provide freedom to traffic researchers and engineers.

In VISSIM there are many functions and parameters which control the VISSIM itself and associated study experiments, these parameters can be assigned through VISSIM GUI and remain fixed during the running of the experiment or manipulated through programing via VISSIM COM interface which gives us the opportunity to change these parameters during the running time. For example, changing the traffic signal during the running time of the experiment to respond to the dynamics of the freeway. VISSIM COM interface can be programmed via any type of programming language with the ability to handle COM object.

Accordingly, the next section will investigate the advance VISSIM simulator possibilities, particularly ramp control development. We give one example of a dense network with three ramps, but our proposed solution can handle any type of networks with any number of ramps.

6 Experimentation: Results and Analysis

The studied network in Fig. 2 consists of a mainstream freeway with three lanes and three-metered on-ramp with one lane each. The network consists of four sources of inflow: O_1 for the mainstream flow, O_2, O_3, and O_4 for the three on-ramp flow; it has only one discharge point D_1 with unrestricted outflow. The mainstream freeway consists of three areas of interests: A_1, A_2, and A_3. In addition, three control agents (RLCA) which are located at the entrance points of each ramp. Each area of interest is associated with one control agent. Each area is about 1350m as follows: 580m before the ramp, 250m as a merging area and 520m after the ramp. There are four uncontrolled sections; three of them before A_1, A_2 (which is about 1500m), and A_3 (which is about 100m), the fourth is after A_3.

Figure 3 shows the demand for both the mainstream freeway and ramps. This demand scenario gives us the

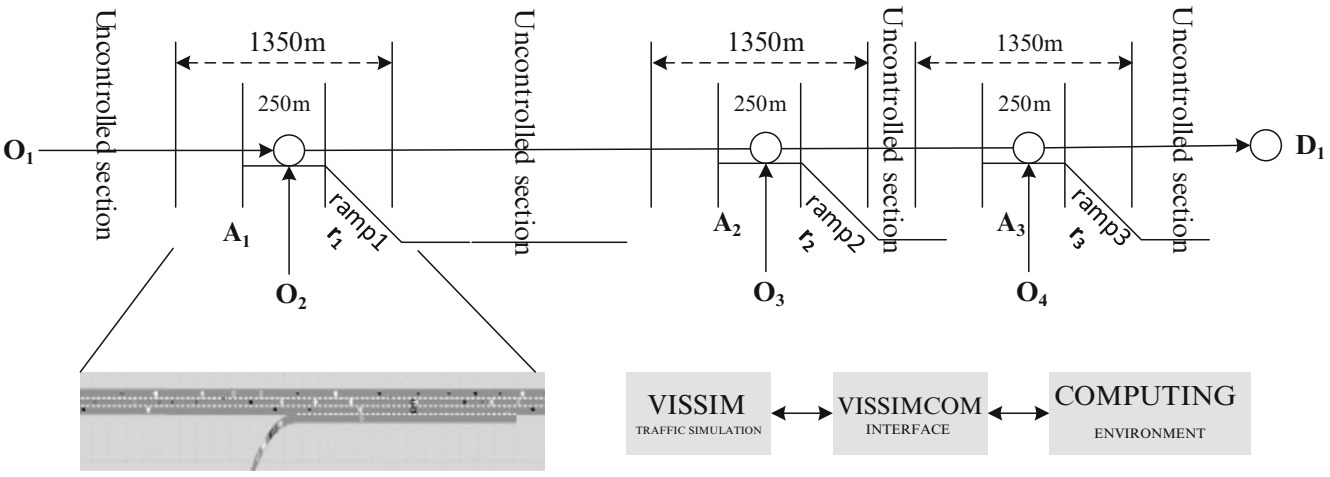


Fig. 2 The studied network

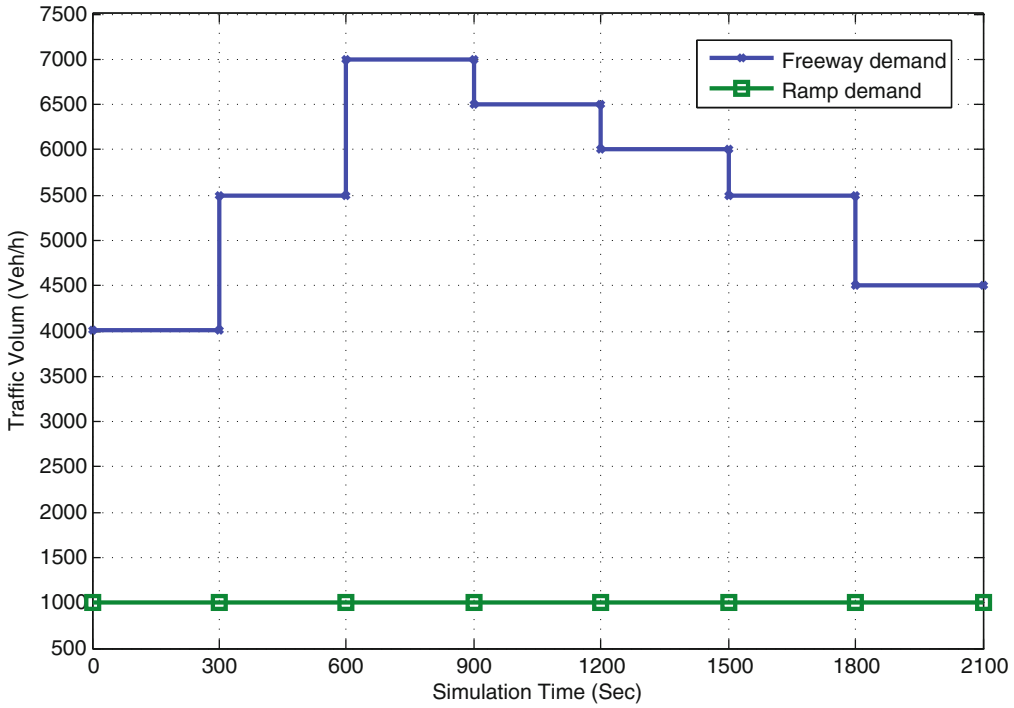


Fig. 3 The demand scenario

Table 1 Proposed framework performance evaluation per agent

	$Agent_1$		$Agent_2$		$Agent_3$			
		A_1	r_1	A_2	r_2	A_3	r_3	A
	IAs	109.05	89.7	112.06	113.07	97.53	80.92	401.79
	Max-plus	112.87	69.18	100.34	77.36	94.18	54.93	388.85
TT(AVG)(s)								3.2%
	CA	110.05	70.47	99.09	77.9	93.68	54.01	385.32
								4%

opportunity to study the effect of the on-ramp smart control. That is because with such scenario the density ρ can exceed the critical density ρ_{cr} which is proven during the learning process random choice of action can lead us to a scenario with density ρ equals 72 (Veh/km) which is higher than the critical one ρ_{cr} which is equal to 62 (Veh/km). The studied network is considered a dense network where the smart control system is highly needed.

Table 1 shows the proposed framework performance evaluation (the three approaches) per agent where A defined as the area of the freeway, which starts from the beginning of A_1 to the end of A_3. And r_1, r_2, r_3 are defined as areas starting from the beginning points of the ramp 1, 2 and 3 to the end points of A_1, A_2 and A_3 respectively. In the IAs the $agent_1$ tries greedily to solve his local congestion only regardless of other agent's, problems. This leads to overpopulating section A_2 of the road, hence increasing the average travel time for A_2 and r_2 to 112 and 113 respectively, and the same for $Agent_3$. In contrary, in coordinated reinforcement learning with max-plus, the agents optimally cooperate to resolve the congestion problem which leads to 3.2% advance over the IAs case. Although the CA gives us some improvements, we do not recommend this costly solution compared to the RL with a max-plus. Figure 4 gives us the density associated with each approach of the framework. Particularly, Fig. 4a shows that $Agent_1$ tries to maintain its density within a small neighborhood of the critical density regardless of other agents' performance.

Table 2 provides the results obtained from the proposed framework with all the three approaches and compares it to the base case (no-metering). It can be seen that coordinated reinforcement learning with max-plus and centralized agent have shown considerable improvement over the base case. Centralized agent and Cooperative Q-learning with max-plus gives 6.5% and 6.9% in terms of total travel time and 6.74% and 7.5% in terms of average speed improvements over the base case respectively.

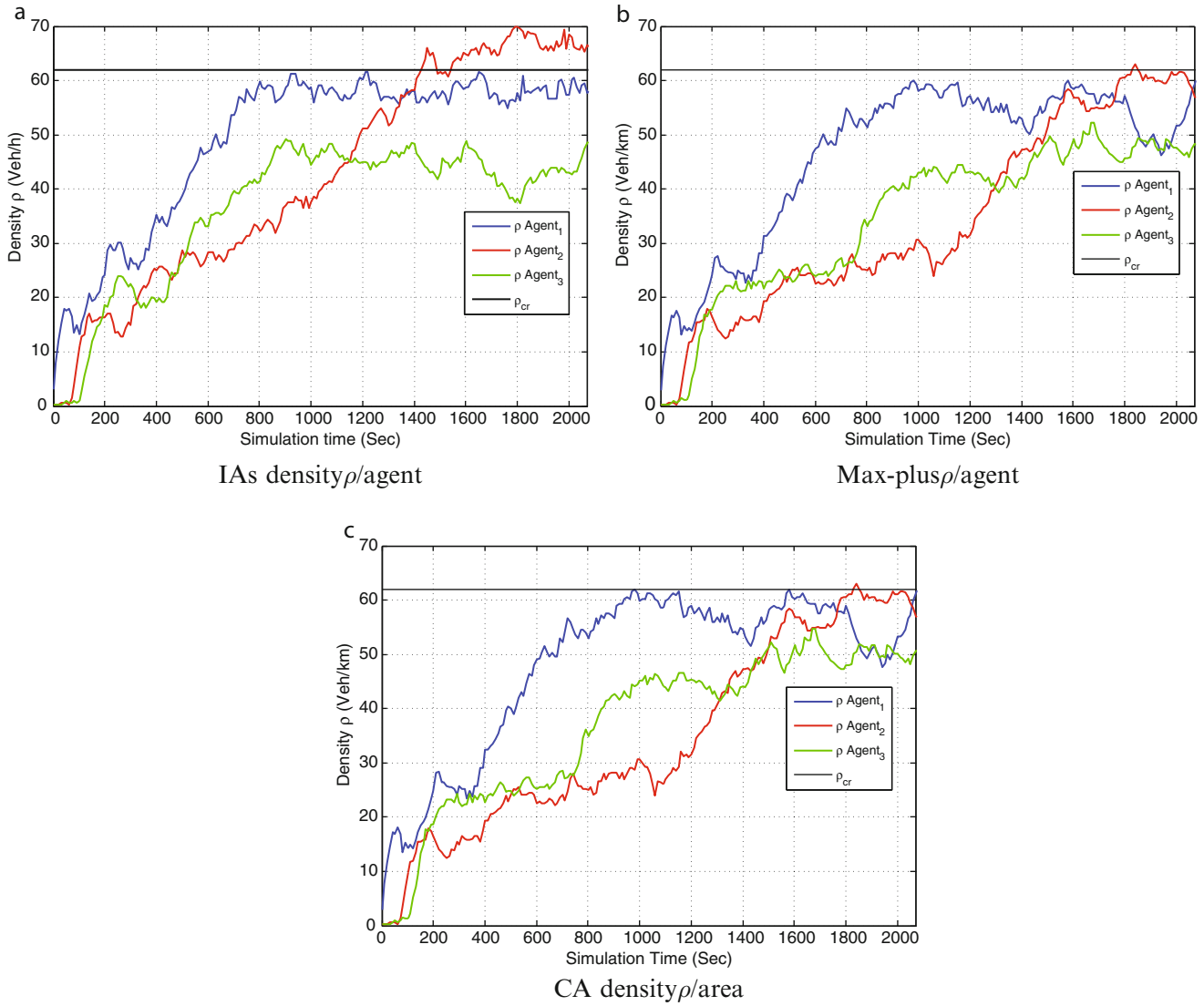


Fig. 4 Freeway density for the proposed framework with the three approaches

Table 2 Proposed framework performance evaluation

	No-metering	IAs	Max-plus	Centralized agent
Total travel time (h)	333	331.5	311.5	308.7
		0.5%	6.5%	6.9%
Average speed (km/h)	45.25	45.01	48.5	48.65
		0.5%	6.74%	7.5%

We give one example of a dense network with three ramps, but our proposed framework can handle whatever type of networks with any number of ramps. In addition, our recommendation for this type of network and similar one is to apply the second approach of our proposed framework which is cooperative Q-learning with max-plus.

7 Conclusion

In this paper, we addressed the problem of traffic congestion in freeways at the network level. A new system for controlling ramp metering was introduced based on a multi-agent

reinforcement-learning framework. Our proposed framework comprises both a Markov Decision Process Modeling technique and an associated cooperative Q-learning technique, which is based on payoff propagation (max-plus algorithm) under the coordination graph framework. For our solution to be as optimal as possible, we considered different modes of operation depending on the network architecture. This framework was tested on the state-of-the-art VISSIM traffic simulator in a dense practical scenario. The experimental results were thoroughly analyzed to study the performance of the proposed framework using a concrete set of metrics, namely, total travel time and average speed while keeping freeway density at optimum level close to the critical ratio for maximum traffic flow. Our findings proved that our system achieved significant enhancement on both features as compared to the base case. The advantages of applying our system include saving car fuel, decreasing air pollution, and could save lives by reducing the chances of car accidents.

Acknowledgments This research is supported by the Ministry of Higher Education (MoHE) of Egypt through PhD fellowships. Our sincere thanks to E-JUST University for guidance and support.

References

1. Yu, X., Xu, W., Alam, F., Potgieter, J., Fang, C.: Genetic fuzzy logic approach to local ramp metering control using microscopic traffic simulation. In: Mechatronics and Machine Vision in Practice (M2VIP), 2012 19th International Conference. (2012) 290–297
2. Ghods, A., Kian, A., Tabibi, M.: Adaptive freeway ramp metering and variable speed limit control: a genetic-fuzzy approach. Intelligent Transportation Systems Magazine, IEEE **1**(1) (2009) 27–36
3. Liang, X., Li, J., Luo, N.: Single neuron based freeway traffic density control via ramp metering. In: Information Engineering and Computer Science (ICIECS), 2010 2nd International Conference on. (2010) 1–4
4. Li, J., Liang, X.: Freeway ramp control based on single neuron. In: Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on. Volume 2. (2009) 122–125
5. Feng, C., Yuanhua, J., Jian, L., Huixin, Y., Zhonghai, N.: Design of fuzzy neural network control method for ramp metering. In: Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on. Volume 1. (2011) 966–969
6. Veljanovska, K., Gacovski, Z., Deskovski, S.: Intelligent system for freeway ramp metering control. In: Intelligent Systems (IS), 2012 6th IEEE International Conference, IEEE (2012) 279–282
7. Davarynejad, M., Hegyi, A., Vrancken, J., van den Berg, J.: Motorway ramp-metering control with queuing consideration using q-learning. In: Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on, IEEE (2011) 1652–1658
8. Ji, X., He, Z.: An optimal control method for expressways entering ramps metering based on q-learning. In: Intelligent Computation Technology and Automation, 2009. ICICTA'09. Second International Conference on. Volume 1., IEEE (2009) 739–741
9. Vlassis, N., Elhorst, R., Kok, J.R.: Anytime algorithms for multiagent decision making using coordination graphs. In: Systems, Man and Cybernetics, 2004 IEEE International Conference on. Volume 1., IEEE (2004) 953–957
10. Guestrin, C.E.: Planning under uncertainty in complex structured environments. PhD thesis, Stanford University (2003)
11. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann (1988)
12. Ernst, D., Glavic, M., Capitanescu, F., Wehenkel, L.: Reinforcement learning versus model predictive control: a comparison on a power system problem. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **39**(2) (2009) 517–529
13. Watkins, C.J., Dayan, P.: Q-learning. Machine learning **8**(3–4) (1992) 279–292
14. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. arXiv preprint cs/9605103 (1996)
15. Ahmed Fares, W.G.: Freeway ramp-metering control based on reinforcement learning. In: Control and Automation, 2014. ICCA 2014. 11th IEEE International Conference on. (in press)
16. Papageorgiou, M., Hadj-Salem, H., Blosseville, J.M.: Alinea: A local feedback control law for on-ramp metering. Transportation Research Record (1320) (1991)
17. Wainwright, M., Jaakkola, T., Willsky, A.: Tree consistency and bounds on the performance of the max-product algorithm and its generalizations. Statistics and Computing **14**(2) (2004) 143–166
18. Maciejewski, M.: A comparison of microscopic traffic flow simulation systems for an urban area. Problemy Transportu **5** (2010) 27–38
19. Prendinger, H., Gajananan, K., Bayoumy Zaki, A., Fares, A., Molenaar, R., Urbano, D., van Lint, H., Gomaa, W.: Tokyo virtual living lab: Designing smart cities based on the 3d internet. Internet Computing, IEEE **17**(6) (Nov 2013) 30–38

Intelligent system concept for high-energy performance and adaptable user comfort

Vladimir Tanasiev, Adrian Badea, Cristian Dinca, and Horia Necula

1 Introduction

Primary energy consumption of any modern society is assured either from domestic or from external production and is distributed over the following sectors: transportation, residential and tertiary, industry and agriculture.

Distribution of consumption is specific to each country. At EU level it was found that the largest consumer is the residential and tertiary sector (Table 1).

Residential and tertiary sector are responsible for 39.68 % of the total consumption of primary energy in the EU [1]. Being the sector with biggest energy share it also has the biggest potential to obtain energy efficiency. In this respect EU launched an energy efficiency trend, which implies that all buildings built after 2020 to be net zero energy buildings [2]. Currently there are a variety of energy efficiency solutions in Europe, among those are included: passive houses [3–5], intelligent buildings [6–10], net zero energy houses [11–14]. Intelligent management solutions, which take into account the technological context and user's needs to be in permanent control, allow better management and greater control over building resources. From the technological point of view current technology allows computing machines to consume energy proportional to the complexity of the calculations. In terms of size of the computing devices, progress is remarkable, current technology allows realization of powerful hardware with low dimensions [15–18].Integrated development platform together with software development environments have reached a high maturity level which allows applications with rich graphical content to be installed on both mobile devices and computers (web or desktop applications). New generation of software application based on SOA (Service Oriented Architecture) and HTML 5 (HyperText Markup Language) will revolutionize HMI (Human Machine Interface) interaction [19].

Installations and equipment (heat pumps, heat recovery units, air-conditioning units, etc.) successfully integrated programmable modules that allow the exchange of information with other devices based on standardized communication protocols (BACnet, KNX, Instabus, RS 232, USB, ZigBee).

From the technological point of view there are opportunities to develop an intelligent solution for managing energy consumption and user comfort in buildings. In order to develop and intelligent system for energy and user comfort management, we must answer to the question: What is an intelligent building?

Currently there are more than 30 definitions, among them are worth mentioned [6]:

- Buildings which have fully automated building service control systems (Cardin 1983)
- Buildings which 'provide information' from an intelligent operator to act upon (Fagan 1985)
- The type of building, which harnesses and integrates all levels of IT form data processing and environmental control and security (David S. Brockfield 1987)
- An intelligent building is one that maximizes the efficiency of the occupants while at the same time minimizing the costs associated with running the building (David Boyd 1994)
- A building that creates an environment that maximizes the efficiency of the occupants of the building while at the same time allowing effective management of the resources with minimum lifetime costs. (IBC – Intelligent Building International)

According to Arkin and Paciuk in [6] intelligent buildings must provide the environment and resources for optimal use of the building in accordance with its purpose. Building

V. Tanasiev (✉) • C. Dinca • H. Necula
Faculty of Power Engineering, University POLITEHNICA of Bucharest, Bucharest, Romania
e-mail: vladimir.tanasiev@energ.pub.ro; crisflor75@yahoo.com; horia@energ.pub.ro

A. Badea
Romanian Academy of Scientists, Bucharest, Romania

Faculty of Power Engineering, University POLITEHNICA of Bucharest, Bucharest, Romania
e-mail: adrian.badea@upb.ro

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_26,

intelligence is not related to the sophistication of service systems in a building but rather integration between different services and between systems and building structure.

In the book "intelligent skin" the authors (Wigginton M., Harris J.) relates the intelligence to the possession of intellectual facilities, which provide a capacity for understanding. There is an inferred ability to provide and comprehend meaning and apply this acquired knowledge through the thinking process of reasoning.

2 Intelligent building concept

Currently, a building is considered Smart or Intelligent if it incorporates the best available concepts, materials, systems and technologies integrating these to achieve or exceed the performance requirements of the buildings stakeholders [20].

Nowdays energy needs and use of users vary considerably compared to 30 years ago. As well as comfort needs become more and more sophisticated, which bring us to the question: how can we respond properly to increasing level of comfort and reduce at the same time energy consumption?

The answer is enclosed in an intelligent building system which is able to collect building information, interact with building owners and at the same time being able to create and execute building policies in order to reduce energy consumption while maintaining users comfort at agreed parameters.

The realization of intelligent building system relies in development of a mathematical linear model which will be able to approximate energy losses through building elements and development of non-linear model for determining user comfort behavior. Figure 1 presents the functional structure of the intelligent management system concept.

Table 1 Primary energy consumption by sectors in TOE, 2012.[1]

Country	Residential	Tertiary	Transport	Others	Industry	Agriculture
Romania	8060,6 (35,51 %)	1763,2 (7,77 %)	5345,1 (23,55 %)	244,7 (1,08 %)	6786,8 (29,90 %)	497,7 (2,19 %)
EU (28)	282316,6 (26,21 %)	148479,3 (13,47)	351.080 (31,85 %)	7924,5 (0,72 %)	282316,6 (25,61 %)	23638 (2,14 %)

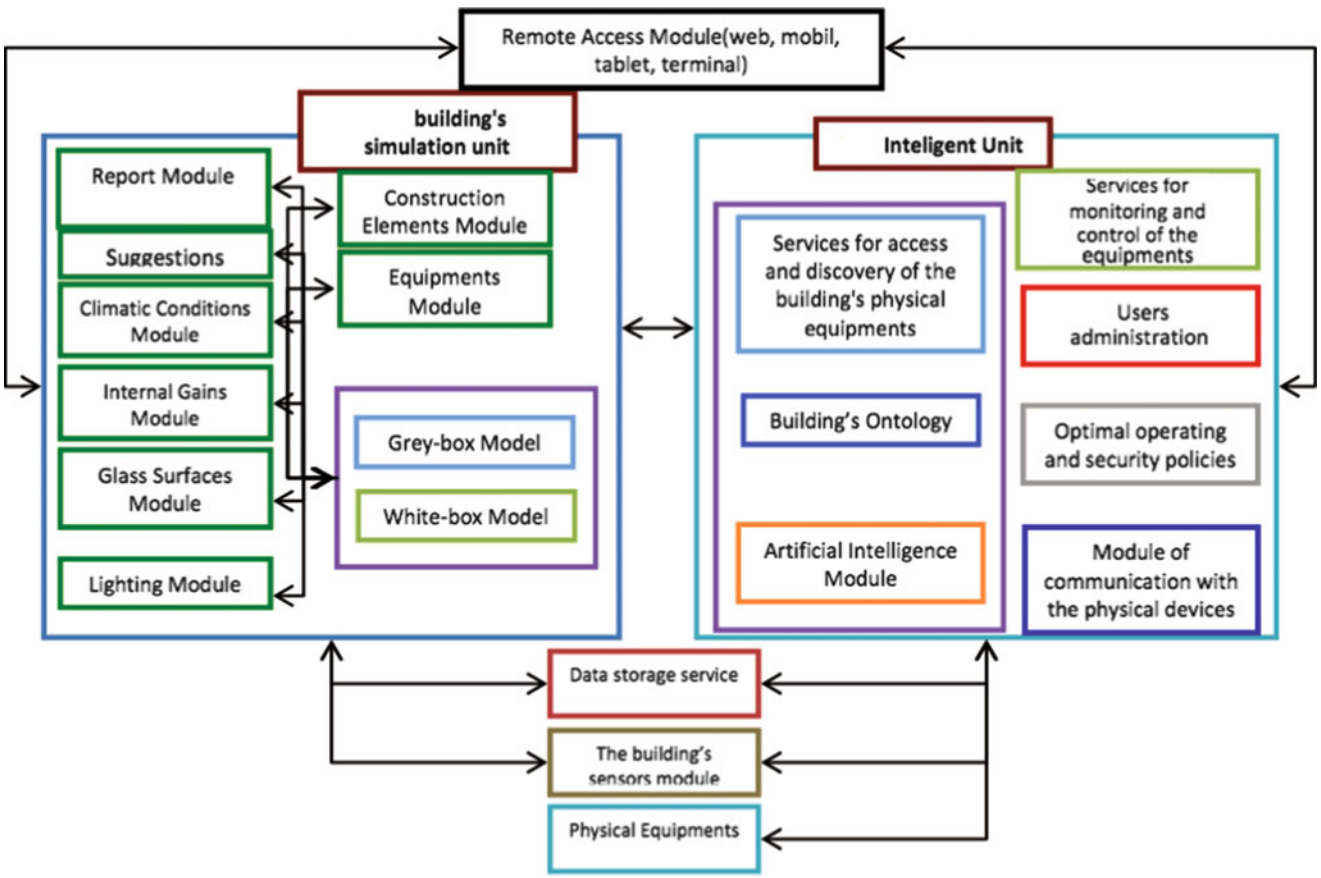


Fig. 1 Functional structure of an intelligent building.[26]

The following infrastructure has as a starting point the identified needs from both segments: buildings equipped with automated systems and dynamic applications for simulation of energy's consumption in buildings.

Intelligent component's infrastructure is organized in two different entities. The structural concept gives the possibility to use simulation unit in conjunction with intelligent unit or separately.

2.1 Simulation unit of the building (SUB)

SUB uses the following modules: "internal input", "climatic conditions", "glass surfaces", "construction elements", "lighting", for determining the energy requirement of the building, but also the visual and thermal comfort of the user.

Reporting module allows users to view the building's consumption and other relevant information related the users. The "suggestions" module has a double role: in projection stage of the building, it can help users to model the building using a cost-benefit analysis, and for the buildings that implement the intelligent system, this module can recommend appropriate temperatures for day and night and different schedules for the installations that can reduce the energy consumption.

The key component of the SUB is represented by the energy prediction algorithm. The most suitable algorithm is "grey-box" prediction model.

The algorithm uses a combined approach of "white-box" and "black-box" algorithm and is based on transfer function model which satisfy a physical model for energy flows in the building structure. The model foundation is based on calculation of conduction transfer function using state space method (Ceylan and Myers 1980; Seem 1987; Ouyang and Haghighat 1991). The basic state space system is defined by the following linear matrix equations:

$$\begin{cases} \dfrac{dx}{dt} = Ax + Bu \\ y = Cx + Du \end{cases} \quad (1)$$

where x is a vector of state variables, u is a vector of inputs, y is the output vector, t is time, and A, B, C, and D are coefficient matrices. Through the use of matrix algebra, the vector of state variables (x) can be eliminated from the system of equations, and the output vector (y) can be related directly to the input vector (u) and time histories of the input and output vectors [23].

For optimal parameters of the building a learning process with three steps minimizes the errors using at the same time a low computational effort:

- Bounds on physical parameters are estimated from a rough description of the building geometry and materials.
- A global direct search algorithm is used to determine estimates of the model parameters that satisfy the constraints on parameters and the physical representation.
- A nonlinear regression algorithm that relies on local derivative information is used to deter- mine optimal parameters that minimize errors between model predictions and measurements and satisfy the constraints.

In the global search, the search for an optimum occurs over the entire feasible parameter space and a discrete set of cost function evaluations are performed for specified points within the chosen parameter space [24].

2.2 Intelligent unit (IU)

The IU is composed of one set of services and modules that allow the discovery, verification, composing, monitoring and control of the building's equipment for creating thermal and visual comfort of the user. The IU uses as input the result of the unit's simulation. The IU has to maintain the efficient consumption of energy, therefore there are three operating systems used for reduced energy usage: the passive operation, which creates ways of achievement thermal comfort by using passive sources (opening the windows, opening or closing the shutters and others), hybrid operations, that allow using both mechanical and passive sources and mechanical operations that allow achievement of thermal comfort only by using the available installations of the building.

The IU manages the security policies that protect the buildings installation, allows editing the comfort policies by logged users and creates new policies for optimum functioning, which have as main objective obtaining energy efficiency in buildings.

The ontology of the building represents a basic entity that gives direction to the intelligent system. The ontology "knows" the following elements: building's resources, security and operating policies set by the users, the availability of the physical equipment connected to the physical environment, access rights of the users and their associate profiles. The ontology of the building is the commanding element of the intelligent system that takes decisions based on data analysis.

The artificial intelligent module "learns" user's behavior based on their actions and it's able to anticipate their future actions. This module uses a set of probabilistic methods that anticipates user's behavior according to the number of the occurrences of their events. The intelligent unit receives data from the building's sensors, but also from the simulation unit. Based on intelligent computation algorithms, the intelligent unit will transmit to the "ontology" what decisions are necessary for event anticipation.

3 Remote access module

Remote access module gathers all ways upon which the users can communicate with the intelligent management system: computer, tablet, mobile phone. An easy way to access building's resources can be realized through internet browsers or through specific applications that can be installed on mobile devices or tablets.

The graphical user interface (GUI) can serve for building editing, user management and building control. Each user with executing rights can edit a building policy for thermal comfort which can be implemented as long as it does not interfere with other existing policies (security policies or other running user policies).

The graphic interface it will be intuitive, easy to access and offers to users a pleasant experience when interacting with the IU. Figure 2 presents an example of a web interface that can be used for editing the constructive elements of the building.

Viewing the consumption reports, the access to the suggestions module, the list of building materials and climate data is organized in the top horizontal menu and editing the construction items is available in the horizontal menu at the bottom or in the vertical menus the right.

The transfer of data between intelligent unit and GUI can be based on **Service oriented architecture (SOA).** SOA brings a design pattern which is based on division of software in small tasks called services, that can be used locally or accessed through Internet (Figure 3).

The biggest advantage of SOA is that it allows flexibility in selecting the programming language and allows intercommunication between two different software products.

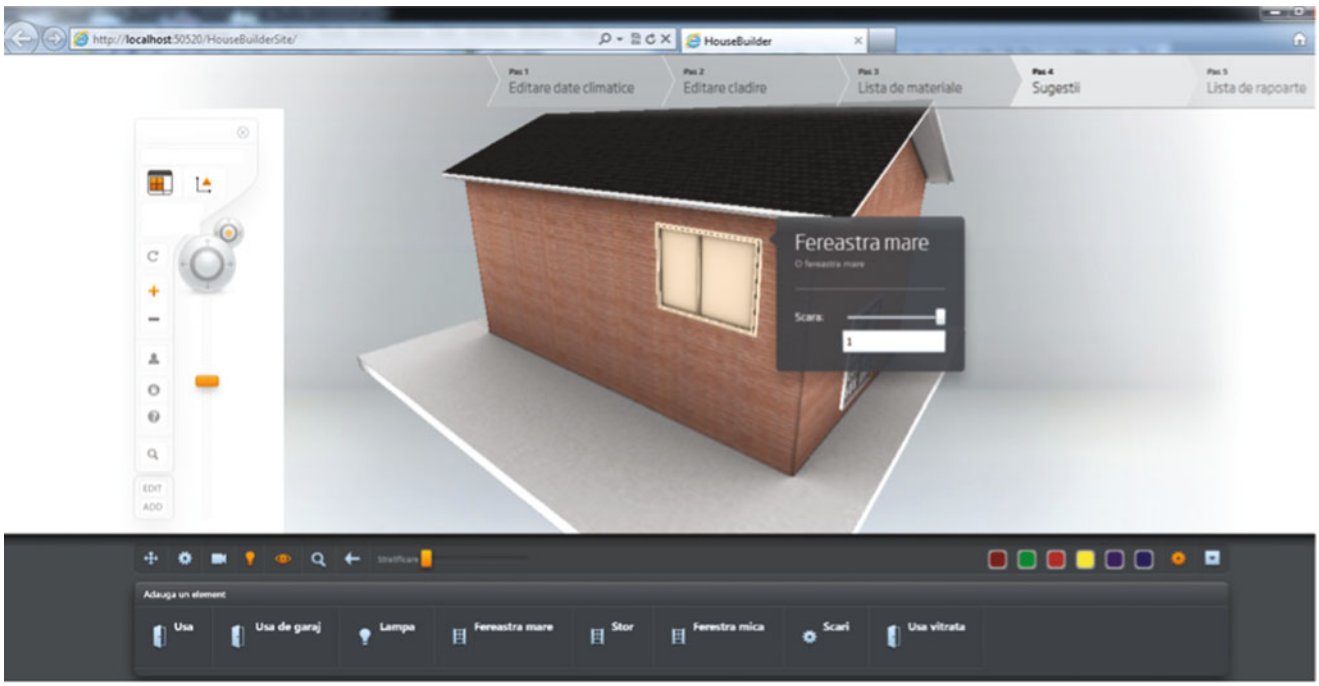


Fig. 2 Interface example of remote access module

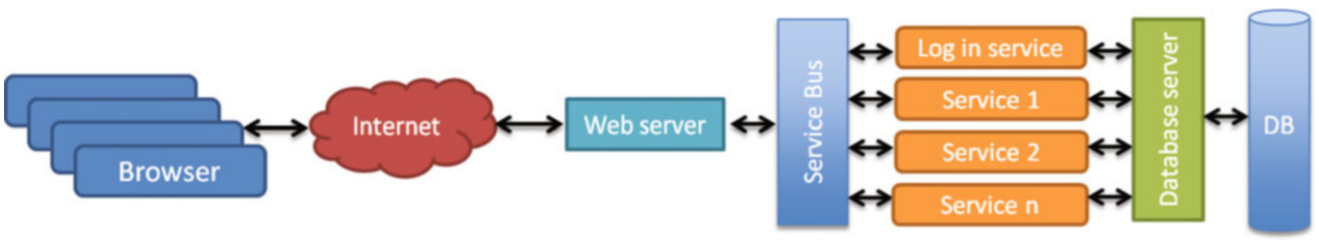


Fig. 3 Service oriented architecture

4 Hardware infrastructure

Hardware infrastructure of an intelligent management system should be adaptable in order to respond to a large variety of cases. Generally the infrastructure of a smart building should be composed of:

- at least one data transfer medium (wireless or cable).
- at least one intelligent unit that can make the most effective decisions.
- a system for command execution.
- at least one method of human interaction with intelligent unit.
- a way to remote access

Currently the cost of the implementation of the smart building's intelligent component decreased mainly for three reasons: lower prices for basic components, lower prices for software products and a more simple way for communication with the physical devices. Figure 4 shows an overview of an intelligent management system hardware infrastructure.

The communication server should be able to establish connections over the internet in order to avoid ip routings and mismatches. Another role of the server should be automatic domain allocation for users at login.

Primary bus should connect high level controllers while secondary bus should connect low level controllers.

Over the internet

5 Conclusion

Present article does not propose to solve all the issues related to the concept of intelligent building, instead it tries to outline a way that can be accomplished intelligent management system of a building. By implementing a set of user policies, in accordance with owners working schedule, the required energy for heating can be reduced with more than 20 % [25]. Smart grids or smart city, a new concept that emerged in the present, have as basic element intelligent management system. Another great advantage of implementation of such system is a better synchronization between energy produced and energy consumed.

Acknowledgements The study has been funded by the UEFISCDI within the National Project number 38/2012 with the title: „Technical-economic and environmental optimization of CCS technologies integration in power plants based on solid fossil fuel and renewable energy sources (biomass)" – CARBOTECH.

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/134398.

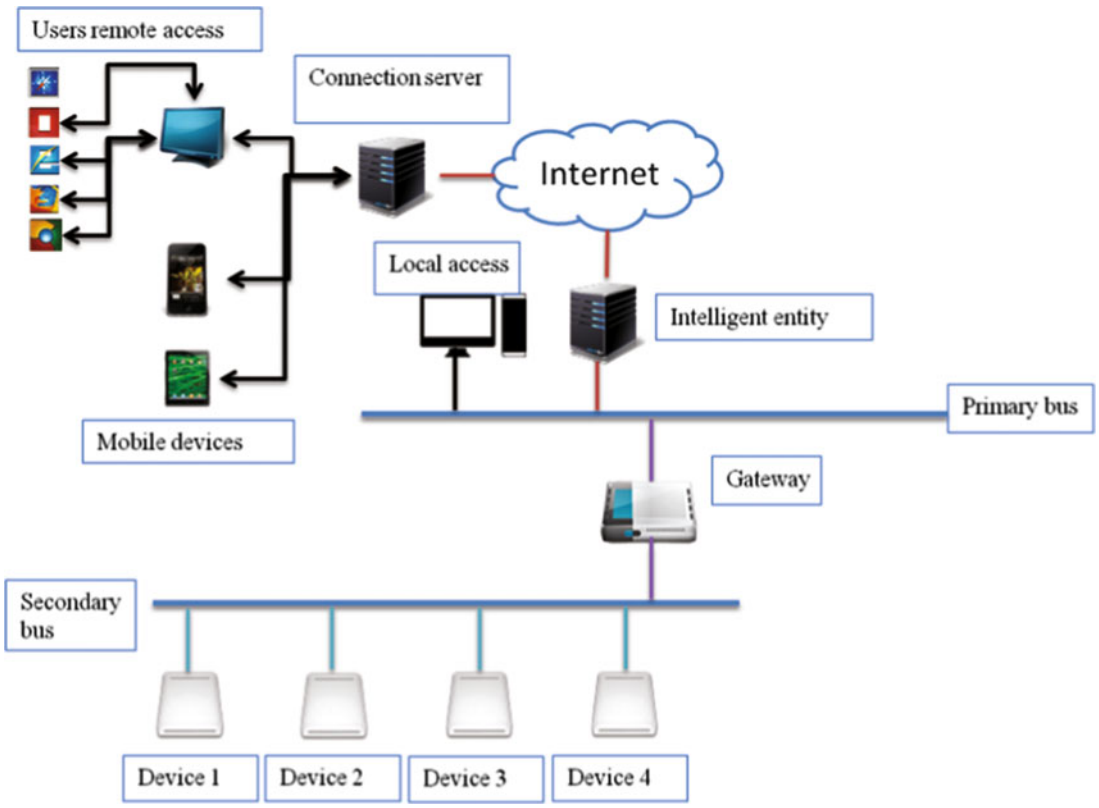


Fig. 4 Hardware infrastructure of the smart building

References

1. Eurostat, Final energy consumption, http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&init=1&plugin=1&language=en&pcode=tsdpc320, 2012.
2. European Parliament and Council. Directive 2010/31/EU of the European Parlia- ment and of the Council of 19 May 2010 on the energy performance of buildings. Official Journal of the European Union 2010.
3. IEEA (The 'Passive-On' project) - The passivhaus standard in european warm climates - Part 1. A review of comfortable low energy homes, EC funded project Passive-on ('Marketable Passive Homes for Winter and Summer Comfort' EIE/04/091/S07.38644, 2004-2007), Nottingham, 2007.
4. About Passive House - What is a Passive House?, http://passiv.de/en/02_informations/01_whatisapassivehouse/01_whatisapassivehouse.htm.
5. Passive Houses in different climates, http://www.passipedia.org/passipedia_en/basics/passive_houses_in_different_climates
6. H. Arkin, M. Paciuk, Evaluating intelligent building according to level of service system integration, Automation in Construction, Issue 6, 471–479, (1997).
7. Johnny K.W. Wong, Heng Li, Application of the analytic hierarchy process (AHP) in multi-criteria analysis of the selection of intelligent building systems, Building and Environment 43, 108–125, (2008).
8. Smith, S., Intelligent buildings, Best, R., Valence, G. (Eds.), Design and Construction: Building in Value, Butterworth Heinemann, 36–58, UK (2002).
9. Kolokotsa, D., Rovas, D., Kosmatopoulos, E., Kalaitzakis, K., A roadmap towards intelligent net zero- and positive-energy buildings, Solar Energy, Issue 85, 3067–3084, (2011).
10. Wonga, J., K., W., Lia, H., Wang, S.,W., Intelligent building research: a review, Automation in Construction, Issue 14, 143–159, (2005).
11. Miller, W., Buys, L., Anatomy of a sub-tropical Positive Energy Home (PEH), Solar Energy, Issue 86, 231–241, (2012).
12. Mitchell Leckner, Radu Zmeureanu, Life cycle cost and energy analysis of a Net Zero Energy House with solar combisystem, Applied Energy, Issue 88, 232–241, (2011).
13. Patxi, H., Kenny, P., From net energy to zero energy buildings: Defining life cycle zero energy buildings (LC-ZEB), Energy and Buildings, Issue 42, 815–821, (2010).
14. Stutterecker, W., Blümel, E., Energy plus standard in buildings constructed by housing associations, Energy, Issue 48, 56-65, (2012).
15. Arduino, http://arduino.cc/en/Main/Products
16. .Net Micro Framework, http://www.ghielectronics.com/catalog/category/7/
17. Android open accessory KIT (AOAA), http://www.embeddedartists.com/products/app/aoa_kit.php
18. BeagleBone low size computer, http://beagleboard.org/bone
19. Anthes, G., HTML5 Leads a Web Revolution, Communications of the ACM, Vol. 55 No. 7, 16-17.
20. Ghiaus, C, Inard, C, Kolokotsa, D., Intelligent buildings – results of Smart Accelerate Project, PREA Workshop, 2006.
21. ISO 13790:2008. Energy performance of buildings? Calculation of energy use for space heating and cooling. Geneva, Switzerland: ISO;(2008).
22. S., Wang, X., Xu. Simplified building model for transient thermal performance estimation using GA-based parameter identification. International Journal of Thermal Sciences;45(4):419–32, (2006).
23. US Department of Energy, EnergyPlus Engineering Reference, pp 35,(2012).
24. Braun,J., E., Chaturvedi, N., An Inverse Gray-Box Model for Transient Building Load Prediction, HVAC&R Research, 8:1, 73-99, (2002).
25. Sicurella, F., Tanasiev, V., Wurtz, E., Achieving energy savings and thermal comfort by coupling insulation, thermal inertia and ventilation strategies: a case study, IBPSA, Chambery, (2012)
26. Tanasiev,V., Badea, A., Energy efficiency through intelligent management of building equipment, U.P.B. Sci. Bull,(2012)

Sparse hidden units activation in Restricted Boltzmann Machine

Jakub M. Tomczak and Adam Gonczarek

1 Introduction

Restricted Boltzmann Machines (RBM) play an important role in deep learning. Single RBM is used as a building block in many deep architectures, such as Deep Boltzmann Machines, which is further used in a layer-wise pre-training [1]. Moreover, RBM has been proven to be a universal approximator of a distribution over binary inputs [12], [15]. Finally, RBM has been applied as feature extractor for further prediction [7] or as standalone non-linear classifiers [9].

Although some efficient learning algorithms have been developed for RBMs, e.g., *Contrastive Divergence* [6], they usually suffer from two issues. First, it is known that the more hidden units are used, the higher likelihood over the training data can be achieved (for reasonable conditions, see [12], [15]). Nonetheless, in practice we face the problem of overcomplete representations, where many of learned features are strongly correlated [13]. Consequently, lots of hidden units have to be used which affects the learning time and increase the risk of overfitting. Second, learning generative models such as RBM allows to obtain good features for reconstruction but weaker for prediction. It has been noticed that one possible way of solving both stated problems is applying techniques which enforce sparse solutions.

Hitherto, several methods for sparsification have been proposed. The simplest technique for obtaining sparse weights matrix utilizes *weight decay* which is equivalent to the application of the ℓ_2 norm to model weights [2]. Recently, a very interesting approach was proposed which aims at introducing a *Tikhonov-type regularization* term that results in a specific kind of correction to the weight decay [3]. Sparse solutions can be also obtained by replacing sigmoid units by *rectified linear units* [17]. It turns out that the rectified linear units incorporate sparseness to the model weights by forcing the response of hidden units to be zero for negative impulses [4]. A completely different approach aims at *sparse activation of hidden units*.[1] It has been advocated that it is beneficial to activate a subset of hidden units to diversify information among hidden units and specialize subgroups of them for specific inputs. In other words, dense representations are highly entangled and any change in the input modifies most of the features [4]. Additionally, it is expected that sparse activation of hidden units would eventually lead to more discriminative features. Techniques within this group of sparse solutions apply a regularization term which try to keep the amount of active hidden units at a fixed small level. One such approach applies the cross-entropy to force sparse hidden units activation [16], while the other – the ℓ_2 norm [10].

Along this line of research, we propose a new approach which utilizes a regularization term using the symmetric Kullback-Leibler divergence to compare the actual and the desired distribution over active hidden units. We evaluate the proposed approach empirically, and show that our approach generalizes the regularization method proposed in [16].

2 Restricted Boltzmann Machine

2.1 The model

Restricted Boltzmann Machine (RBM) is a bipartite Markov random field with one layer consisting of visible variables, $\mathbf{x} \in \{0, 1\}^D$, and the other of hidden variables, $\mathbf{h} \in \{0, 1\}^M$ [18]. In the case of classification, an extension of the RBM can be used, Classification Restricted Boltzmann Machine (ClassRBM), in which an additional unit is added, $\mathbf{y}$,

J.M. Tomczak (✉) • A. Gonczarek
Institute of Computer Science, Wrocław University of Technology, wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: jakub.tomczak@pwr.wroc.pl; adam.gonczarek@pwr.wroc.pl

[1] In [5] such approach is called *selectivity*.

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_27,

connected to the hidden units only. This additional unit corresponds to the class label represented by the 1-to-K coding scheme. With each state ($\mathbf{x}$, $\mathbf{y}$, $\mathbf{h}$) we associate the following energy [9]:

$$E(x,y,h|\theta) = -b^\top x - c^\top h - d^\top y - x^\top W h - h^\top U y \quad (1)$$

where $\theta = \{b, c, d, W, U\}$ denotes parameters. A ClassRBM with M hidden units is a parametric model of the joint distribution over visible and hidden variables that takes the following form:

$$p(x,y,h|\theta) = \frac{1}{Z(\theta)} exp\{-E(x,y,h|\theta)\} \quad (2)$$

where $Z(\theta) = \sum_{x,y,h} exp\{-E(x,y,h|\theta)\}$ is a partition function. An important fact is that the conditional probabilities $p(h_j|x,y,W,U,c)$, $p(x_i \mid \mathbf{h}, \mathbf{W}, \mathbf{b})$, $p(y_k \mid \mathbf{h}, \mathbf{V}, \mathbf{d})$ can be analytically calculated (see [9] for details).

2.2 Prediction

For given parameters θ it is possible to compute exactly the distribution $p(y|x,\theta)$ by marginalizing out hidden variables which can be further used in predicting class label for a new example $\mathbf{x}$. This conditional distribution takes the following form [9]:[2]

$$p(y_k = 1|x,\theta) = \frac{exp\{d_k\}_j (1 + exp\{c_j + (W_{\cdot j})^\top x + U_{jk}\})}{\sum_l exp\{d_l\}_j (1 + exp\{c_j + (W_{\cdot j})^\top x + U_{jl}\})}. \quad (3)$$

2.3 Learning

For given N training examples $\mathcal{D} = \{(x_n, y_n)\}$ to train the generative model we consider minimization of the negative log-likelihood:

$$\mathcal{L}(\theta) = -\sum_{n=1}^{N} \log p(x_n, y_n|\theta). \quad (4)$$

Since it is impossible to calculate exact gradient of (4), the minimization can be done by applying *Contrastive Divergence* [6], where the gradient is approximated using Gibbs sampling. See [9] for its simple extension for the ClassRBM.

To prevent the model from overfitting, additional regularization term can be added to the learning objective:

$$\mathcal{L}_\Omega(\theta) = \mathcal{L}(\theta) + \lambda \Omega(\theta), \quad (5)$$

where $\lambda > 0$ is the regularization coefficient, and $\Omega(\theta)$ is the regularization term.

In words, this form of the regularization is simply adding a penalty to the objective function which provides convenient fashion to control the model capacity. In general, in the context of RBM, one most widely used regularization technique is *weight decay* which controls the ℓ_2 norm of the parameters, and results in smooth models. On the other hand, it is advocated to obtain *sparse* models, where small amount of hidden units is active at the same time, by enforcing low probability of hidden units activation. In the next section we discuss the latter approach.

3 Sparse hidden units activation

In the following section we consider different choices of the regularization term $\Omega(\theta)$ which lead to different learning rules. For brevity, we use the following notation: activation of the j^{th} hidden unit $p_{jn} \stackrel{\Delta}{=} p(h_j = 1|x_n, y_n, \theta)$, mean activation $p_j \stackrel{\Delta}{=} \frac{1}{N}\sum_n p_{nj}$, and $s_j \stackrel{\Delta}{=} c_j + (W_{\cdot j})^\top x + U_{j\cdot} y$.

3.1 Cross-entropy

In [16] the regularization term is chosen to be the cross-entropy between the desired level of activation μ and the mean activation of the j^{th} unit:

$$\Omega_{CE}(\theta) = -\sum_j \Big(\mu \log p_j + (1-\mu)\log(1-p_j)\Big). \quad (6)$$

Then, the derivative wrt the parameter θ is the following:

$$\frac{\partial \Omega_{CE}(\theta)}{\partial \theta} = \sum_j (p_j - \mu) \frac{\partial s_j}{\partial \theta}. \quad (7)$$

Note that this regularization term can be also obtained from the Kullback-Leibler divergence between the desired and the current hidden unit activation distribution by dropping the entropy of the desired distribution which is constant.

[2] $\mathbf{A}_{i\cdot}$ denotes the i^{th} row of matrix $\mathbf{A}$, $\mathbf{A}_{\cdot j}$ denotes the j^{th} column of matrix $\mathbf{A}$, and A_{ij} is the element of matrix $\mathbf{A}$.

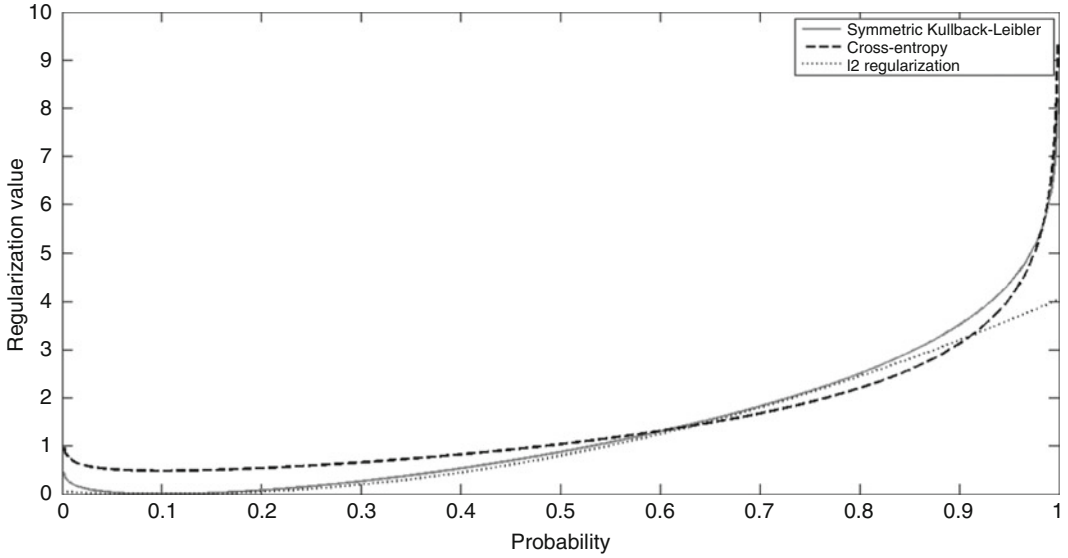


Fig. 1 Comparison of the regularization terms for $\mu = 0.1$. The cross entropy and the ℓ_2 regularization are scaled by factor 1.5 and 5, respectively, for easy of visual comparison

3.2 ℓ_2 norm

In [10] the regularization term is chosen to be the sum of squared differences between the fixed activation level and the mean hidden units activation:

$$\Omega_{L2}(\theta) = \sum_j (\mu - p_j)^2. \tag{8}$$

Then, the derivative wrt the parameter θ is as follows:

$$\frac{\partial \Omega_{L2}(\theta)}{\partial \theta} \propto \sum_j (p_j - \mu) \quad p_j(1 - p_j) \frac{\partial s_j}{\partial \theta}. \tag{9}$$

Notice that the obtained gradient of the regularization term can be seen as the gradient of the cross-entropy corrected by the variance of the hidden unit activation $p_j(1 - p_j)$.

3.3 Symmetric Kullback-Leibler

We propose to use a regularization term based on the symmetric Kullback-Leibler divergence between the desired and the actual distribution of the hidden unit activation:

$$\Omega_{KL}(\theta) = \sum_j (\mu - p_j) \left(\log \frac{\mu}{1 - \mu} - \log \frac{p_j}{1 - p_j} \right). \tag{10}$$

Then, the derivative wrt the parameter θ is as follows:

$$\frac{\partial \Omega_{KL}(\theta)}{\partial \theta} = \sum_j \left\{ p_j(1 - p_j) \left(\log \frac{p_j}{1 - p_j} - \log \frac{\mu}{1 - \mu} \right) + p_j - \mu \right\} \frac{\partial s_j}{\partial \theta}. \tag{11}$$

Notice that omitting in (8) the first term in the sum leads to the same solution as in (7). Thus, the first term can be seen as small corrections in regularization strength depended on the current distance from the desired activation level. Hence, the regularization term based on the symmetric Kullback-Leibler can be seen as a modification of the cross-entropy.

3.4 Remarks

Regularizers comparison. In fact, a closer look on the regularization terms (6), (8), and (10) provide a corollary that these regularizers behave in a very similar manner, i.e., they return similar values. These regularization terms are depicted in Figure 1. It can be noticed that there are small differences among them and thus we put a hypothesis that their influence on learning is almost indistinguishable. Nonetheless, the small dissimilarities may play an important role in the learning process. We verify this issue during experiments.

Optimization. The regularization terms in (6), (8), and (10) require summing over the whole training set which is troublesome. To alleviate this issue mini-batches can be used but this procedure results in biased estimates anyway.

Therefore, in [10], [11] it is advocated to update only the bias terms which directly control the hidden units activation level, and thus their sparsity. Further in the experiments, we penalize the biases $\mathbf{c}$ only.

4 Experiments

Datasets. We evaluate the presented regularization techniques on the image corpora, MNIST[3], and the document classification benchmark, 20-newsgroup dataset[4].

The MNIST dataset contains 60,000 training and 10,000 test images (28×28 pixels) of the ten hand-written digits (from 0 to 9). Of the training images, we set aside 10,000 for validation.

The 20-newsgroup dataset contains 8,500 training, 1,245 validation, and 6,497 test documents. We used 100 most frequent words describing a document for the binary inputs. The task is to classify a document to one of four classes (newsgroup meta-topics).

Learning details. In the experiment we used the ClassRBM and the following learning methods: learning without any regularization term, and using *Dropout* [8], cross-entropy regularization (6), ℓ_2 norm regularization (8), and symmetric Kullback-Leibler regularization (10). Additionally, in the case of the MNIST, we compared the considered methods with the Tikhonov-type regularization type 1 and type 2 [3], however, the reported results were obtained where RBM was used as a feature extractor for the logistic regression classifier. Similarly, in the context of 20-newsgroups, we compared the proposed learning method with the SVM fed up with the features learnt by RBM [14].

We performed learning using Constrastive Divergence on mini-batches of 10 examples for both datasets. We used 4900 hidden units for MNIST and 500 – 20-newsgroup. For the MNIST dataset the learning rate was equal 0.005 while for the 20-newsgroup – 0.001. In order to choose the regularization coefficient we performed the model selection using the validation set for symmetric Kullback-Leibler regularization (10) where $\lambda \in \{0.0001, 0.0005, 0.001, 0.005, 0.01\}$, and for the sparse regularization (6) and (8) – 1.5λ and 5λ, respectively. Additionally, for all sparse regularization terms for hidden units activation we set μ to $\{0.005, 0.05, 0.1, 0.3\}$. The number of iterations over the training set was determined using early stopping according to the validation set classification error, with a look ahead of 15 iterations.

Table 1 Detailed test results for MNIST. Best result in bold.

Method	Error [%]	Mean active
ClassRBM	3.42	60
Dropout	2.87	88
cross-entropy regularization (6)	**2.06**	29
ℓ_2 regularization (8)	2.18	31
symmetric Kullback-Leibler regularization (10)	2.17	29
Logistic reg. + Tikhonov-type regularization 1 [3]	*2.53*	-
Logistic reg. + Tikhonov-type regularization 2 [3]	*2.47*	-

Table 2 Detailed test results for 20-newsgroups. Best result in bold.

Method	Error [%]	Mean active
ClassRBM	20.5	146
Dropout	21.07	50
cross-entropy regularization (6)	19.98	86
ℓ_2 regularization (8)	19.84	68
symmetric Kullback-Leibler regularization (10)	**19.81**	74
SVM + RBM [14]	*18.95*	-

Evaluation methodology. We consider two evaluation metrics. First, we use the classification error (**Error**, expressed in %). Second, for trained RBM, we check the mean number of active units (**Mean active**). The classification error and the mean number of active units are evaluated on the test set only.

Results and Discussion. The results are presented in Table 1 and 2. First, we notice that the ClassRBM without any regularization performs worse than with regularization. Only exception is application of *Dropout* in the case of 20-newsgroup (see Table 2). The symmetric Kullback-Leibler as the sparse regularization term performs similarly to the cross-entropy and the ℓ_2 norm regularization on both considered datasets. Moreover, its application results in sparse solutions, i.e., mean activation of hidden units is the smallest among considered models on both datasets. In fact, it has to be noticed that the data in the 20-newsgroup are sparse itself (on average about 4 out of 100 words per document) and thus the problem becomes controversial whether any fashion of sparsity is useful. Nonetheless, the mean number of hidden units is higher for the ClassRBM than for the one with any regularization term, see Table 2, and the results (i.e. errors) show positive effect of applying any regularization term forcing sparse hidden units activation.

[3] http://yann.lecun.com/exdb/mnist/

[4] In the experiments we used the small version of the original dataset: http://www.cs.nyu.edu/~roweis/data.html.

5 Conclusions

We have presented new regularization term, called the symmetric Kullback-Leibler sparse regularization, which aims at obtaining small amount of active hidden units and hence sparse solution. The experiments showed that the application of the proposed schema allows to extract better discriminative features with sparser hidden activation, which resulted in decrease in the classification error comparing to ClassRBM or other methods. However, as we pointed out in the Section 3.4, differences among the considered regularizers are rather negligible and thus any of these regularizers can be used for sparse hidden unit activation. This corollary was supported by the carried out experiments.

References

1. Bengio, Y.: Learning Deep Architectures for AI. Foundations and Trends® in Machine Learning 2(1):1-127. (2009).
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer New York. (2006).
3. Cho, K., Ilin, A., & Raiko, T.: Tikhonov-Type regularization for Restricted Boltzmann Machines. In Artificial Neural Networks and Machine Learning (*ICANN 2012*). pp. 81-88. Springer Berlin Heidelberg. (2012).
4. Glorot, X., Bordes, A., & Bengio, Y.: Deep Sparse Rectifier Networks. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (*AISTATS 2011*). Journal of Machine Learning Research Workshop & Conference Proceedings 15:315-323. (2011).
5. Goh, H., Thome, N., & Cord, M.: Biasing restricted Boltzmann machines to manipulate latent selectivity and sparsity. In NIPS workshop on deep learning and unsupervised feature learning. (2010).
6. Hinton, G. E.: Training products of experts by minimizing contrastive divergence. Neural Comput 14:1771-1800. (2002)
7. Hinton, G. E.: A practical guide to training Restricted Boltzmann Machines. In Neural Networks: Tricks of the Trade. pp. 599-619. Springer Berlin Heidelberg. (2012).
8. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580. (2012).
9. Larochelle, H., & Bengio, Y.: Classification using discriminative restricted Boltzmann machines. In Proceedings of the 25th International Conference on Machine learning (*ICML 2008*). pp. 536-543. (2008, July).
10. Lee, H., Ekanadham, C., & Ng, A.: Sparse deep belief net model for visual area V2. In Advances in Neural Information Processing Systems (*NIPS 2007*). pp. 873-880. (2007).
11. Le, Q.V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., & Ng, A.: On optimization methods for deep learning. In Proceedings of the 28th International Conference on Machine Learning (*ICML 2011*). pp. 265-272. (2011).
12. Le Roux, N., & Bengio, Y.: Representational power of Restricted Boltzmann Machines and deep belief networks. Neural Computation 20(6):1631-1649. (2008).
13. Lewicki, M. S., & Sejnowski, T. J.: Learning overcomplete representations. Neural Computation 12(2):337-365. (2000).
14. Marlin, B.M., Swersky, K., Chen, B., & Freitas, N.D.: Inductive principles for restricted Boltzmann machine learning. In International Conference on Artificial Intelligence and Statistics (*ICML 2010*). pp. 509-516. (2010).
15. Martens, J., Chattopadhya, A., Pitassi, T., & Zemel, R.: On the Expressive Power of Restricted Boltzmann Machines. In Advances in Neural Information Processing Systems (*NIPS 2013*). pp. 2877-2885. (2013).
16. Nair, V., & Hinton, G. E.: 3D object recognition with deep belief nets. In Advances in Neural Information Processing Systems (*NIPS 2009*). pp. 1339-1347. (2009).
17. Nair, V., & Hinton, G. E.: Rectified linear units improve Restricted Boltzmann Machines. In Proceedings of the 27th International Conference on Machine Learning (*ICML 2010*). pp. 807-814. (2010).
18. Smolensky, P.: Information processing in dynamical systems: foundations of harmony theory. In: Parallel distributed processing: explorations in the microstructure of cognition, Vol. 1: Foundations, pp. 194281. MIT Press, Cambridge, MA, USA. (1986)

Accelerated learning for Restricted Boltzmann Machine with momentum term

Szymon Zaręba, Adam Gonczarek, Jakub M. Tomczak, and Jerzy Świątek

1 Introduction

Deep learning has recently become a field of interest due to its ability of automatic high-level features extraction [1]. Deep architectures are powerful models that achieve high performance on difficult pattern recognition problems, such as image analysis [2], motion tracking [3], speech recognition [4], and other application, e.g., collaborative filtering [5], text analysis [6].

Typically, building blocks of deep models are Restricted Boltzmann Machines (RBM). RBM are generative models with hidden variables which aim at modelling a distribution of visible variables. In the case of classification problems, RBM are used as standalone feature extractors, or as a parameter initialization for deeper models. However, when RBM are trained in an unsupervised fashion, there is no guarantee they provide discriminative features. To address this problem information about class label can be involved in the RBM, which leads to the Classification Restricted Boltzmann Machine (ClassRBM) [12].

Although the representational power of deep models, including RBM, is very tempting, their broad applicability was limited until recent past because of the difficulty of learning deep architectures. Deep learning became a topic of high interest thanks to the breakthrough learning algorithm called *Contrastive Divergence* [13]. The Contrastive Divergence allows to efficiently perform approximate stochastic gradient descent learning procedure. Since the Contrastive Divergence serves as an elemental learning procedure, different techniques can be applied in order to speed up the convergence or to obtain more reliable estimates. The most common approach is to modify the learning objective by adding an extra regularization term, e.g., ℓ_2 regularizer (weight decay) to keep the parameters below some threshold, or sparse regularizer to enforce sparse activation of hidden units [14, 15, 16]. Another regularization procedure is based on randomly dropping a subset of hidden units (or alternatively subset of weight parameters) during one updating epoch. This leads to techniques such as *dropout* or its extensions [17, 18, 19, 20]. A different approach is to apply optimization techniques, e.g., momentum method [15] or the Nesterov's accelerated gradient (Nesterov's momentum) [21].

In this paper, we aim at accelerating learning of RBM by applying the momentum term and the Nesterov's momentum. We formulate the following research questions:

Q1: Does the application of the momentum term or the Nesterov's momentum speed up the convergence of learning?

Q2: Does the application of the momentum term or the Nesterov's momentum increase the classification accuracy?

Q3: Does the Nesterov's momentum perform better than the momentum term?

In order to verify these issues we carry out experiments on the image dataset MNIST.

2 Classification using Restricted Boltzmann Machine

Let $\mathbf{x} \in \{0, 1\}^D$ be the D-dimensional vector of visible variables, $\mathbf{h} \in \{0, 1\}^M$ be the M-dimensional vector of hidden variables, and $\mathbf{y}$ be the vector coding the class label using 1-of-K coding scheme, i.e., $y_k = 1$ iff observation $\mathbf{x}$ belongs to the class k, where $k = 1, \ldots, K$. Joint dependency between observed and hidden variables, $(\mathbf{x}, \mathbf{y}, \mathbf{h})$, is described by the following energy function:

$$E(\mathbf{x}, \mathbf{y}, \mathbf{h}|\boldsymbol{\theta}) = -\mathbf{b}^T\mathbf{x} - \mathbf{c}^T\mathbf{h} - \mathbf{d}^T\mathbf{y} - \mathbf{h}^T\mathbf{W}\mathbf{x} - \mathbf{h}^T\mathbf{U}\mathbf{y} \quad (1)$$

where x_i, y_k and h_j are binary state of observed unit i, k-th entry in the vector coding class label, and the binary state of

S. Zaręba (✉) • A. Gonczarek • J.M. Tomczak • J. Świątek
Institute of Computer Science, Wroclaw University of Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: szymon.zareba@pwr.wroc.pl; adam.gonczarek@pwr.wroc.pl; jakub.tomczak@pwr.wroc.pl; jerzy.swiatek@pwr.wroc.pl

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_28,

hidden unit j, respectively. Further, b_i, d_k and c_j denote bias parameters associated with x_i, y_k and h_j, respectively. Finally, W_{ji} is the weight parameter modeling the relationship between x_i and h_j, whereas U_{jk} is the weight between h_j and y_k. For brevity, let $\boldsymbol{\theta} = \{\mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{W}, \mathbf{U}\}$ denotes the model parameters. ClassRBM defines the joint probability distribution over observed and hidden variables as follows:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{h}|\boldsymbol{\theta}) = \frac{1}{Z} e^{-E(\mathbf{x}, \mathbf{y}, \mathbf{h}|\boldsymbol{\theta})}, \tag{2}$$

where Z is the partition function obtained by summing out all possibles pairs of observed and hidden variables, $Z = \sum_{\mathbf{x}} \sum_{\mathbf{y}} \sum_{\mathbf{h}} e^{-E(\mathbf{x}, \mathbf{y}, \mathbf{h}|\boldsymbol{\theta})}$.

The inputs, hidden, and label variables are conditionally independent given the other variables, and the conditional probabilities can be written as follows:

$$p(x_i = 1|\mathbf{h}, \boldsymbol{\theta}) = \text{sigm}(b_i + \mathbf{h}^T \mathbf{W}_{\cdot i}) \tag{3}$$

$$p(h_j = 1|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = \text{sigm}(c_j + \mathbf{W}_{j\cdot}\mathbf{x} + U_{jl}) \tag{4}$$

where $\text{sigm}(\cdot)$ is the logistic sigmoid function, and $\mathbf{W}_{i\cdot}$, $\mathbf{W}_{\cdot j}$ denote rows and columns of matrix $\mathbf{W}$, respectively. Finally, l denotes index such that $l = \{k\colon y_k = 1\}$.

It turns out that calculation of the exact form of the conditional probability distribution $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ is tractable:

$$p(y_l = 1|\mathbf{x}, \boldsymbol{\theta}) = \frac{exp(d_l) \prod_{j=1}^{M} (1 + exp(c_j + \mathbf{W}_{j\cdot}\mathbf{x} + U_{jl}))}{\sum_{k=1}^{K} exp(d_k) \prod_{j=1}^{M} (1 + exp(c_j + \mathbf{W}_{j\cdot}\mathbf{x} + U_{jk}))}. \tag{5}$$

Further, in the paper, we refer learning joint distribution $p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})$ of ClassRBM to as *generative ClassRBM*, while conditional distribution $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ of ClassRBM – *discriminative ClassRBM*.

3 Learning

Let $\mathcal{D}$ be the training set containing N observation-label pairs, $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}$. In generative learning we aim at minimizing the following negative log-likelihood function:

$$\begin{aligned} \mathcal{L}_G(\boldsymbol{\theta}) &= -\sum_{n=1}^{N} \ln p(\mathbf{x}_n, \mathbf{y}_n|\boldsymbol{\theta}) \\ &= -\sum_{n=1}^{N} \ln p(\mathbf{y}_n|\mathbf{x}_n, \boldsymbol{\theta}) - \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\boldsymbol{\theta}) \end{aligned} \tag{6}$$

Therefore, the generative learning objective consists of two components, namely, the supervised learning objective $\sum_{n=1}^{N} \ln p(\mathbf{y}_n|\mathbf{x}_n, \boldsymbol{\theta})$, where we fit parameters to predict label given the observation, and the unsupervised objective $\sum_{n=1}^{N} \ln p(\mathbf{x}_n|\boldsymbol{\theta})$, which can be seen as a data-dependent regularizer.

Hence, omitting the last component in (6) leads to the following objective:

$$\mathcal{L}_D(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \ln p(\mathbf{y}_n|\mathbf{x}_n, \boldsymbol{\theta}). \tag{7}$$

Notice that we obtained the negative conditional log-likelihood which is the natural objective in discriminative learning.

4 Classical momentum term and Nesterov's momentum

Application of the classical momentum term to modify the search direction is a simple method for increasing the rate of convergence [15]. In general, the idea is to modify the update step of the gradient-based method by adding previous value of the gradient:

$$\mathbf{v}^{(new)} = \alpha \mathbf{v}^{(old)} - \eta \Delta(\boldsymbol{\theta}), \tag{8}$$

where $\Delta(\boldsymbol{\theta})$ is the step dependent on current parameters value, $\mathbf{v}$ denotes *velocity*, i.e., the accumulated change of the parameters, α is the momentum parameter determining the influence of the previous velocity included in calculating new velocity.

Recently, it has been shown that the modified version of the momentum term, which is based on the Nesterov's accelerated gradient technique, gives even better results [21]. The Nesterov's accelerated gradient (henceforth called *Nesterov's momentum*) is based on the idea to include velocity in the objective function which follows the gradient calculation. Such approach allows to avoid instabilities and respond to inappropriately chosen direction. The Nesterov's momentum is calculated as follows:

$$\mathbf{v}^{(new)} = \alpha \mathbf{v}^{(old)} - \eta \Delta(\boldsymbol{\theta} + \alpha \mathbf{v}^{(old)}). \tag{9}$$

4.1 Learning generative ClassRBM

In ClassRBM learning one optimizes the objective function (6) with respect to the parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{U}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}$. However, gradient-based optimization methods cannot be directly applied because exact gradient calculation is intractable. Fortunately, we can adopt *Constrastive Divergence*

ALGORITHM 1: Contrastive Divergence algorithm for generative ClassRBM

Input: data $\mathcal{D}$, learning rate η, momentum parameter α

Output: parameters $\boldsymbol{\theta}$

for each example $(\mathbf{x}, \mathbf{y})$ **do**

1. Generate samples using Gibbs sampling:
   Set $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}) := (\mathbf{x}, \mathbf{y})$.
   Calculate probabilities $\hat{\mathbf{h}}^{(0)}$, where $\hat{h}_j^{(0)} := p(h_j = 1 | \mathbf{x}^{(0)}, \mathbf{y}^{(0)}, \boldsymbol{\theta}^{temp})$.
   Generate sample $\bar{\mathbf{h}}^{(0)}$ for given probabilities $\hat{\mathbf{h}}^{(0)}$.

   for $t = 0$ to $\tau - 1$ **do**
      Calculate probabilities $\hat{\mathbf{x}}^{(t+1)}$, where $\hat{x}_i^{(t+1)} := p(x_i = 1 | \bar{\mathbf{h}}^{(t)}, \boldsymbol{\theta}^{temp})$.
      Generate sample $\bar{\mathbf{x}}^{(t+1)}$ for given probabilities $\hat{\mathbf{x}}^{(t+1)}$.
      Calculate probabilities $\hat{\mathbf{y}}^{(t+1)}$, where $\hat{y}_k^{(t+1)} := p(y_k = 1 | \bar{\mathbf{x}}^{(t)}, \boldsymbol{\theta}^{temp})$.
      Generate sample $\bar{\mathbf{y}}^{(t+1)}$ for given probabilities $\hat{\mathbf{y}}^{(t+1)}$.
      Calculate probabilities $\hat{\mathbf{h}}^{(t+1)}$, where
      $\hat{h}_j^{(t+1)} := p(h_j = 1 | \hat{\mathbf{x}}^{(t+1)}, \hat{\mathbf{y}}^{(t+1)}, \boldsymbol{\theta}^{temp})$.
      Generate sample $\bar{\mathbf{h}}^{(t+1)}$ for given probabilities $\hat{\mathbf{h}}^{(t+1)}$.
   end for

2. Compute step $\Delta(\boldsymbol{\theta}^{temp})$. In particular:
   $\Delta\mathbf{b} := \mathbf{x}^{(0)} - \bar{\mathbf{x}}^{(\tau)}$
   $\Delta\mathbf{c} := \hat{\mathbf{h}}^{(0)} - \hat{\mathbf{h}}^{(\tau)}$
   $\Delta\mathbf{d} := \mathbf{y}^{(0)} - \bar{\mathbf{y}}^{(\tau)}$
   $\Delta\mathbf{W} := \hat{\mathbf{h}}^{(0)}\mathbf{x}^{(0)\mathrm{T}} - \hat{\mathbf{h}}^{(\tau)}\bar{\mathbf{x}}^{(\tau)\mathrm{T}}$
   $\Delta\mathbf{U} := \hat{\mathbf{h}}^{(0)}\mathbf{y}^{(0)\mathrm{T}} - \hat{\mathbf{h}}^{(\tau)}\bar{\mathbf{y}}^{(\tau)\mathrm{T}}$
3. Update momentum term $\mathbf{v}^{(new)} := \alpha\mathbf{v}^{(old)} - \eta\Delta(\boldsymbol{\theta}^{temp})$ and set $\mathbf{v}^{(old)} := \mathbf{v}^{(new)}$.
4. Update parameters:
   $\boldsymbol{\theta} := \boldsymbol{\theta} + \mathbf{v}^{(new)}$
   and set $\boldsymbol{\theta}^{temp} := \boldsymbol{\theta}$ in case of the classical momentum (8) or
   $\boldsymbol{\theta}^{temp} := \boldsymbol{\theta} + \alpha\mathbf{v}^{(old)}$ in case of the Nesterov's momentum (9).

end for

algorithm which approximates exact gradient using sampling methods.

In fact, the Contrastive Divergence algorithm aims at minimizing the difference between two *Kullback-Leibler divergences* [14]:

$$\mathrm{KL}(Q||P) - \mathrm{KL}(P_\tau||P) \tag{10}$$

where Q is the empirical probability distribution, and P_τ is the probability distribution over visible variables after τ steps of the Gibbs sampler, P is the true distribution.

The approximation error vanishes for $\tau \to \infty$, i.e., when P_τ becomes stationary distribution, and $\mathrm{KL}(P_\tau \mid P) = 0$ with probability 1. Therefore, it would be beneficial to choose a large value of τ, however, it has been noticed that choosing $\tau = 1$ works well in practice [1]. The procedure of the Contrastive Divergence for the generative ClassRBM is presented in Algorithm 1.

4.2 Learning discriminative ClassRBM

In discriminative learning we need to minimize the objective (7) wrt the parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{U}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}$. Unlike the generative model, here the gradient can be calculated analytically. Hence, the parameters of the discriminative ClassRBM can be determined using the stochastic gradient descent algorithm. The learning procedure for the discriminative case is presented in Algorithm 2.

ALGORITHM 2: Stochastic gradient algorithm for discriminative ClassRBM

Input: data $\mathcal{D}$, learning rate η, momentum parameter α
Output: parameters $\boldsymbol{\theta}$
for each example $(\mathbf{x}, \mathbf{y})$ **do**

1. Compute step $\Delta(\boldsymbol{\theta}^{temp})$. In particular:
   Set $l := \{k : y_k = 1\}$.
   Calculate probabilities $\hat{\mathbf{y}}$, where $\hat{y}_k := p(y_k = 1|\mathbf{x}, \boldsymbol{\theta}^{temp})$.
   Calculate $\boldsymbol{\sigma}$, where $\sigma_{jk} = \text{sigm}(c_j^{temp} + \mathbf{W}_{j\cdot}^{temp}\mathbf{x} + U_{jk}^{temp})$.
   $\Delta\mathbf{c} := \sum_{k=1}^{K} \hat{y}_k \boldsymbol{\sigma}_{\cdot k} - \hat{y}_l \mathbf{1}$
   $\Delta\mathbf{d} := \hat{\mathbf{y}} - \mathbf{y}$
   $\Delta\mathbf{W} := \sum_{k=1}^{K} \hat{y}_k \boldsymbol{\sigma}_{\cdot k}\mathbf{x}^T - \boldsymbol{\sigma}_{\cdot l}\mathbf{x}^{\mathrm{T}}$
   $\Delta\mathbf{U} := \boldsymbol{\sigma}_{\cdot l}(\hat{\mathbf{y}} - \mathbf{1})^{\mathrm{T}}$
2. Update momentum term $\mathbf{v}^{(new)} := \alpha\mathbf{v}^{(old)} - \eta\Delta(\boldsymbol{\theta}^{temp})$ and set $\mathbf{v}^{(old)} := \mathbf{v}^{(new)}$.
3. Update parameters:
   $\boldsymbol{\theta} := \boldsymbol{\theta} + \mathbf{v}^{(new)}$
   and set $\boldsymbol{\theta}^{temp} := \boldsymbol{\theta}$ in case of the classical momentum (8) or
   $\boldsymbol{\theta}^{temp} := \boldsymbol{\theta} + \alpha\mathbf{v}^{(old)}$ in case of the Nesterov's momentum (9).

end for

5 Experiments

5.1 Details

Dataset. We evaluate the presented learning techniques on the image corpora MNIST[1]. The MNIST dataset contains 50,000 training, 10,000 validation, and 10,000 test images (28×28 pixels) of ten hand-written digits (from 0 to 9).

Learning details. We performed learning using Constrastive Divergence with mini-batches of 10 examples. Both generative and disriminative ClassRBM were trained with and without momentum and the Nesterov's momentum. The learning rate was set to $\eta = 0.005$ and $\eta = 0.05$ for generative and disriminative ClassRBM, respectively. The momentum parameter was set to $\alpha \in \{0.5, 0.9\}$. The number of iterations over the training set was determined using early stopping according to the validation set classification error, with a look ahead of 15 iterations. The experiment was repeated five times.

Evaluation methodology. We use the classification accuracy as an evaluation metric. We compare the learning procedure using Contrastive Divergence (**CD**), stochastic gradient descent (**SGD**) and learning with the classical momentum term (**CM**) and the Nesterov's momentum (**N**). The classification accuracy (expressed in %) is calculated on test set.

5.2 Results and Discussion

The results (mean values and standard deviations over five repetitions) for the generative ClassRBM are presented in Table 1 while for the discriminative ClassRBM in Table 2. The mean classification accuracy for generative ClassRBM is presented in Figure 1 and for disriminative ClassRBM – in Figure 2.[2]

We notice that the application of the momentum term and the Nesterov's momentum indeed increases the speed of the convergence (see Figure 1 and 2). This phenomenon is especially noticeable in the case of the disriminative ClassRBM and $\alpha = 0.9$ (Figure 2). The Nesterov's momentum performs slightly better than the momentum term in terms of the speed of convergence only during the first epochs (see Figure 2).

On the other hand, for larger number of hidden units, i.e., for M larger than 400 in the case of the generative ClassRBM and M larger than 100 for the discriminative ClassRBM, application of the momentum term and the Nesterov's momentum resulted in the increase of the classification accuracy and more

[1] http://yann.lecun.com/exdb/mnist/

[2] In both cases the number of hidden units was equal 900.

Table 1 Classification accuracy for generative ClassRBM with the Contrastive Divergence (**CD**) and the classical momentum term (**CM**) and the Nesterov's momentum (**N**).

	CD		CM $\alpha = 0.5$		CM $\alpha = 0.9$		N $\alpha = 0.5$		N $\alpha = 0.9$	
Hidden units	mean	std	mean	std	mean	std	mean	std	mean	std
9	54.79	4.15	60.79	4.47	63.60	1.33	63.69	2.09	63.04	3.13
25	81.21	0.33	80.32	1.26	82.21	1.01	80.51	0.48	81.76	0.60
100	90.81	0.36	90.74	0.32	90.70	0.48	90.67	0.30	91.11	0.50
400	94.02	0.29	94.34	0.27	94.51	0.35	94.54	0.20	94.61	0.23
900	95.08	0.23	95.62	0.08	95.53	0.42	95.45	0.19	95.25	0.23

Table 2 Classification accuracy for discriminative ClassRBM with the stochastic gradient descent (**SGD**) and the classical momentum term (**CM**) and the Nesterov's momentum (**N**).

	SGD		CM $\alpha = 0.5$		CM $\alpha = 0.9$		N $\alpha = 0.5$		N $\alpha = 0.9$	
Hidden units	mean	std	mean	std	mean	std	mean	std	mean	std
9	92.39	0.14	92.34	0.26	91.91	0.32	92.30	0.46	91.41	0.35
25	95.74	0.32	95.49	0.36	95.05	0.14	95.51	0.29	95.15	0.09
100	97.36	0.09	97.35	0.17	97.46	0.05	97.40	0.06	97.67	0.11
400	97.63	0.07	97.67	0.02	97.96	0.05	97.77	0.10	97.86	0.06
900	97.52	0.11	97.72	0.01	97.85	0.14	97.73	0.04	97.82	0.02

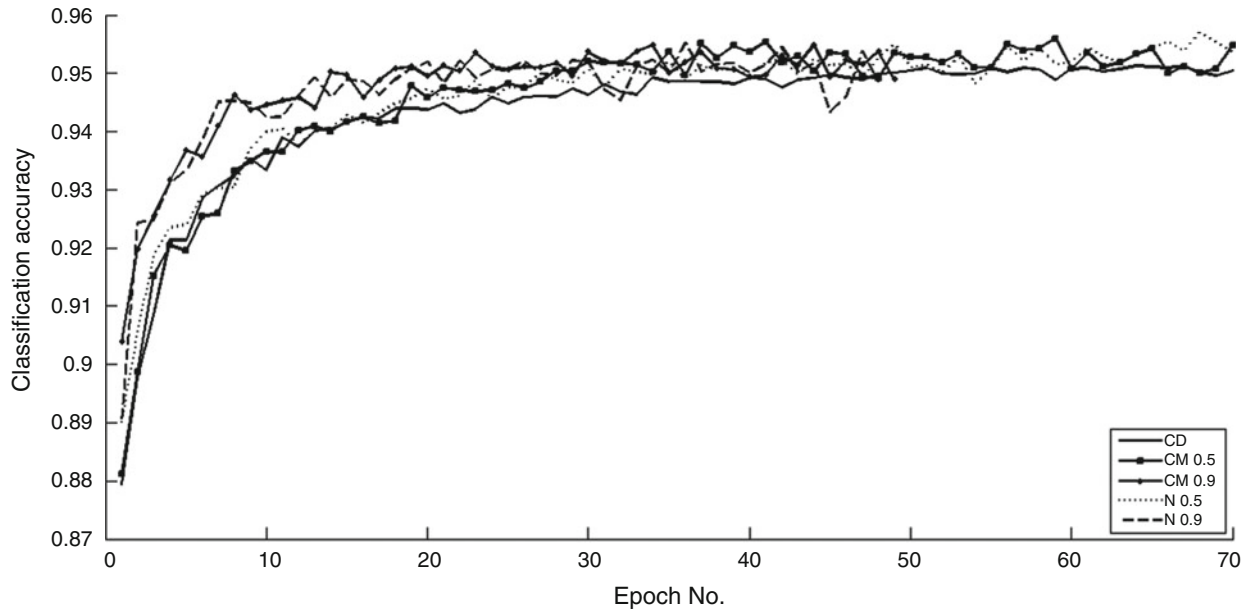


Fig. 1 Convergence of the generative ClassRBM with the Contrastive Divergence (**CD**), and with the classical momentum term (**CM**) and Nesterov's momentum (**N**) with $\alpha = 0.5$ and $\alpha = 0.9$ for the classification accuracy measured on the test set.

stable outcomes (smaller standard deviations). However, the Nesterov's momentum is slightly more robust than the momentum in terms of the standard deviations (see Table 1 and 2). This effect can be explained in the following way. The Nesterov's momentum calculates a partial update of parameters first, and then computes gradient of the objective wrt to partially updated parameters. Such procedure allows to change the velocity quicker if undesirable increase in the objective occurs. Our result is another empirical justification of this phenomenon, previously reported in [21].

6 Conclusions

In this paper, we have presented the accelerated learning of the Restricted Boltzmann Machine with the classical momentum and the Nesterov's momentum term. We have applied the outlined learning procedure to generative and discriminative Classification Restricted Boltzmann Machine. In order to evaluate the approach and verify stated research questions we have performed experiments using MNIST image corpora

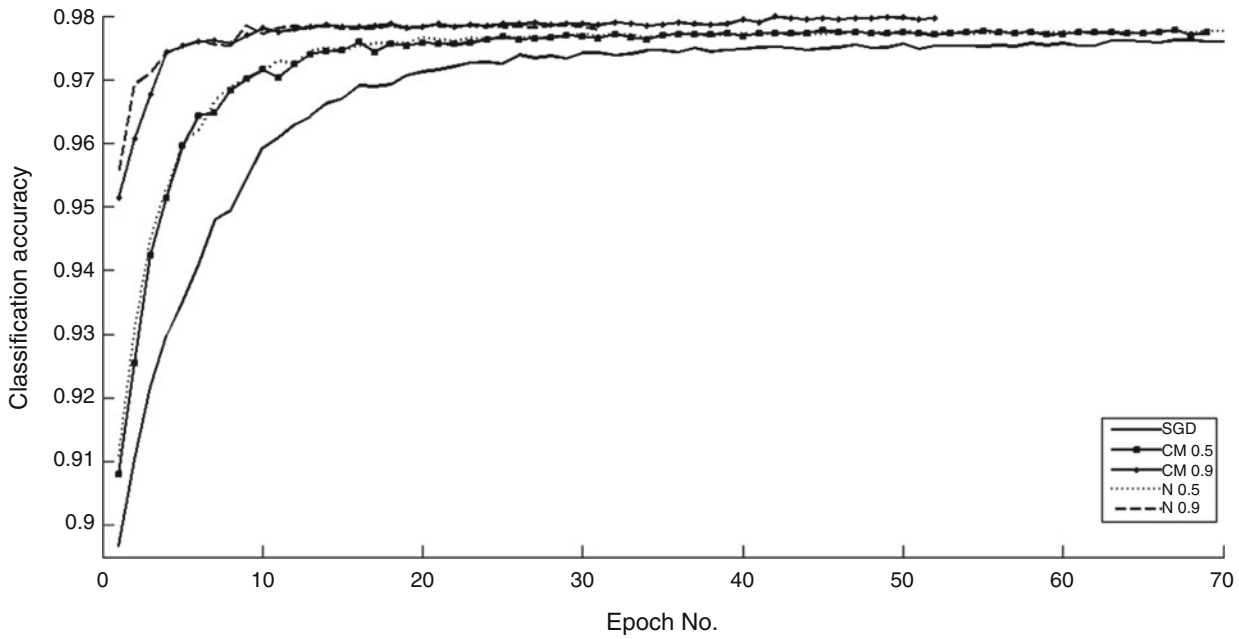


Fig. 2 Convergence of the discriminative ClassRBM with the stochastic gradient descent (**SGD**), and with the classical momentum term (**CM**) and Nesterov's momentum (**N**) with $\alpha = 0.5$ and $\alpha = 0.9$ for the classification accuracy measured on the test set.

for different number of hidden units. The obtained results shows that the application of the momentum term and the Nesterov's momentum indeed accelerates convergence of learning and increase the classification accuracy. However, our comparative analysis does not indicate the superiority of the Nesterov's momentum over the momentum term; these two techniques behave alike.

References

1. Bengio, Y.: Learning deep architectures for AI. Foundations and Trends in Machine Learning **2**(1) (2009) 1–127
2. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786) (2006) 504–507
3. Taylor, G.W., Hinton, G.E., Roweis, S.T.: Modeling human motion using binary latent variables. In Schölkopf, B., Platt, J.C., Hoffman, T., eds.: NIPS, MIT Press (2006) 1345–1352
4. Mohamed, A.R., Hinton, G.E.: Phone recognition using restricted boltzmann machines. In: ICASSP, IEEE (2010) 4354–4357
5. Salakhutdinov, R., Mnih, A., Hinton, G.E.: Restricted boltzmann machines for collaborative filtering. In Ghahramani, Z., ed.: ICML. Volume 227 of ACM International Conference Proceeding Series., ACM (2007) 791–798
6. Salakhutdinov, R., Hinton, G.E.: Replicated softmax: an undirected topic model. In Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A., eds.: NIPS, Curran Associates, Inc. (2009) 1607–1614
7. Neapolitan, R.E.: Probabilistic reasoning in expert systems - theory and algorithms. Wiley (1990)
8. Pearl, J.: Probabilistic reasoning in intelligent systems - networks of plausible inference. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann (1989)
9. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences of the United States of America **79**(8) (1982) 2554–2558
10. Hopfield, J.J.: The effectiveness of neural computing. In: IFIP Congress. (1989) 503–507
11. Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for Boltzmann Machines. Cognitive Science **9**(1) (1985) 147–169
12. Larochelle, H., Bengio, Y.: Classification using discriminative restricted boltzmann machines. In Cohen, W.W., McCallum, A., Roweis, S.T., eds.: ICML. Volume 307 of ACM International Conference Proceeding Series., ACM (2008) 536–543
13. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Computation **14**(8) (2002) 1771–1800
14. Fischer, A., Igel, C.: An introduction to Restricted Boltzmann Machines. In Álvarez, L., Mejail, M., Déniz, L.G., Jacobo, J.C., eds.: CIARP. Volume 7441 of Lecture Notes in Computer Science., Springer (2012) 14–36
15. Hinton, G.E.: A practical guide to training restricted boltzmann machines. In: Neural Networks: Tricks of the Trade (2nd ed.). (2012) 599–619
16. Swersky, K., Chen, B., Marlin, B.M., de Freitas, N.: A tutorial on stochastic approximation algorithms for training restricted boltzmann machines and deep belief nets. In: ITA, IEEE (2010) 80–89
17. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Improving neural networks by preventing co-adaptation of feature detectors. CoRR **abs/1207.0580** (2012)
18. Wager, S., Wang, S., Liang, P.: Dropout training as adaptive regularization. CoRR **abs/1307.1493** (2013)
19. Wan, L., Zeiler, M.D., Zhang, S., LeCun, Y., Fergus, R.: Regularization of neural networks using dropconnect. In: ICML (3). (2013) 1058–1066
20. Wang, S., Manning, C.D.: Fast dropout training. In: ICML (2). (2013) 118–126
21. Sutskever, I., Martens, J., Dahl, G.E., Hinton, G.E.: On the importance of initialization and momentum in deep learning. In: ICML (3). Volume 28 of JMLR Proceedings., JMLR.org (2013) 1139–1147

Optimizing Interface Area of Percolated Domains in Two Dimensional Binary Compound: Artificial Neural Network Modeling on Monte Carlo Experiments

Yongyut Laosiritaworn and Wimalin Laosiritaworn

1 Introduction

In compound materials, interfaces between different materials is known to yield important material behavior. This is as when interface changes, the associate interface free energy also changes and so do thermodynamic properties, e.g. levels of energies being exchanged between two sides of the interface [1], or the energy conversion efficiencies in some excitonic solar cell [2]. Therefore, the understanding of how to optimize the interface area by controlling morphologies of the material is vital.

In the case of binary compounds, each type of the compound can form braches of same atom leading to many enhanced properties, e.g. hardness of materials [3]. Note that in this binary compound system, if the cohesive interaction dominates, a state of material where same type of atoms inside gets connected (the ordered state) presents at low temperature. However, at high temperature, the ordered state changes to disordered state due to thermal fluctuation and some properties become deteriorated. On the other hand, if the adhesive interaction is of preference, the ordered state where neighboring atoms are different is preferred at low temperatures. Similarly, this ordered state changes to disordered state due to thermal fluctuation. Therefore, the competition between cohesive and adhesive interaction as well as the system temperature and size induces various complicate domain interfaces, where understanding of how to optimize the domain interface is useful for enhancing some material properties. For instance, the domains that percolates from and to electrodes, see Fig. 1, with large interface area is useful for enhancing efficiencies in polymer solar cell [4].

Nevertheless, the study of this mixed adhesive and cohesive interaction effect on interface problem (especially for the percolated domain) is not trivial due to the complex kinetic diffusion of the atoms at the interface. Furthermore, since cohesion and adhesion among different types of materials are usually temperature and system size dependent, enormous patterns of interface morphologies are possible at the thermodynamic limit, i.e. the system size approaches infinity. Therefore, to enhance the fundamental understanding, large number of experimental investigations should be carried out which is not possible when resources are limited. Consequently, numerical simulation should be firstly taken as a first guidance to the experiments. Therefore, in this work, Monte Carlo simulations were used to investigate behavior the self-soluble binary AB compounds. The two dimensional Ising model, with both the cohesive interaction (which prefers same atom connecting) and adhesive interaction (which prefer different atom connecting), were considered in representing the interaction among pairs of A and B atom. In the simulation, the Kawasaki algorithm [5] was used to update the system where the number of each kind (A or B specie) was conserved. The thermal equilibrium observables were taken in terms of number of percolated domains and their interface area as function of temperature, system size and the adhesive interaction strength. In addition, to develop database of how the dependent domain properties relate to the independent parameters, the artificial neural network was used to formulate relationship between input and output parameters, with an aim to another step fulfilling understanding in the topic. All these calculation techniques, the obtained results, and the discussion are elaborated in details in the following sections.

Y. Laosiritaworn (✉)
Department of Physics and Materials Science, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand
e-mail: yongyut_laosiritaworn@yahoo.com

W. Laosiritaworn
Department of Industrial Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand
e-mail: wimalin@hotmail.com

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_29,

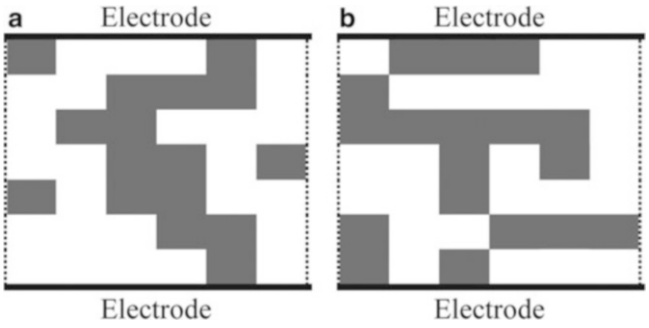


Fig. 1 Example of domain of the same atom arrangement where in (a) there are three percolating clusters linking between the two electrodes whereas in (b) there is none. Note that the grey and the white dots refers to A and B types in the binary compound.

2 Background and Methodologies

2.1 Monte Carlo Simulation and Ising Model

Monte Carlo simulation is a stochastic method which uses random numbers to find the most probable outcomes from associated probability of the system [6]. The Monte Carlo can be used to solve many problems in various fields ranging from finance [7] to physics [6],[8]. On the other hand, the Ising model is a mathematical model inspired by simple ferromagnetic, where the system energy or Hamiltonian is arisen from interaction among members of the Ising system, called spins. Each Ising spins can be only one of the two possible discrete states and interacts with other spins with some interaction strength. The Ising model has been used to study many applicable problems, e.g. the lattice gas [9], binary alloy [10], or neurons in the brain [11].

In this work, the thermal equilibrium Monte Carlo simulation was applied on the investigation of binary compound system, where the compound materials consist of A and B species (atoms) respectively. The atoms A and B were assumed to interact among themselves via the binary alloy like Hamiltonian i.e. [12]

$$H = -\sum_{\langle ij \rangle} \left(J_{co}\delta_{S_iS_j} + J_{ad}\left(1 - \delta_{S_iS_j}\right)\right), \qquad (1)$$

where J_{co} and J_{ad} refers to the cohesive interaction (same neighboring atoms preferred) and the adhesive interaction (different neighboring atoms preferred). In Eq. (1), the Ising spins $S_i = \pm 1$ refer to the A or B atoms on site i of the lattice, the notation $< ij >$ indicates that only nearest neighbor pairs are considered in the sum, and δ is the kronecker delta function (where $\delta_{S_iS_j} = 1$ if $S_i = S_j$ and zero otherwise). Therefore, the interaction J_{co} is for pairs of same spin direction and J_{ad} is for pairs of opposite direction spins. Note that, in Eq. (1), if $J_{ad} = J_{co}$, the Hamiltonian reduces to $H = -\sum_{\langle ij \rangle} J_{co}$ which is a constant, causing A and B atoms to randomly align. Furthermore, with Hamiltonian in Eq. (1), one can vary the cohesive interaction J_{co} and adhesive interaction J_{ad} in mimicking real systems.

In preparing the system, the considered binary compound was set on two-dimensional square lattice having $N = L \times L$ atoms (for both A and B atoms with 50:50 concentration). In this work, L was ranged from 20 to 100 with a step of 20. The cohesive interaction J_{co} was used as a unit of energy, so unit of the energy and temperature changes to J_{co} and J_{co}/k_B (where k_B is the Boltzmann's constant) respectively. Then, adhesive to cohesive interaction ratio J_{ad}/J_{co} was varied from -1.0 to 0.5 with a step of 0.5. The temperature T was varied from 0.05 to 3.30 J_{co}/k_B with a step of 0.05 J_{co}/k_B. In updating the system, the Kawasaki algorithm [5] was considered, i.e. neighboring atoms exchange with a probability of

$$P = \exp\left[-\frac{\Delta H}{k_{\mathrm{B}}T}\right] \qquad (2)$$

where ΔH is the energy difference due to exchange. In details, a pair of neighboring (different) atom is chosen and exchanged if a uniform random number r in the range [0,1) follows $r \leq P$, otherwise the exchange is discarded. In this simulation work, the system was initially set to clustered pattern having all A atoms in one cluster and the other B atoms in another cluster. The unit for time used in the simulation was defined in term of 1 Monte Carlo step (mcs) which is N trials of atom exchanging, either successful or unsuccessful. In each simulation, for particular sets of $\{J_{ad}/J_{co}, L, T\}$, the first 500 mcs were discarded for equilibration and the next 20,000 mcs were used to find the average number of percolated domains (the domains that span from one boundary to the opposite boundary) and its interface area (defined as number different atoms staying adjacent). Therefore, based on this Monte Carlo simulation, the process of cluster (of same spin) forming and decaying was investigated with a purpose to model the domain of the same spin behavior in two dimensional binary compound.

2.2 Artificial Neural Network

The artificial neural network (ANN) is a data mining technique which model relationship between inputs and the outputs via the 'experience' analysis. Specifically, the ANN consists of processing neurons connected together, where in multilayer perceptron [13, 14] neurons are placed in the input layer, hidden layers and the output layer. In the ANN, the network is formed and trained to realize input and output relationship by varying weights between pairs of neurons. During the training, sets of input-output are passed

to the network while the weights are adjusted to minimize the error between the predicted outputs and the targeted outputs. In this work, the weight adjustment took Back Propagation (BP) learning algorithm [14]. In this BP algorithm, inputs are firstly supplied to the neurons in the input layer, where the weighted sum $x_j = \sum_i k_j w_{ij}$ being calculated from all neural i in the current layer, is passed to the neuron j in the next layer. In the equation, k_i is the input to the neuron i, and w_{ij} is the weight. Next, the sigmoid transfer function $g(x_j) = \frac{1}{(1+e^{-x_j})}$ is applied on x_j, and this $g(x_j)$ becomes the input to neuron j. This so called forward pass procedure repeats for all neurons in the input and hidden layers. After that, the BP performs a backward pass, where an error is calculated to adjust the weight for each neuron. The error takes the form $\delta_j = (t_j - a_j)g'(x_j)$ for the output layer while in the hidden layer the error takes the form $\delta_j = \left[\sum_k \delta_k w_{kj}\right] g'(x_j)$. In these error equations, t_j is the target value for neuron j, a_j is the output value for neuron j, $g'(x)$ is the derivative of the sigmoid function g, and x_j is weighted sum of inputs (from all neuron i in the previous layer) to neuron j in the next layer. Then, the weight adjustment is calculated from $\Delta w_{ji} = \eta \delta_j a_i$ where η is the learning rate. These forward and backward processes repeat with new input sets until stopping criteria are met. Details of how to apply the ANN to other materials science problems can be found elsewhere, e.g. consider [15].

3 Results and Discussion

From Monte Carlo simulations, the competition between the same atom and different atom get connected on the domain characteristic was obtained, with an emphasis on the percolating domains. To illustrate this, the main effect plot of the number of percolated cluster ($N_{cluster}$) and interface of the percolated cluster (*interface*, the number of pairs of different atoms on the percolated cluster interface normalized by N) were drawn and presented as in Fig. 2. Note that to calculate each data point in the main effect plot, only one independent variable (the one being in investigation) is kept fixed while the others are varied, where the main effect plot can be extracted from the mean (average) of dependent variables obtained from the varied independent variables.

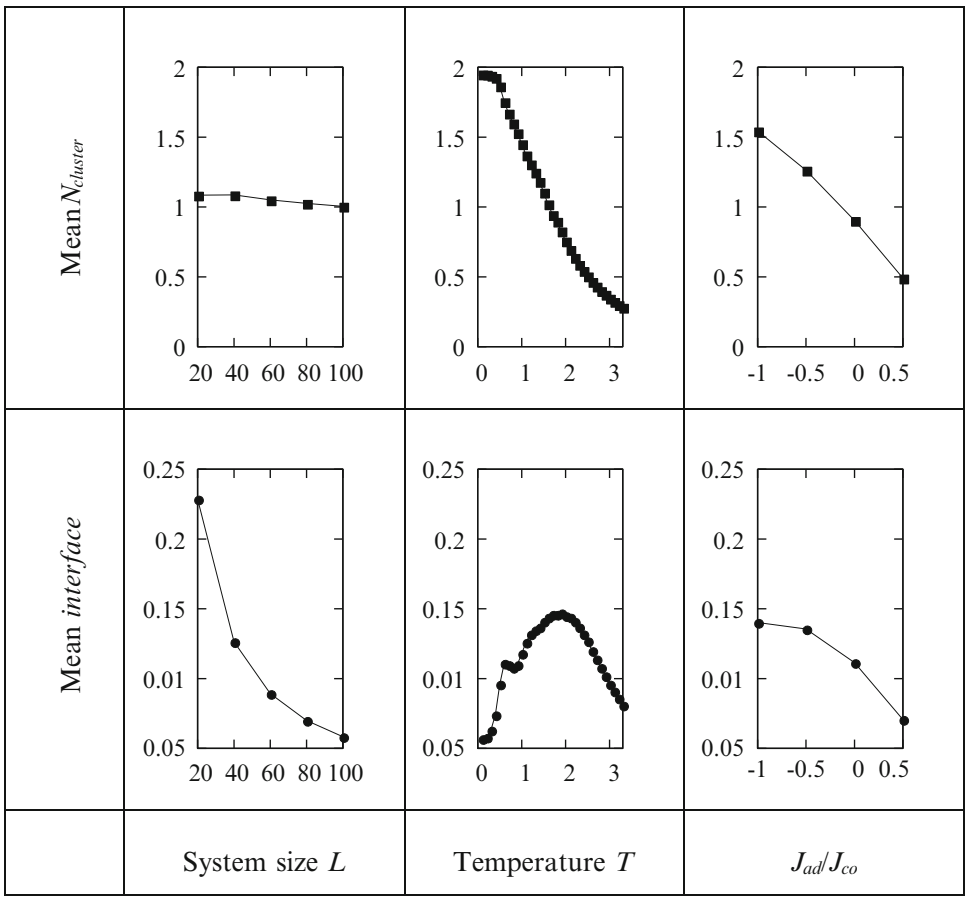


Fig. 2 Main effect plot of $N_{cluster}$ (top row) and interface (bottom row) with varying system size L, temperature T and the ratio J_{ad}/J_{co}.

Table 1 I Input and output data used in ANN training

Symbol	Description	Type	Scaling range	Range	Scaling factor
L	System size	Input	[-1,1]	[20,100]	0.025
T	Temperature	Input	[-1,1]	[0.05,3.30]	0.615385
J_{ad}/J_{co}	Ratio of adhesive to cohesive interaction	Input	[-1,1]	[-1,0.5]	1.333333
$N_{cluster}$	Number of percolated cluster	Output	[0,1]	[0.0001,2]	0.500025
interface	Interface area of all percolated cluster	Output	[0,1]	[0.00002,0.3124]	3.201226

In Fig. 2, the main effect plots of $N_{cluster}$ and *interface* as functions of the system size L, the temperature T and the adhesive to cohesive interaction ratio J_{ad}/J_{co} are presented. It is found that independent parameters (L,T,J_{ad}/J_{co}) have influences on the results ($N_{cluster}$,*interface*). Specifically, although the system size L has only small effect on $N_{cluster}$, it has strong effect on the interface. These results tell that although $N_{cluster}$ does not change much with L, the *interface* changes a lot but only for the L between 20 and 60. Therefore, it could be implied that the results for L larger than 100 are not significantly different. However, for the temperature effect, $N_{cluster}$ is close to 2 at small temperature, which means that there are only 2 big percolated domains (for A and B domains). Nevertheless, with increasing T, the cluster of the same atom reduces in size since smaller cluster forms. This therefore reduces $N_{cluster}$. For the *interface*, at low temperatures, the *interface* is rather small since there are just 2 big percolated domains. However, on increasing T, the 2 big percolated clusters reduce in size where its interface region changes to less smooth (become rougher) and this enhances the interface area. On further increasing the temperature, many small domains occur where $N_{cluster}$ ceases, so *interface* drops as a consequent. Then, for the effect of J_{ad}/J_{co} ratio, the increase of the adhesive interaction certainly decrease the same atom domain so both $N_{cluster}$ and *interface* drop as results.

Next, to create an extensive database for future use, e.g. for L, T and J_{ad}/J_{co} not considered in the simulation, the ANN was used to construct {L,T,J_{ad}/J_{co}} and {$N_{cluster}$,*interface*} relationship. This was firstly done by scaling inputs and outputs into appropriate ranges. Then, the inputs (which are L, T and J_{ad}/J_{co}) and outputs (which are $N_{cluster}$ and *interface*) were fed to the ANN. However, before feeding the data to the ANN, the data has to be scaled in an appropriate range. Therefore, input data were preprocessed by scaling into the range between -1 to 1 and output data was scaled to the range of 0 to 1. This process was done by using scaling factors (SF). The SF is calculated from (SR_{max}-SR_{min})/(X_{max}-X_{min}) where SR_{max} is the upper scaling range limit, SR_{min} is the lower scaling range limit, X_{max} is the maximum actual value, and X_{min} is the minimum actual value. Then the preprocessed value is calculated from $SR_{min} + (X\text{-}X_{min}) \times SF$, where X is the actual numeric value. The used SF and other important parameters are shown in Table 1.

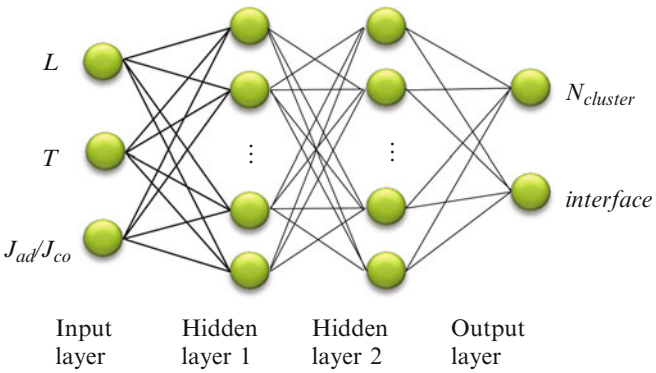


Fig. 3 The schematic diagram of the ANN network, with best network as 3-16-20-2.

The ANN network including number of hidden layers and hidden nodes were searched to determine the appropriate setting. The search was carried out for up to 2 hidden layers and up to 20 nodes in each hidden layer. The results of best network are 3-16-20-2, which these numbers represent number of neuron in input layer, in first and second hidden layer, and in output layer respectively, i.e. see Fig. 3.

In the ANN training, the dataset were separated into 3 sets for training, validating, and testing with the percentage of approximately 68 %, 16 %, and 16 % respectively. Training data were used to train the network, validating data were used to prevent overtraining, and testing data were used as 'unseen' dataset to test the network accuracy. As a result, 1320 records were separated to 898, 211, and 211 records. Afterwards, the best network obtained previously was used to train ANN, where the ANN accuracy was judged on mean absolute error $\mathrm{MAE} = \frac{\sum_{i=1}^{n} |f_i - y_i|}{n}$ and the R-square $R = \frac{n\sum_{i=1}^{n} f_i y_i - \left(\sum_{i=1}^{n} f_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\left(\sum_{i=1}^{n} f_i^2\right) - \left(\sum_{i=1}^{n} f_i\right)^2}\sqrt{n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2}}$, where f_i is the ANN predicted output, y_i is the actual target, and n is the total number of dataset used. Note that MAE can be used to indicate error on the average and the R-square tells how well the outputs are replicated by the ANN model. Therefore, the lower of the MAE and the closer of the R-square to 1 are required to guarantee the network accuracy. In this work, the MAE obtained is very low while the R-square are very close to 1, i.e. see Table 2.

Table 2 ANN training results using the best architecture 3-16-20-2.

	Training datasets		Testing datasets	
Output	MAE	*R*-square	MAE	*R*-square
$N_{Cluster}$	0.013833	0.99907	0.016013	0.998449
Interface	0.002685	0.998097	0.003154	0.99747

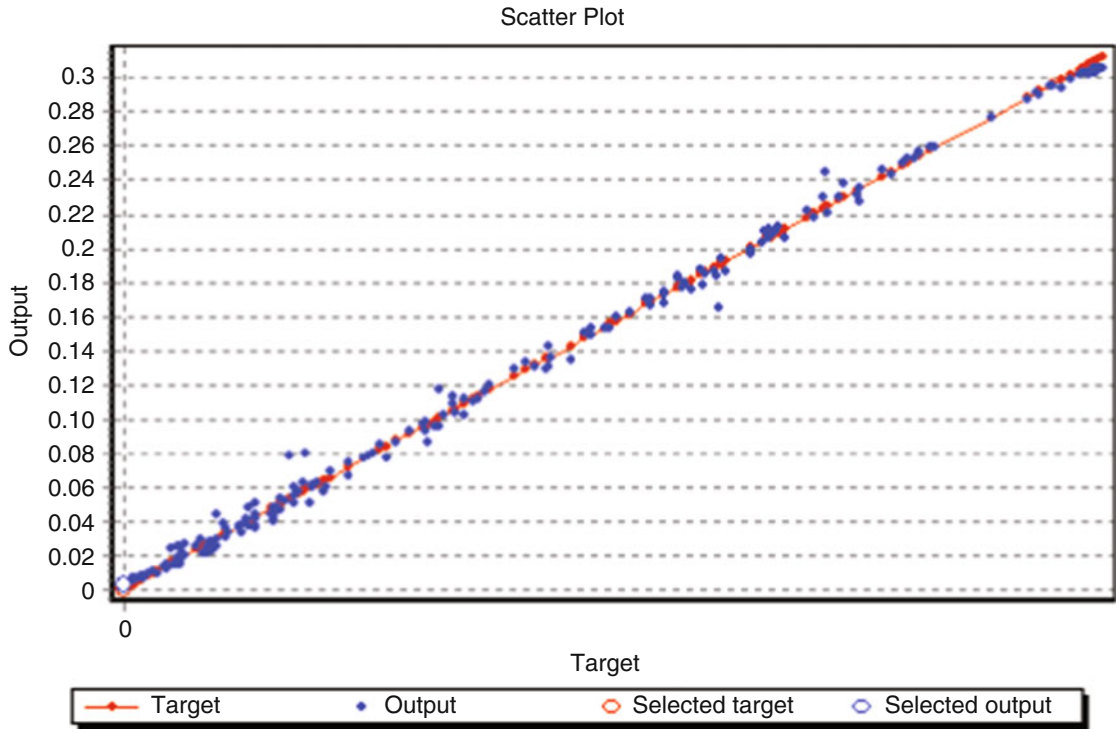


Fig. 4 The scattering plot for *interface* from ANN testing datasets. The linear fit between ANN output and the target has R-square = 0.99747 implying high accuracy of the ANN modeling.

For visual verification on the ANN accuracy, the scattering plot between the actual data and the ANN predicted data for interface from testing datasets was shown as an example in Fig. 4, which confirms ANN accuracy. Furthermore, to compare the results in a more informative way, the actual data for both $N_{cluster}$ and *interface* are plotted as discrete data points in Fig. 5 while the ANN data were plotted as solids line. As can be seen, the ANN data can be used as the representative of the actual data (as the lines pass through all actual data with somewhat high confidence). Therefore, it can be concluded that the proposed ANN models can accurately predict the results for the unseen inputs.

4 Conclusion

In this study, the effect of competition between adhesive and cohesive interaction, the system temperature, and the system size on the characteristic of domain forming, with an emphasis on percolating domain and their interface, in binary compound was investigated using Artificial Neural Network and Monte Carlo simulation. The average number of percolated domain and its interface profiles were obtained which can be used to identify the preferred configuration of the system. In addition, artificial neural network was used to construct

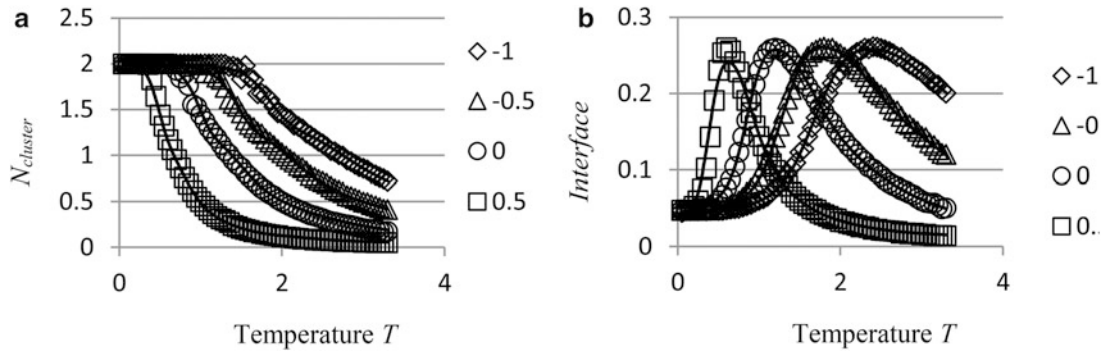


Fig. 5 (**a**) The average number of percolated cluster ($N_{cluster}$) and (**b**) the normalized percolated cluster interface (*interface*) as a function of temperature for J_{ad}/J_{co} = -1, -0.5, 0 and 0.5 calculated from $L = 40$ system. The discrete data points are from Monte Carlo results and the solid lines are from ANN results.

relationship between inputs and outputs in a form of database for future use. Good agreement between the predicted and targeted outputs confirms the reliabilities of the network and suggests the artificial neural network is applicable for modeling the interplay between adhesive and cohesive interaction in mixed binary compound system. Therefore, from the relationship provided, the work reported could be of some benefits for designing application based on binary compound materials where the material interface is crucial.

References

1. Gan, D., Lyon, L.A.: Interfacial Nonradiative Energy Transfer in Responsive Core-Shell Hydrogel Nanoparticles. J. Am. Chem. Soc. 123, 8203-8209 (2001).
2. Springer, J., Poruba, A., Vanecek, M.: Improved Three-Dimensional Optical Model for Thin-Film Silicon Solar Cells. J. Appl. Phys. 96, 5329-5337 (2004).
3. Hörling, A., Hultman, L., Odén, M., Sjölén, J., Karlson, L.: Mechanical Properties and Machining Performance of $Ti_{1-x}Al_xN$-Coated Cutting Tools. Surf. Coat. Technol. 191, 384-392. (2005).
4. Correia, H.M.G., Barbosa, H.M.C., Marques, L., Ramos, M.M.D.: Understand the Importance of Molecular Organization at Polymer–Polymer Interfaces in Excitonic Solar Cells, Thin Solid Films, http://dx.doi.org/10.1016/j.tsf.2013.10.057 (2014).
5. Kawasaki, K.: Diffusion Constants Near the Critical Point for Time-Dependent Ising Models I. Phys. Rev. 145, 224-230 (1966).
6. Newman, M.E.J., Barkema, G.T.: Monte Carlo Methods in Statistical Physics. Oxford University Press, Oxford (1999).
7. Jäckel, P: Monte Carlo Methods in Finance. John Wiley & Sons Ltd., Chichester (2002).
8. Landau, D.P., Binder, K.: A Guide to Monte Carlo Simulations in Statistical Physics (2nd Ed.). Cambridge University Press, Cambridge (2005).
9. Heringa, J.R., Blöte, H.W.J.: The Simple-Cubic Lattice Gas with Nearest-Neighbour Exclusion: Ising Universality. Physica A. 232, 369-374 (1996).
10. Krachler, R: Order–Disorder in Binary Alloys: A Theoretical Description by Use of The Ising Model. J Alloy Compd. 319, 221-227 (2001).
11. Schneidman, E., Berry, M.J., Segev, R., Bialek, W.: Weak Pairwise Correlations Imply Strongly Correlated Network States in A Neural Population. Nature. 440, 1007–1012 (2006).
12. Vives, E., Castán, T., Planes, A.: Unified Mean-Field Study of Ferro- and Antiferromagnetic Behavior of The Ising Model with External Field. Am. J. Phys. 65, 907-913 (1997).
13. Dayhoff, J.E.: Neural Network Architectures: An Introduction. Van Nostrand Reinhold, New York (1990).
14. Swingler, K.: Applying Neural Networks: A Practical Guide. Academic Press, London (1996).
15. Laosiritaworn, W., Khamman, O., Ananta, S., Yimnirun, R., Laosiritaworn, Y.: Artificial Neural Network Modeling of Ceramics Powder Preparation: Application to $NiNb_2O_6$. Ceram. Inter. 34, 809-812 (2008).

Cognitive Science Based Scheduling In Grid Environment

N.D. Iswarya, M.A. Maluk Mohamed, and N. Vijaya

1 Introduction

Grid is a form of distributed computing, it reaches the goal by collecting resources from multiple locations and perform manipulation of intensive and large scale data set problems [16]. Huge amount of data sets are generated throughout the world for specific research purpose. Data grid is one type of grid which aims to generate large number of data sets and has the ability to access, manage and transfer data sets securely. Different people need grids for different reasons like National grids, project grid, private grid, goodwill grid, peer-to- peer grid, cloud grid. Some of the international collaborative are AP GRID, EGI and OPEN GRID FORUM etc.. Some of the national grid initiatives are D grid, national grid, tera grid etc [9]. Grid applications often involve large amounts of data or computing resources that require secure resource sharing across organizational boundries. This makes complex understanding of Grid Application management. Grid middleware helps users with continual computing ability and uniform access to resources [17]. A number of Grid middleware and management tools such as Globus, UNICORE, Legion and Grid bus etc.. provides the users to access remote resources transparently over a world-wide Network. Without these middleware projects and products, grid computing simply could not exist.eg EMI, 3G BRIDGE, GLOBUS, IGE, and UNICORE. The middleware uses information about the different "jobs" to calculate the optimal allocation of resources [18].

Scheduling is the important part to assign a job (allocating suitable resources to workflow tasks) to the requested site [10]. Work flow scheduling mainly focus on assigning jobs to site in a sequence manner. In order to improve the scheduling decision by focusing on overall constraints, intelligence is added to the grid by means of cognitive artificial intelligence. Cognitive science is the process of knowing or study of mind. Artificial intelligence involves problem solving, generalization and perception which revolves around the study of high level cognition [11]. In this work it is used to make the scheduling system more intelligent, with particular reference to intelligent behaviour as computation. AI is used to study cognitive process in machine.

Many issues encircle the scheduling process like Heterogeneity, granularity, replication, storage, data locality, fault tolerance, security etc., of these issues this work concentrates on data locality issue since it is an important factor which influences the response time [12]. The main focus of this work is to make the system to learn, think and perceive the system environment before making scheduling decisions. For this a Cognitive engine (CE) is designed and distributed to all grid sites. Cognition describes the process of knowing altogether. For learning CE is used to recognize and process the incoming request by providing the information to learning machine. It makes thinking be easy and perfect, when the same request is coming again. Perception is generally based on association rules. CE can also keeps only the needed replicas and it replaces replica by less important with most important. This approach of scheduling will be helpful to the researchers to achieve optimization. It is the type of duplicating mind by implementing right program. The proposed algorithm is implemented using Java. The proposed intelligence cognitive mode algorithm will give better performance and it can also reduce the make span. It will be useful for the current emerging data intensive

N.D. Iswarya (✉)
Software System Group, Dept of CSE, MAM College of Engineering, Tiruchirappalli, Tamil Nadu, India
e-mail: ammuishu50@gmail.com

M.A.M. Mohamed
Software System Group, Dept of CSE, M.A.M college of engineering, Tiruchirappalli, Tamil Nadu, India
e-mail: ssg_malukmd@mamce.org

N. Vijaya
Software System Group, M.A.M college of engineering, Tiruchirappalli, Tamil Nadu, India
e-mail: ssg_vijaya@mamce.org

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_30,

applications such as high energy physics, data mining, climate models, etc.,

This paper is organised as follows: In Section 2 describes the papers related to Grid scheduling and Cognitive Science. In Section 3 shows the issues in Grid Scheduling and innovation work to overcome the issues. Section 4 explains the architecture of cognitive mode algorithm. Section 5 experiments and discussions. Section 6 conclusion and future plans.

2 Related Works

In [1] five different replication strategies for 3 different access patterns such as random, temporal, and geographical are explained. It focuses on bandwidth consumption and access latency which is different for each access pattern. Based on the access pattern the dynamic replication strategy will automatically create and delete replicas. Cascading strategy is one among the five and it works better in all cases. In [2] it decouples job scheduling and data replication algorithm and are addressed and optimized separately to simplify the design and implementation. With the help of the chicksim the performance of four different job scheduling algorithms and 3 different replication algorithms are evaluated and it is a decentralized implementation. In [3] by using optorsim optimization is done at two stages, it deals with queue access cost based on the data and status of network links. To process a large amount of data, map reduce is used. By using cost matrix (task, slot) optimal data locality based on weights is achieved [4]. Close to file job scheduling algorithm, select the best site based on the delay and minimum computation. It also allocates the job to the site closer to the best site with processing capacity [5]. Dynamic hierarchical replication strategy is used to decrease access latency to avoid the replica storage at many sites, it stores replica in the site where most of the file access occurs. It reduces the job execution time [6]. Combined scheduling strategy select a site based on the jobs and computing capacity. Modified dynamic replication strategy provides less access time based on the file access time. Here the replica was replaced based on the number of access of the file. It is an enhanced version of dynamic replication strategy, it also selects best replica based on the storage queue, distance between the nodes and data transfer time [7]. Bandwidth hierarchy based replication avoids network congestion by maximizing network level locality. It also decreases the access time by dividing the site into several regions. Fetching time is reduced only if the requested file is in the same region, so it works better only for the small amount of storage elements [8]. Modified bandwidth hierarchy based replication is the extension of bandwidth hierarchy based replication since it replicates the most accessed files [13].

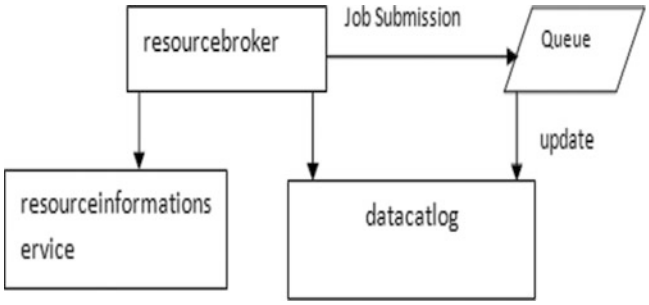


Fig. 1. General scheduling

Data intensive and network aware meta scheduling approach (DIANA) gives importance to network characteristics while making scheduling decisions. It also considers data and computing power to process the request. It create global cost matrix for all the sites and based on the least cost it allocates the job. For selecting best replica data location service is used. Decision is based on the overall cost which optimizes the entire scheduling process [15]. AI research have contributed to make different kinds of cognitive modelling. When a system cannot be evaluated by quantitative comparisons, accurate methods to evaluate the work contribution [14] helps to correlate brain with machines. Many research have contributed several approaches to improve the intelligency in machine.

3 Design Issues

Data Grid environment is comprised of storage elements (SE's) and compute elements (CE's). Based on scheduling it stores the data and execute the jobs. Three main steps of scheduling in complex system [19] are: Discovery of resources, suitable place for running a job, execution of a job. The procedure for scheduling jobs is shown in Fig. 1. The Resource Broker gathers the available resources through a Data catalog and Resource information Services respectively. Then, it dispatch a job based on the cost of resource and data transfer time based on scheduling mechanism. Results are transferred to the Resource Broker or another storage resource. Until all the jobs are scheduled this process is repeated. This work is designed to be used for grid environments. Many data intensive e-science applications uses grid environment for computation and storage of huge data sets. While storing and retrieving the huge data sets delay experienced by the grid users will get minimized, when this work is deployed in those environments. Imagine in neuroscience community, size of one brain image with higher resolution is in the order of 1 TB. If 500 images are to be analyzed and examined across different geographical locations by different researchers in grid environment, then this solution will be of much helpful in providing better

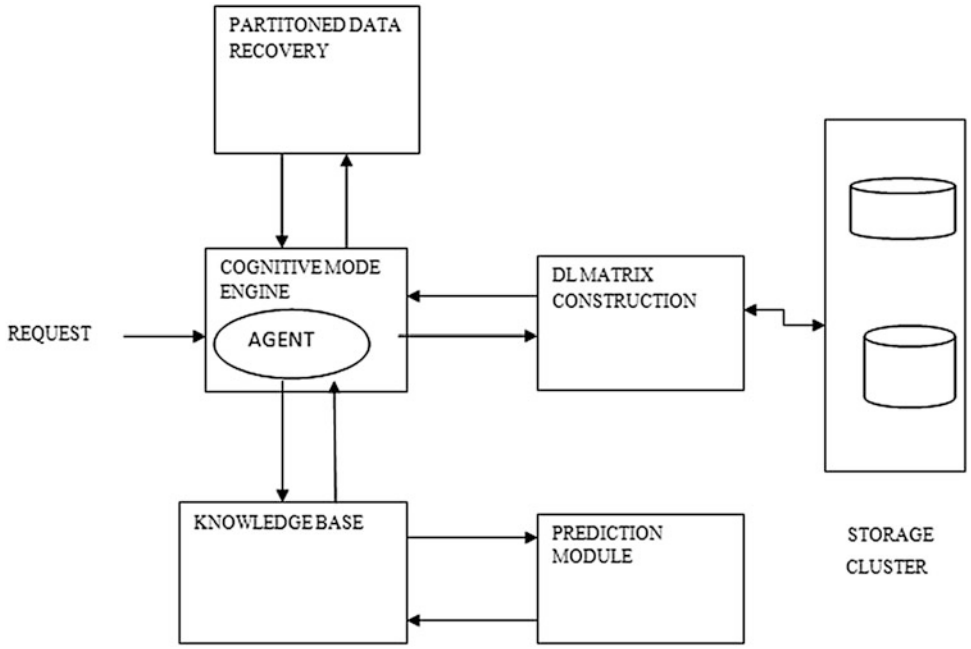


Fig. 2 System architecture

performance. Since these images cannot be transferred to the requested site for analysis because of its huge size and storage capacity of the requested site will not permit to store these huge size images. At this instant, usage of cognitive engine will intelligently handle the data sets and minimize the make span. In the fast moving IT business community, people don't have time to wait. This cognitive based scheduling in grid environment will reduce the response time of the customer.

4 Cognitive Mode Algorithim

Assigning jobs to a proper site is an important criterion in Grid Scheduling. In grid Scheduling each site in the grid contain cognitive engine which performs cognitive mode algorithm. Cognitive engine performs scheduling based on the data sets. The objectives for adding intelligence to the grid scheduling is to minimize and/or maximize one or more parameters like jobs' completion time, execution cost, system throughput, resources utilization, data availability time and response time. In Grid environment, many techniques and algorithms have been presented so far to handle scheduling effectively. But still make span is not optimized. In this work, a novel idea of introducing cognitive science is made. Here the problem solving skills of human beings were studied and applied to the scheduling problem in the grid. The Cognitive Engine deployed in the grid has the capability of learning, thinking and perceiving things from the environment and take decisions intelligently. Since this work optimizes the performance i.e., make span, it can be applied to cloud environment to tackle the challenging issues in scheduling.

```
//Learning and thinking
for each request R1 in site
LA: CE Get R1
{
if R1 already exist
Find d of R1 in DLM
}
Else If(R1 is available in only
one site )
{
CE checks BW
If BW=MAXIMUM
      send data
else if (BW=MINIMUM)
{
   Partition the data as chunks
   Store chunks in meta data
repository
   Send data
}
Else If(R1 is available in more
than one site)
{
   CE checks Best site based on BW
   Send the data
}
//perception
AH: If R2 arrives
    check KB
```

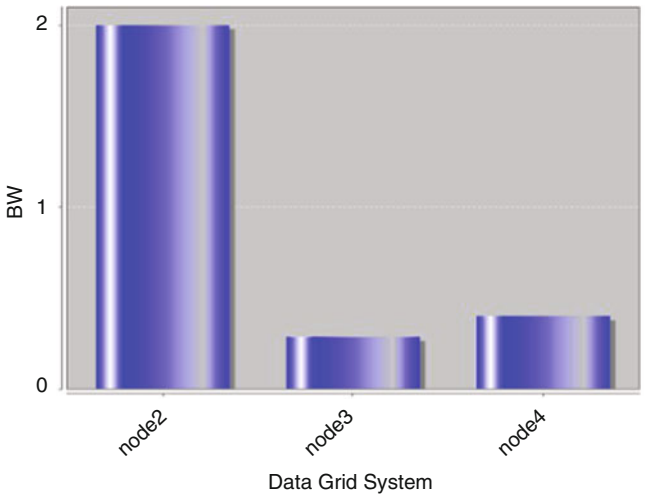


Fig. 3 Bandwidth Evaluation

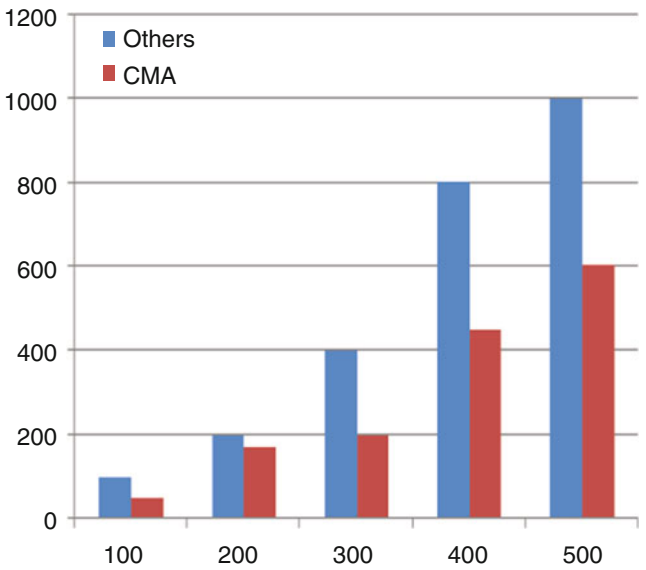


Fig. 4 Execution time Evaluation

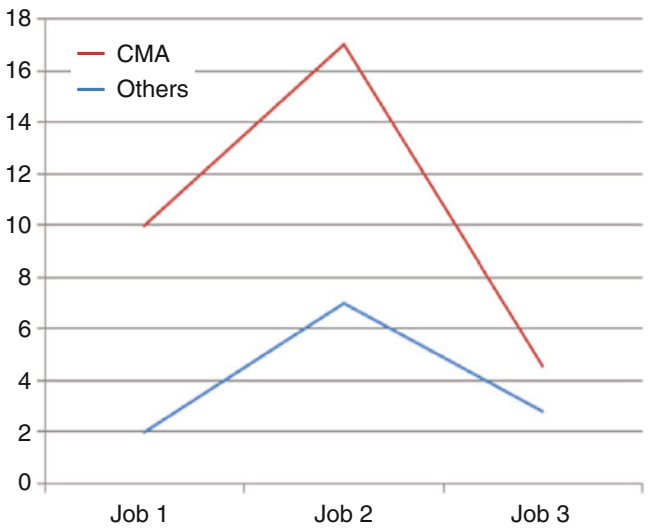


Fig. 5 Performance analysis of CMA with other scheduling algorithm.

```
R1==R2 SEND DATA along with relevant data stored in meta data repository
```

The core functions of the cognitive engine are Data locality matrix holds the distance of the requested datasets and the requesting site.Partitioned meta-data repository will hold the details of the data sets partitioned and the location of the grid data bases.The entire requests for data sets from various sites will be stored in the knowledge Base.When a request arrives the Learning Agent (LA)will learn the knowledge to see whether the request exists already if so then the subsequent requests made by the same sites are stored in the anticipation history.The Anticipation History(AH) otherwise called as prediction module which contains the requests predicted in advance by the site which contains the data sets. This will reduce the make span and delay, since the requests are predicted in advance and the required data sets are transferred well in advance to the corresponding site where it is required.

5 Discussion And Experiments

Proposed algorithm is implemented using java. Consider a grid environment with 4 nodes. Each node is refered as different sites in a grid. For example in the proposed grid environment the amino acid database is used. Node 1 send a request to grid environment for example asking for a glycine. Each node contain cognitive engine responsible for processing the resources based on the Bandwidth (BW)and Running time. The execution time of the jobs is monitored. If node 2 already processed this request the cognitive engine sends the executed job along with perception proline. Fig. 3 shows the Bandwidth Evaluation of 4 nodes which is efficiently used. Fig. 4 shows the execution time Evaluation of nodes where the overall time to execute a job is increased when the number of jobs are increased. This is due to the better selection of the resources. The proposed model is evaluated and compared with existing scheduling criterion. Fig. 5 shows the performance analysis of CMA compared to other scheduling algorithm. Fig. 6 represents the performance analysis of CMA based on the Time and Bandwidth of nodes used for the execution. This will reduce the execution time. It reduce make span and gives better improvement.

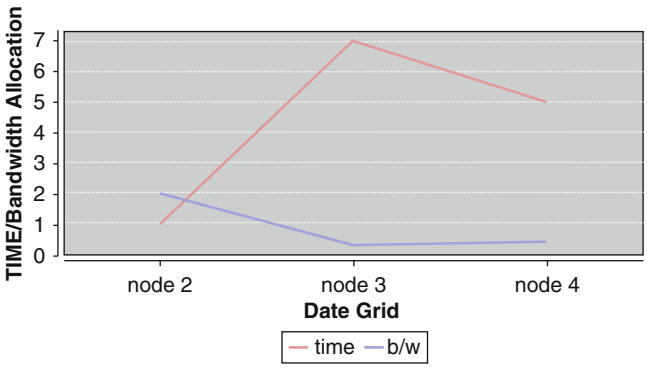


Fig. 6 Performance analysis of CMA based on time and BW

6 Conclusion And Future Plan

During scheduling, transferring huge data sets from one site to another site requires more network bandwidth. In order to mitigate this issue this work focuses on incorporating cognitive science in scheduling. Here a cognitive engine is designed and deployed in various grid sites. The intelligent agents present in CE will help in analyzing the request and creating the knowledge base. Depending upon the link capacity, decision will be taken whether to transfer data sets or to partition the data sets. Prediction of next request is made by the agents to serve the requesting site with data sets in advance. This will reduce the data availability time and data transfer time. Replica catalog and Meta data catalog created by the agents assist in decision making process. Nowadays cognitive science is playing major role in every aspect of business from customer understanding, team management to supplier relationships. The main aim of this work is to incorporate cognitive science to improve the performance in scheduling. It was initially designed to give solution in grid environment. Still some aspects like resource failure and fault recovery is not addressed here. In future it can be extended to solve these issues and may be applied to the current scheduling problems in cloud.

References

1. Manish parashar,"Grid computing introduction and overview"
2. Kavitha Ranganathan" Decoupling Computation and Data Scheduling in Distributed Data-Intensive Applications"
3. David G. Cameron, Rub´en Carvajal-Schiaffino David G. Cameron, Rub´en Carvajal-Schiaffino"Evaluating Scheduling and Replica optimization strategies in optor sim" CERN, European Organization for Nuclear Research, 1211 Geneva, Switzerland.
4. Mohamed HH, Epema DHJ. An evaluation of the close-to-files processor and data Co-allocation policy in multiclusters. In: International conference on cluster computing. USA: IEEE Society Press; 2004. p. 287–98.
5. Mansouri N, Dastghaibyfard GH. A dynamic replica management strategy in Data Grid. Journal of Network and Computer Applications 2012;35(4):1297–303.
6. Najme Mansouri, Gholam Hosein Dastghaibyfard, Ehsan Mansouri," Combination ofdata replication and scheduling algorithm for improving data availability in Data Grids"
7. Park S-M, Kim J-H, Go Y-B, Yoon W-S. Dynamic Grid replication strategy based on internet hierarchy, in international workshop on grid and cooperative com- puting. Lecture Note in Computer Science 2003;1001:1324–31
8. Sashi K, Thanamani A. Dynamic replication in a Data Grid using a modified BHR region based algorithm. Future Generation Computer Systems 2011;27(2): 202–10.
9. The Data Grid project. /http://eu-datagrid.web.cern.ch/eu-datagrid/S.
10. Carsten Ernemann, Volker Hamscher, Uwe Schwiegelshohn, Ramin Yahyapour" On Advantages of Grid Computing for Parallel Job Scheduling"
11. Stuart J. Russell and Peter Norvig" Artificial Intelligence A Modern Approach".12 Stefka Fidanova"Simulated Annealing for Grid Scheduling Problem"
12. McClatchey R, Anjum A, Stockinger H, Ali A, Willers I, Thomas M, "scheduling in Data intensive and network aware (DIANA) grid scheduling", Journal of Grid Computing,2012
13. Solomonoff"Some recent work in artificial intelligence"
14. Jim Davies" The Role of Artificial Intelligence Research Methods in Cognitive Science".
15. Jiehai Cheng, Wei Li,"Research of the application of Grid computing on geographical information system"
16. Parvin Asadzadeh, Rajkumar Buyya
17. Chun Ling Kei, Deepa Nayar, and Srikumar Venugopal" Global Grids and Software Toolkits: A Study of Four Grid Middleware Technologies".
18. www.gridcafe.org
19. McClatchey R, Anjum A, Stockinger H, Ali A, Willers I, Thomas M. Data intensive and network aware (DIANA) Grid scheduling. Journal of Grid Computing 2007;5:43–64.

Vulnerability evaluation of multiplexing PUF for SVM attacks

S. Kiryu, K. Asahi, and M. Yoshikawa

1 Introduction

Following the advancement of reverse engineering technologies, electronic component counterfeiting has become an increasingly serious problem. This counterfeiting issue results in financial damage to companies, such as impairment of the brand image and a decrease in sales. It has been pointed out that if counterfeit electronic components are mounted onto electric vehicles or medical appliances, a fatal accident may occur. A feature article for Nikkei Electronics in April 2010 discussed the damage caused by counterfeits[1]. At present, counterfeit semiconductors are said to occupy approximately 5 % of the global semiconductor market. Measures that are necessary to prevent counterfeiting include the management of electronic components by assigning a specific identification to each electronic component and the creation of identification that is difficult to clone. The physical unclonable function (PUF)[2]-[7] is now attracting attention as a technique to prevent counterfeiting. The PUF extracts and uses arbitrary dispersion as specific identification, which is generated during semiconductor manufacturing. Arbitrary dispersion during semiconductor manufacturing is impossible to artificially control and is difficult to clone. Various conventional PUFs have been reported to exist, and an arbiter PUF has been used in practice. The vulnerability of the arbiter PUF to machine learning attacks has been pointed out[5].

On the other hand, we have developed a new 4-MUXs PUF[2] and proposed its modeling method for machine learning attacks[3]. In addition, we discussed the vulnerability for attack using Neural Network.

This paper discusses the vulnerability of the proposed 4-MUXs PUF for attack using support vector machine (SVM). Experiments show superior resistance for SVM attacks in comparison with conventional PUF.

2 Preliminaries

2.1 Multiplexing PUF

Figure 1 shows a conventional PUF. As shown in Fig.1, a unit is composed of 2 selectors in a conventional PUF. By contrast, a unit is composed of 4 selectors in the 4-MUXs PUF. In conventional 2-MUXs PUF, since a unit is composed of two selectors, the signal to select 2 selectors requires 1 bit. Therefore, 2 N units are used for the 2 N-bit challenge. In the 4-MUXs PUF, since a unit is composed of four selectors, the signal to select 4 selectors in each unit requires 2 bits. Therefore, N units are used for the 2 N-bit challenge.

In the 4-MUXs PUF, 4-bit output can be obtained by inputting one start signal. The difference in delay between every pair of the selector block's four outputs (4 bits) is detected using DFFs. By detecting the difference in delay between every pair of signals output from the selector block, the order of arrival of four signals can be specified. In addition, to detect the difference in delay stably, a pair of the earliest and latest arriving signals is used as a response. A 2-bit identification number is assigned to each of four signals, and 2-bit identification numbers of the earliest and latest arriving signals are coupled to generate a 4-bit response.

2.2 Modeling for attacks

This paper adopts the modeling method[3], which was based on reference[6]. The present study uses four classes and creates a challenge model that can express the operation of

S. Kiryu (✉) • K. Asahi • M. Yoshikawa
Dept. of Information Engineering, Meijo university, Nagoya, Japan
e-mail: dpa_cpa@yahoo.co.jp

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_31,

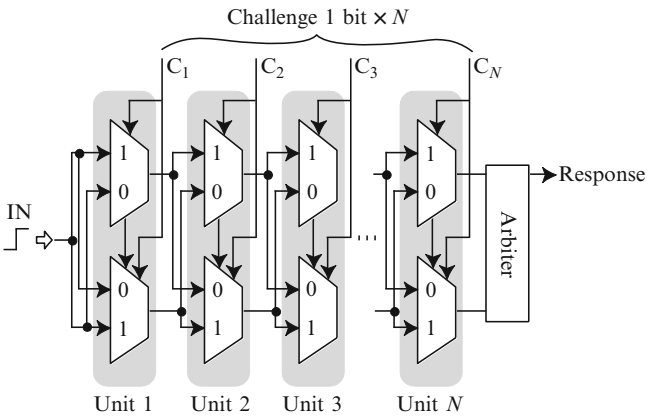


Fig. 1 Example of conventional PUF

the 4-MUXs PUF in a manner similar to that of conventional 2-MUXs PUF.

Figure 2 shows the operation obtained when challenge 00111101 is input into the 4-MUXs PUF. Class 4 is used in the first step. And then, since challenge 11 is input into the selector, class 1 is used in the second step. Therefore, the class when inputting the challenge changes in such a manner as 4, 4, 1, 2, and 1. Since signals to be noticed are assumed to pass through selector 1, the final class is always 1. Identification numbers are given to all the classes except the final class, and an 8-bit challenge model is obtained. In the case of Figure 2, the challenge model is 11110001.

3 Vulnerability evaluation

3.1 Learning using a support vector machine

Methods. Using a support vector machine (SVM), an experiment was performed for a conventional PUF and the proposed PUF, and the correct answer rate was compared between the two PUFs. The challenge lengths were set at 8, 16, 32, and 64 bits (bit lengths), and each case was examined. For the experiment, a simulation program was created for each of the conventional and proposed PUFs (4-MUXs PUFs). The experimental procedure was as follows.

1. Challenges for learning data were created. The numbers of data sets were 100, 1000, and 2000.
2. For test data, 10000 challenges were created.
3. For each data set, a challenge model was obtained.
4. Using the simulation PUF, the response corresponding to each of the challenges for learning and testing was obtained. In the case of the proposed PUF, because of the 4 bits, the response was converted into a decimal number.
5. The challenge model for learning data was forced to learn using the SVM. The response was used as a label.
6. After learning, a learning model was output from the SVM. Using the learning model, the challenge model for the test data was identified as a label.
7. The identification results were compared with the response of the test data to obtain the correct answer rate.
8. For each PUF, whether the response to the challenge agreed with the learning results was examined to obtain the correct answer rate.

Results. Figure 3 shows the results obtained by performing machine learning attacks against the conventional PUF using the SVM.

In the case of using the model, the correct answer rate exceeded 95 % at all the bit lengths when the number of learning data sets was more than 3000. Therefore, the conventional PUF was revealed to be vulnerable to machine learning attacks using the SVM. The correct answer rate was higher when the challenge model was forced to learn than when the challenge was directly forced to learn. Therefore, the modeling was confirmed to be effective.

Figure 4 shows the correct answer rate obtained when the proposed PUF was forced to learn the challenge model and that obtained when the proposed PUF was directly forced to learn the challenge. At each bit length, the correct answer rate was higher when the challenge model was forced to learn than when the challenge was directly forced to learn. Therefore, the proposed modeling was confirmed to be effective. In the case where the correct answer rate was compared between the conventional and proposed PUFs, when the challenge was 8 bits, the correct answer rates of both PUFs were 100 %. However, the correct answer rate of the proposed PUF was significantly lower than that of the conventional PUF as the number of bits increased. Therefore, the resistance to machine learning attacks using the SVM was revealed to be higher in the proposed PUF than in the conventional PUF.

3.2 Changes in the correct answer rate due to different learning methods

Methods. When the proposed PUF was forced to learn using the SVM, the 4-bit response was converted into a decimal number, as described in the experimental procedure (4) in section 3.1.1 (Methods). Therefore, even when the proposed PUF did not learn at all, the correct answer rate was $100/12 = 8.33$ %. When the conventional PUF did not learn at all, the correct answer rate was approximately $100/2 = 50$ %. In This section, we describe the correct answer rate to the response but not that to the 1-bit response.

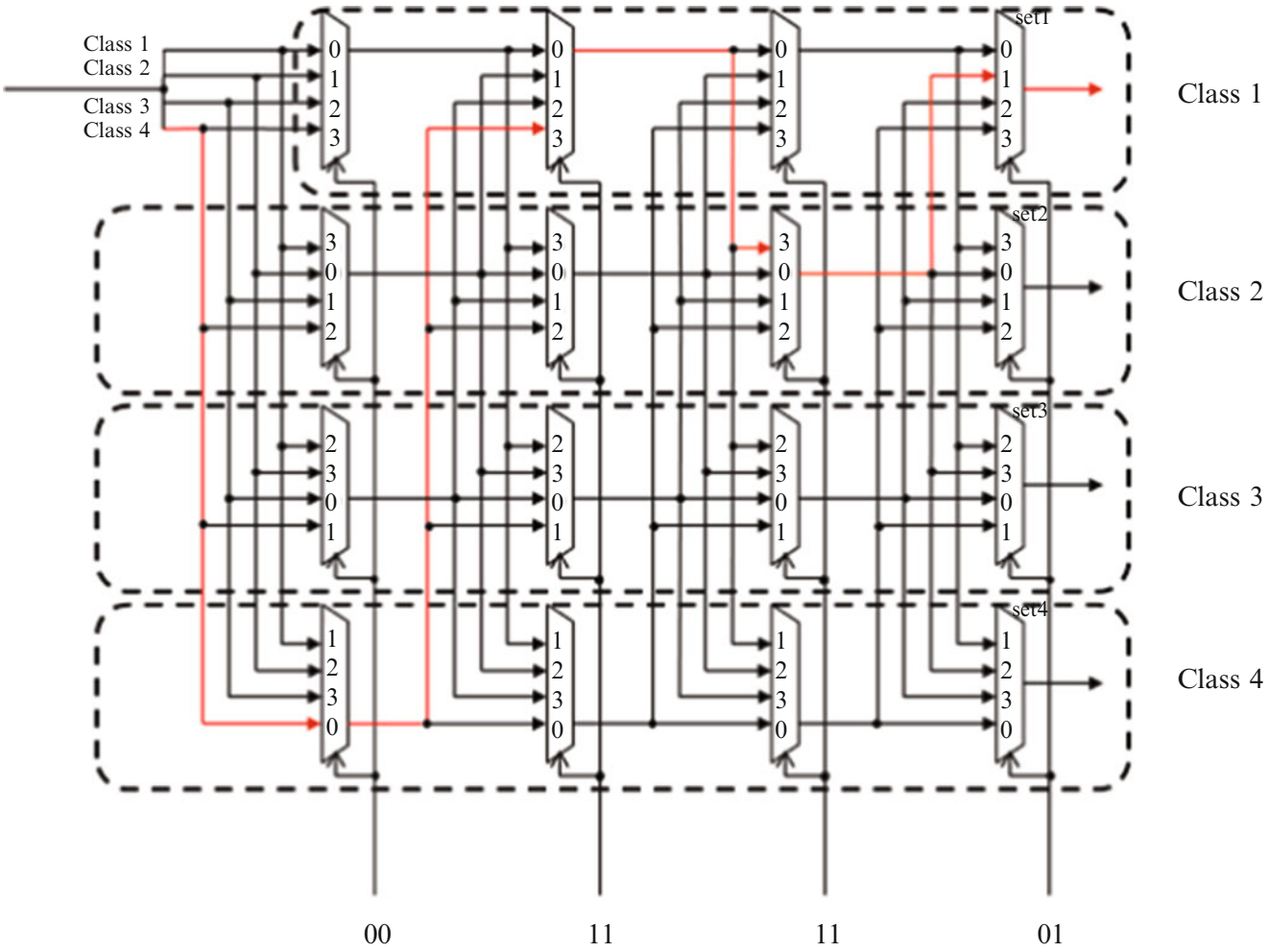


Fig. 2 Example of an operation with challenge 00111101

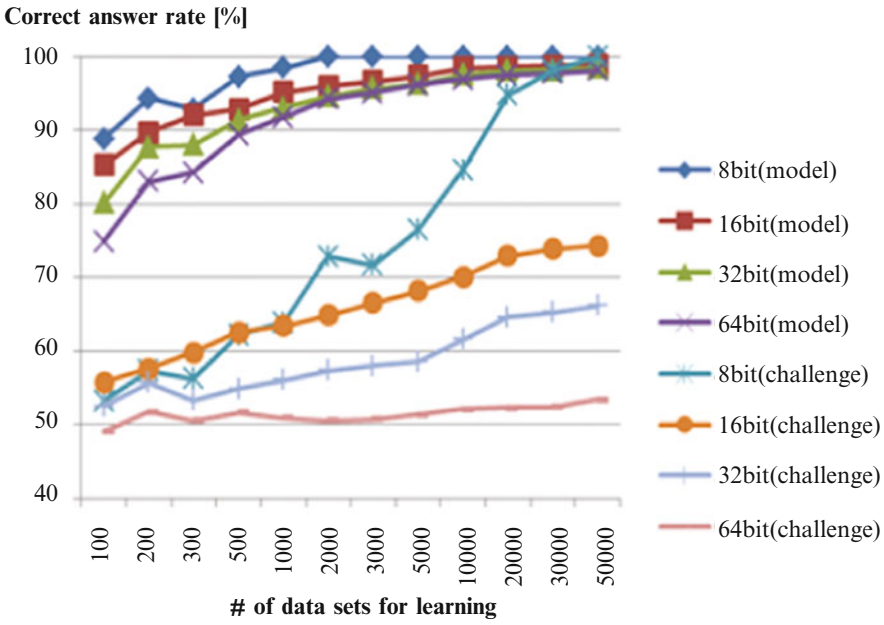


Fig. 3 Results obtained by performing SVM attacks against the conventional PUF

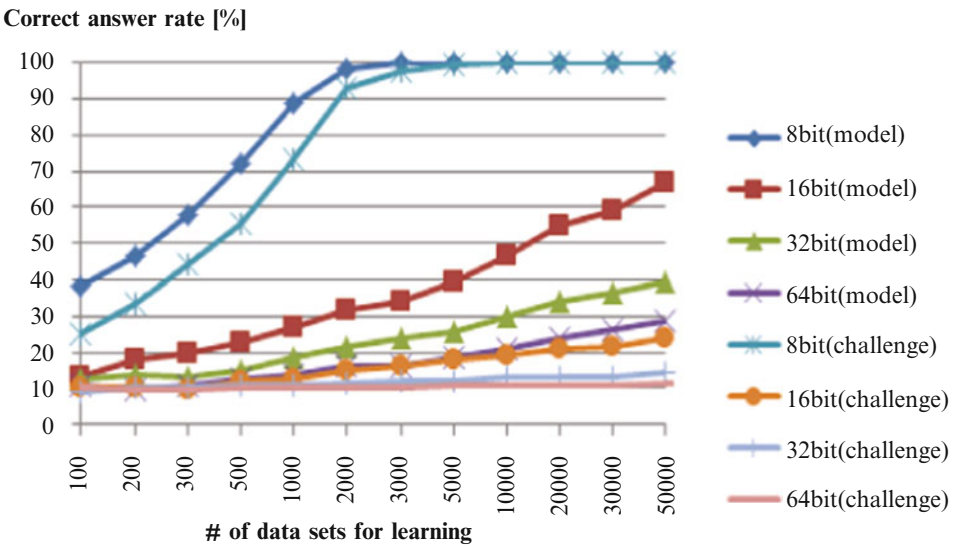


Fig. 4 Correct answer rate of the proposed PUF

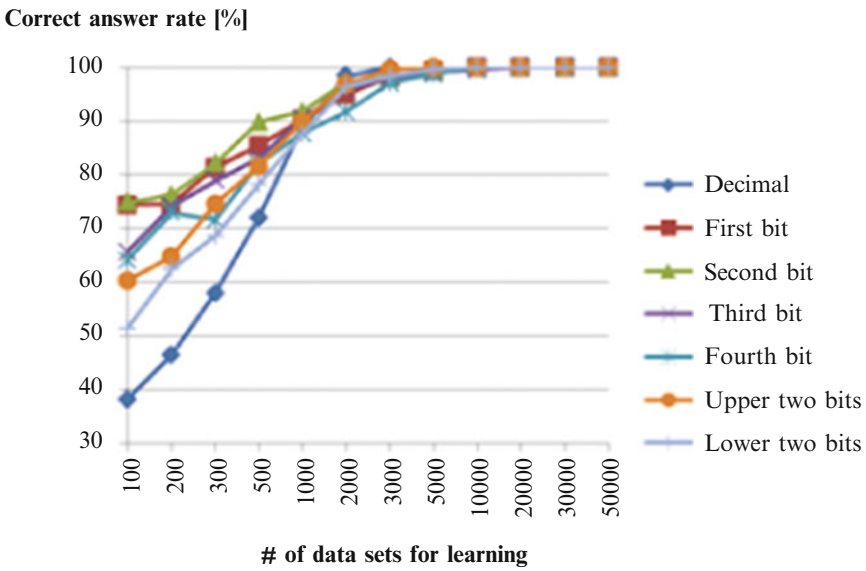


Fig. 5 Results using 8-bit challenge

Therefore, the correct answer rate to the 1-bit response was obtained.

To examine the changes in the correct answer rate due to different learning methods, an experiment was performed. The experimental procedure was as follows.

1. In the 4-bit response, the first bit was forced to learn.
2. The second bit was forced to learn.
3. The third bit was forced to learn.
4. The fourth bit was forced to learn.
5. The first and second bits (the upper two bits) were converted into decimal numbers, and the converted numbers were forced to learn.
6. The third and fourth bits (the lower two bits) were converted into decimal numbers, and the converted numbers were forced to learn.

Results. Figures 5-8 show the experimental results obtained using four bit lengths. In the case of 8 bits, when the number of learning data sets was 10000, the correct answer rate exceeded 99 %.

Since only 256 patterns existed for the challenge, searching the objective challenge and response from the learning data was considered to be better than forcing to learn. For bit lengths other than 8 bits, the difference in the correct answer

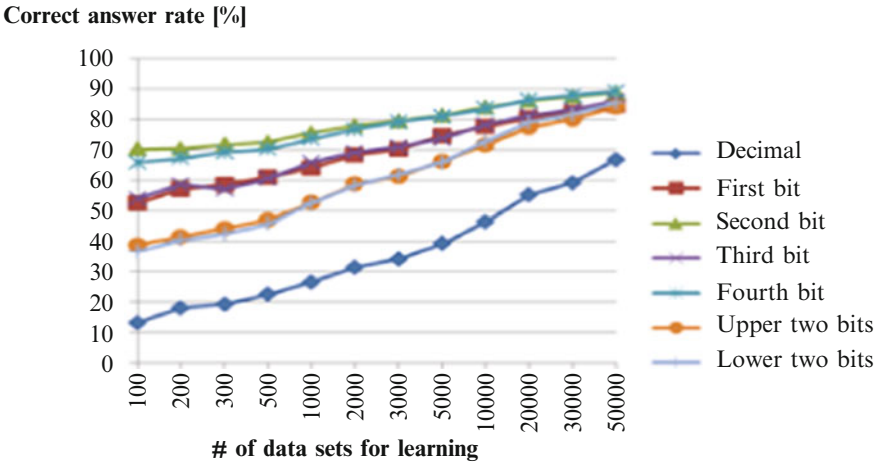


Fig. 6 Results using 16-bit challenge

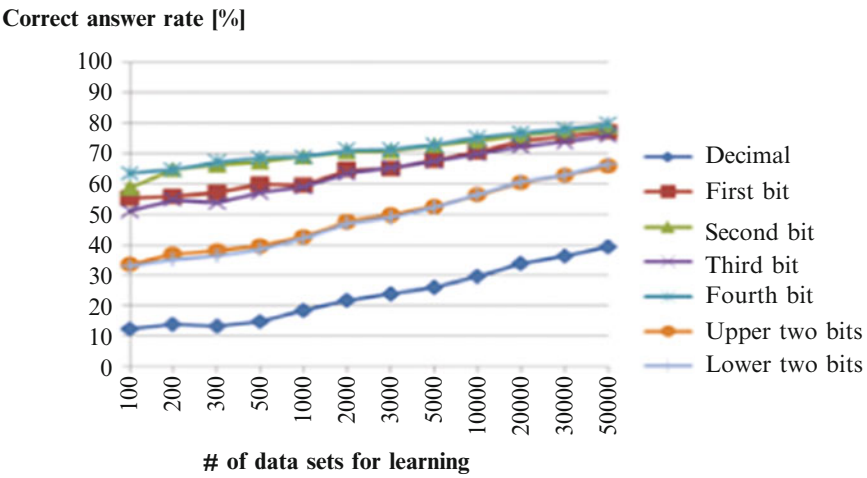


Fig. 7 Results using 32-bit challenge

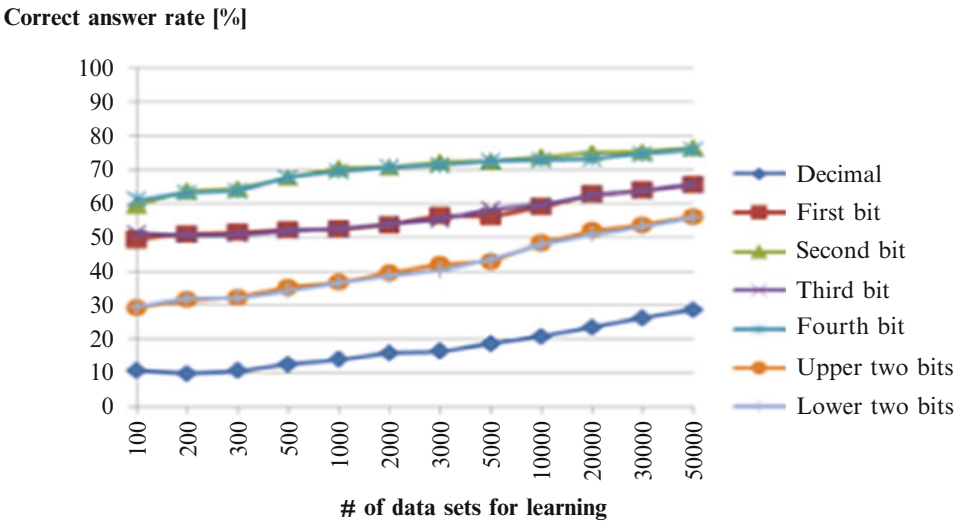


Fig. 8 Results using 64-bit challenge

rate between learning the first and second bits and between learning the third and fourth bits was less than 2 % in almost all cases.

A similar tendency was observed in the correct answer rate when learning the upper and lower two bits. In all cases, the correct answer rate was higher when learning the second and fourth bits than when learning the first and third bits. To examine whether this phenomenon was peculiar to the simulation PUF used, a similar experiment was performed using the simulation PUF with different parameters.

However, the results obtained in this experiment were similar to those shown in Figures 6-8. Therefore, it was revealed that the proposed PUF was more vulnerable to machine learning attacks using the SVM at the second and fourth bits than at the first and third bits.

Based on these results, it was presumed that a model formula that is easily used to attack the second and fourth bits could be created, although the solution of the proposed challenge model was difficult to obtain.

4 Conclusion

This paper discussed the vulnerability of the proposed 4-MUXs PUF for SVM attacks. Experiments show that the resistance to machine learning attacks using the SVM was revealed to be higher in the proposed 4-MUXs PUF than in the conventional 2-MUXs PUF. Moreover, the vulnerable points were also clarified.

In the future, we will attack to an actual device and compare the results of the proposed modeling with those of an actual device.

Acknowledgment This research was supported by Japan Science and Technology Agency (JST), Core Research for Evolutional Science and Technology (CREST).

References

1. Nikkei Electronics(in Japanese) http://techon.nikkeibp.co.jp/NE/
2. M.Yoshikawa, A.Naruse, "Multiplexing aware arbiter physical unclonable function", Proc. of IEEE. International Conference on Information Reuse and Integration, pp.639-644, 2012.
3. S.Kiryu, K.Asahi, M.Yoshikawa, "Modeling and Attack for 4-MUXs based PUF", Proc. International Conference on Security and Management, 2014.
4. Jae W. Lee,D. Lim, B. Gassend, G. E. Suh, M. vanDijk, and S. Debadas, "A Technique to Build a SecretKey in Integrated Circuits for Identification and Authentication Applications", in Proceedings of the IEEE VLSI Circuits Symposium, pp.176-179, 2004.
5. U. Ruhrmair, F. Sehnke, J. Solter, G. Dror, S. Devadas, J. Schmidhuber, "Modeling Attacks on Physical Unclonable Functions", in Proc. of ACM Conference on Computer and Communications Security, pp.237-249, 2010.
6. Daihyun Lim., "Extracting Secret Keys from Integrated Circuits.2, Msc thesis, MIT, 2004.
7. S. Kumar, J. Guajardo, R. Maes, G.J. Schrijen and P. Tuyls, "The Butterfly PUF: Protecting IP on every FPGA", Proc. of IEEE International Workshop on Hardware Oriented Security and Trust, pp.67–70, 2008.

Autonomous Visualization for Mitigating Lack of Peripheral Vision in Remote Safe Teleoperation

J. K. Mukherjee

1 Introduction

Human operators of Teleoperated remote robots working in radioactive or contaminated and toxic confines[1][2] face problem of limited visual perception owing to variety of reasons like constrained view from fixed locations at long stand-off distance owing to image sensors' susceptibility to radiation inflicted damage [3], likely occlusion owing to limited camera location feasibility(fig.1) and absence of peripheral vision capability available in natural biological vision. Often tactile fee back is relied upon for overcoming the limitation[4].

The paper proposes a substitute for aforesaid constrained vision, through intelligent view synthesis of 'pertinent workspace zones' falling out of the limited view field in real-time. The criteria of pertinence is context dependant. For example for guessing the trajectory of mobile robot [5] or a near future position of end-effecter, 'look ahead modality' has great utility for navigating. Similarly attracting attention to possible obstacle hit in invisible zone is another modality. The later enhances work throughput by relieving operator from robot damage concerns. The present work devises a fast method for detecting conditions of impending hit between robot in dynamic state and workspace parts by vicinity space coding and also assists autonomous visualization system in synthesizing graphic view of the spatial condition by pointing 'virtual viewer' at hit prone position. High speed execution needs innovative elimination of complex nondeterministic search of surface intersections[6].

J.K. Mukherjee (✉)
BARC, Mumbai 400085, India
e-mail: jkmukh@yahoo.co.uk

2 Related work

Part modeling has matured over time [7]. Workspace modeling using CAD is now common. Several frameworks like open GL, Direct-X [8] etc. are rich platforms for forming virtual reality 'VR'[9]. The robot modeling is straightforward as it is synthesized using component CAD models and its spatial state computed online through its kinematics model and joint sensors [10]. Collision detection has received wide attention [11] and for minimization of cumbersome geometric intersection search optimization have been attempted [12]. Hit avoidance in mobile robot navigation is gross in nature. Associating 'potential' to space around a part helps in attracting or dissuading a robot that acts as charged body [13]. In the context of the present objective, two remarkable difference are worth noting 1; impending hit and not actual collision needs to be detected and 2; the method should be precise unlike that for mobile robot. This implies that objects have to be modeled precisely.

3 New approach

The proposed approach takes recourse to activating the vicinity of parts to allow sensing the *susceptibility* of the robot to collision with workspace part, identify the spatial location of such occurrence and form input to Virtual- Reality 'VR' environment for simulating the pertinent space zone with details in a 'see pertinent zone' SPZ console on-line as an aid to operator of the tele-control system.

3.1 Requirements of Model

The activation model must enable sensing of susceptibility of robot to hit and support determination of suitable viewer orientation for synthesizing local scene. For a structured

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_32,

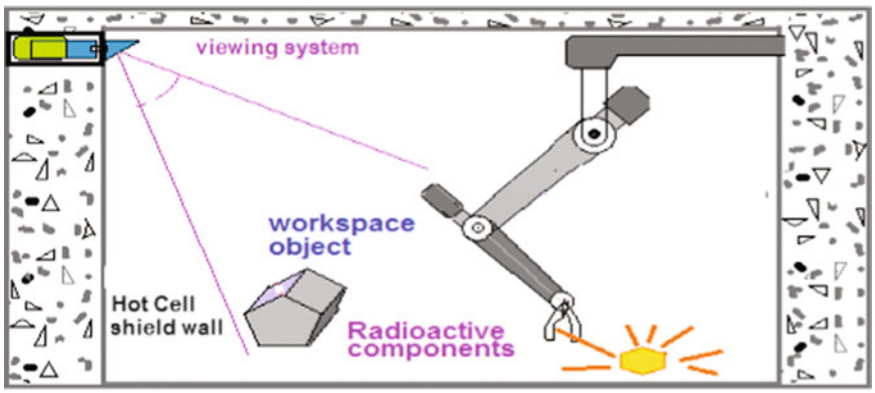


Figure 1 Tele-working hot cell with long stand-off camera and limited peripheral coverage in on-line view field

discussion we consider a point robot ***R*** which is located on robot body. Its position in space is computed by robot kinematics model and joint parameters at a given time 't'. Since all points on a sequential robot's body have unambiguous kinematic definition, our prerequisite on position computability of $\boldsymbol{R_t}$ at time 't' is feasible and $\boldsymbol{R_t}$(x,y,z) is known. At time 't' the nearest distance of ***R*** from any object surface '**S**' determines susceptibility to hit. High Number of parts and complex shapes give rise to large number of surfaces and escalate difficulty in finding the minimum distance [11].

In a space segment that is coded as 'vicinity space', existence of $\boldsymbol{R_t}$ confirms 'Susceptibility to Hit' *STH* condition. Therefore if space in contact with object body is segmented as neighbor or vicinity space a susceptibility sensing modality can emerge. This space is a shell all around 3D object with finite uniform thickness and its inner surface is in contact with the outer surface of the object. ***R*** entering the vicinity space at high speed will trigger little time before hit surface '**S**', needing vicinity shell to be thicker for trigger to occur reasonably earlier than hit. A slow ***R*** will then cause STH alert spuriously. To solve this, multiple shells of increasing size can be deployed around object and their sensitivity may be related to speed of ***R***. We call these shells 'Susceptibility Test Shells' $STS_{(n)}$. Order 'n' is 1 at shell in immediate contact with body and 'q' at farthest shell.

3.2 Vicinity Model

The popular CSG (constructive solid geometry) representation evolves from description for polyhedral objects [7]. As CAD systems mostly do not expose internal details, the method is not preferred. Standard Stereo lithographic format 'STL' data is always available and has been used in our approach here. 'STL' format offers smallest surface entity as plane and so offers ease of geometric operations like intersection.

4 Voxelised workspace model

Voxel oriented approaches have delivered very encouraging results in 3D path finding [14]. In such approaches workspace is represented as 3 dimensional array WVA(x,y, z) of equal cube volume elements named as Voxel and designated as $V_{(XYZ)}$ in further descriptions. It represents cube with diagonal nodes X_i,Y_j,Z_k and X_{i+1}, Y_{i+1} Z_{i+1} (fig. 2a).the indices range from 0 to a max integer value suitable for representing full extent of workspace with desired granularity. The voxels was assigned 'potential' property for path finding problem in [14]. Here we use them for representing spatial occupancy of space by parts in the workspace and other data. A $V_{(XYZ)}$ not occupied by object is empty type; where a triangular surface passes is boundary type (fig. 2b) and where body exists, is inner type. Voxel data record (fig.3) supports linking of other properties to it. Coding of space as STS layer, explained earlier, is also attached to voxel data.

4.1 Forming voxel based model :

Entire workspace voxel array is initialized as 'empty' type. Object model is placed as per its pose and position in work space using node set rotations and translation[8]. In phase I, triangular surface segments (CAD- STL faces) are intersected by grid lines of Cartesian coordinate system, spaced at separation equal to voxel size. Each intersection point's coordinates are binned using the real space to voxel map scale factor for use as index to workspace voxel data array '$WVA_{(x,y,z)}$' and type 'boundary' is assigned to the voxel (fig.3) indexed by x,y,z. Voxels at constant z (integer) levels in Z max >=Z >=Zmin range have closed 2D contour formed by boundary voxels of object (fig. 2c). By using raster scan method [15] the occupied voxel are assigned 'inner' type. Operation over Z min to Z max, results in full voxelisation of part. Object serial number is appended as 'object index' for back referencing to voxel.

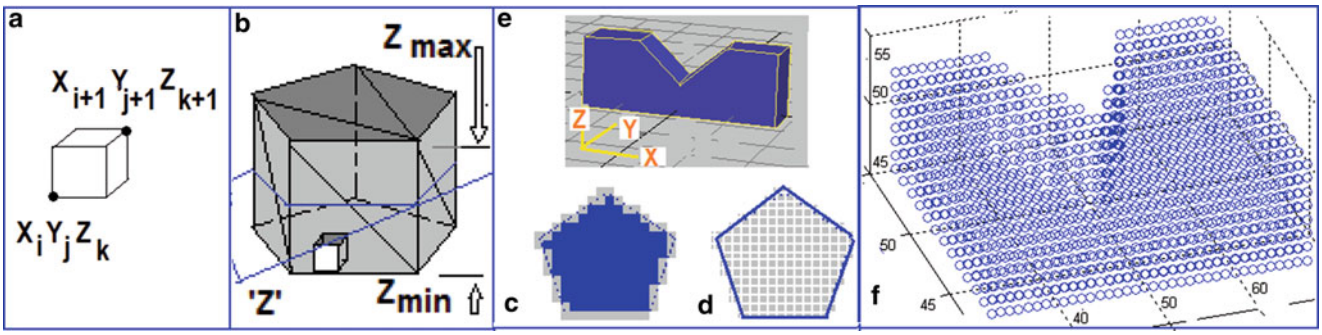


Figure 2 a: voxel definition, b: surface voxel of cylinder object. c: top layer surface voxel and d: intermediate layer in cylinder with body voxel. (e) object (f)its voxelised version.

Grid Pt. X,Y,Z	Object index	Type (empty/ boundary/ inner)	STS (order)	Susceptibility Level β

Figure 3 Data associated with a voxel in 3D workspace voxel array $WVA_{(x,y,z)}$

4.2 A new algorithm for STS coding:

Subspace of cube shape around chosen object is formed from 3D voxel array and referred as object voxel array OVA(x,y, z) [16]. Its size is based on the extent of maximum STS of order 'q' A floating analyzer zone 'FZ' is formed as 8 connected neighbor voxel set around a center voxel FZ(i, j, k) (fig.4a). It has 2 m + 1 layers of (2 m + 1) x (2 m + 1) voxels. This 3D zone is indexed by indices i,j,k varying in range –m*i to m*i, –m*j to + m*j and –m*k to + m*k along x,y and z. Size of FZ is determined by integer 'm' around center voxel $FZ_{(0,0,0)}$.

The object voxel array $\mathbf{OVA_{(x,y,z)}}$ with data as in figure 3, is scanned. On finding a voxel with property as 'boundary', **FZ** is aligned with $\mathbf{FZ_{(i=0,\ j=0,\ k=0)}}$ at $\mathbf{OVA_{(x,y,z)}}$. A search within local neighbourhood at $\mathbf{OVA_{(x,y,z)}}$ is run within **FZ** limits and voxels of 'empty' type are converted to type 'STS' order '1'. Note that '0' at STS field means 'not STS' voxel. This process is continued over the entire x,y,z index space $\mathbf{OVA_{(x,y,z)}}$. The resulting OVA with new 'STS' type of order n = 1 in 'order' field is shown in figure 4d. For higher order STS($1 < n < q$), formation method is repeated. All objects are treated similarly and OVAs aremapped back to WVA(x,y,z).

4.3 Variants of STS formation:

a. Shaping the FZ : The binary content of 'FZ' can be used for coding digitized 3D shape of the effective FZ. A sphere 'Shape Element' *SE* can be formed in 5x5x5 cube by setting mask values as 1 for only those cells in FZ, which are included completely inside sphere of radius 2 voxels. Radius has integer values. Digitized shape approximation is more accurate for higher 'm'.

b. Nonlinear 'Susceptibility' assignment: Last field in voxel record (fig.4) is assigned susceptibility property β. A one dimensional array '$\boldsymbol{\beta}_n$' is used for assigning discrete susceptibility values for STS(n). The **β** assignment may be linear or nonlinear with higher value near object surface.

5 On-line Process for Fast Susceptibility Detection

5.1 Implementation Scheme

A susceptibility estimation framework is created by forming an on line computation scheme (fig.5). The master (fig 5a) sends control signals to a similar slave (fig. 5c). A 'virtual workspace' *VWS* is formed by WVA(x,y,z). A stick model of slave robot, which hosts a (or multiple) susceptibility probe (fig 5b) is maintained in same state as the slave in VWS to impart spatial and dynamic conditions of telecontrolled robot to the sensor. It's position (x,y,z) and velocity in VWS is computed by the kinematics of slave robot and joint angles sensed by encoders.

5.2 'Susceptibility to hit' Trigger

At the probe position 'x,y,z', the susceptibility value is compared with a threshold (fig.6a) and if found alarming, the position of probe is latched and trigger is generated for visualisation (fig.5d). The threshold can be made speed sensitive.

5.3 Enhanced Sensor Implementation

A 'Susceptible Local Array' (SLA) is formed to implement enhanced susceptibility processor (fig 7b) as 8 connected

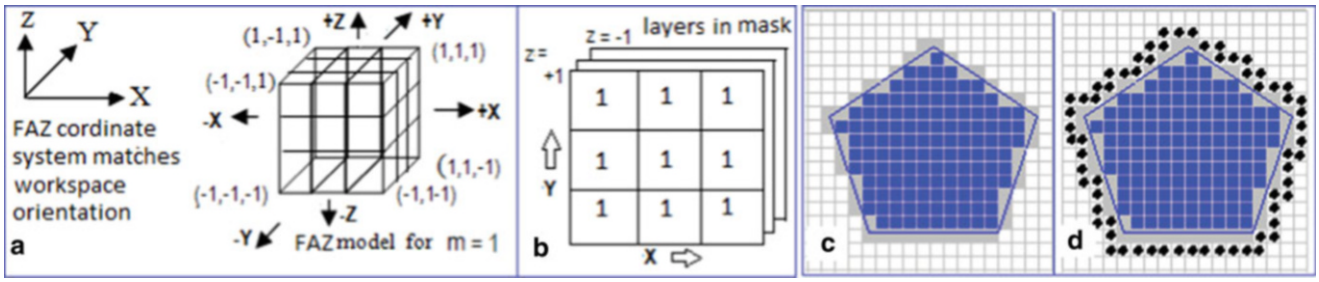


Figure 4 (a) Floating analyzer zone, (b) SE mask values, (c) Object, (d) neighbour processing results showing added outer layer (black dots)

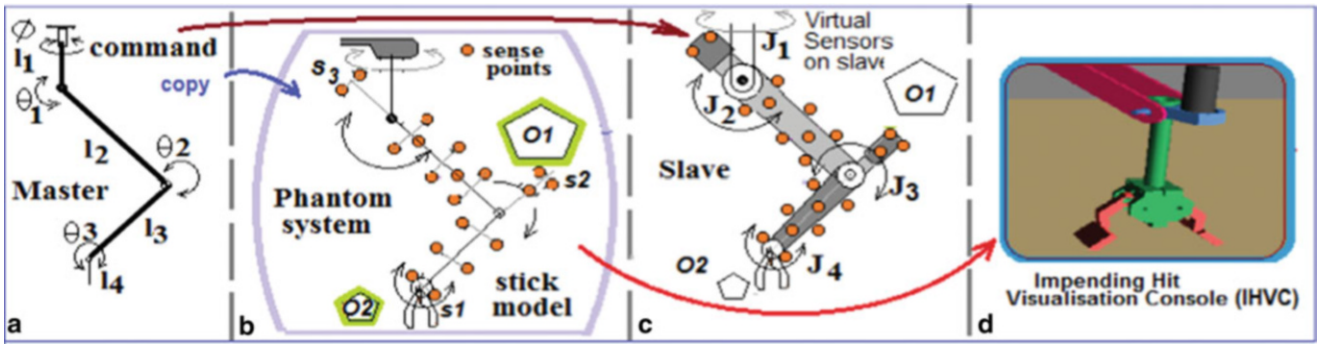


Figure 5 On-line susceptibility assessment framework a; master c; slave b; phantom

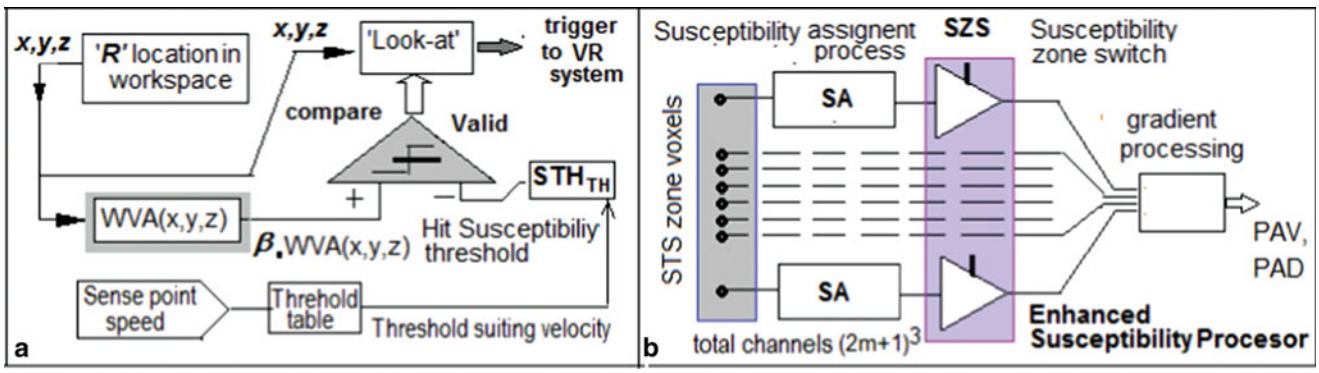


Figure 6 a: render trigger to VR system. b: Enhanced probe function - PAV and PAD

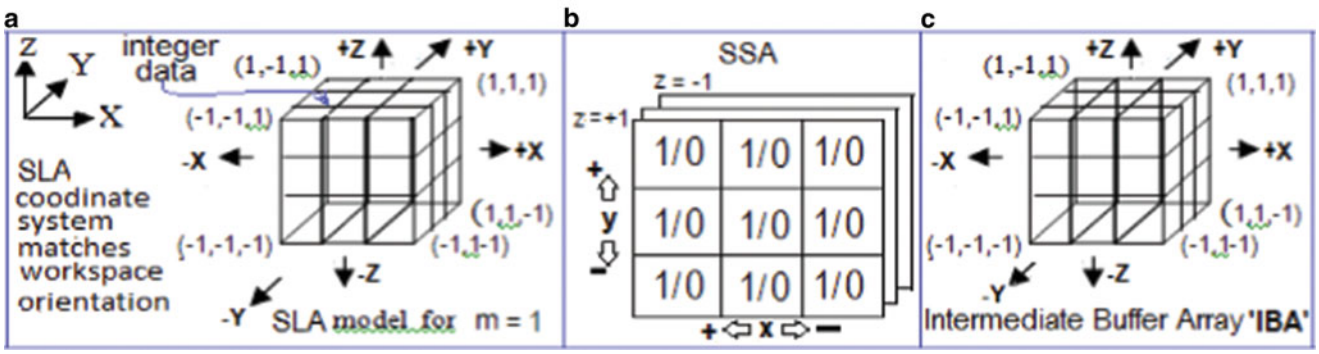


Figure 7 Data organisation for SLA and cell data

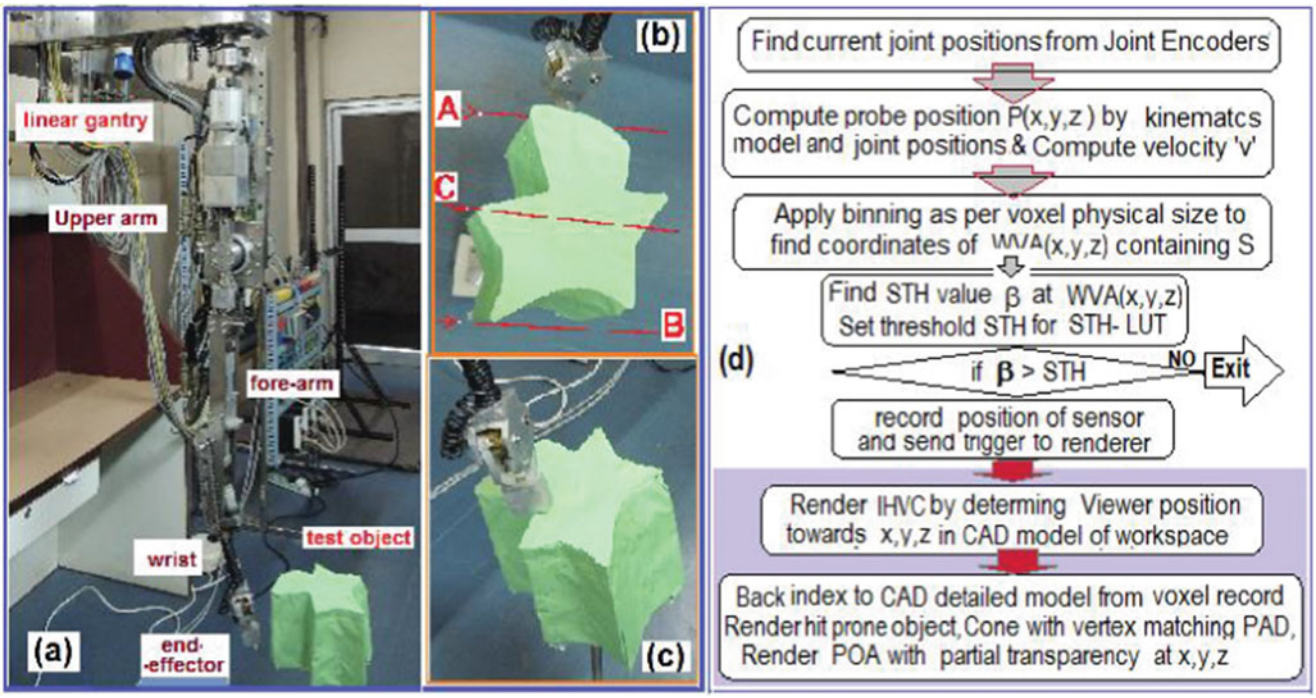


Figure 8 **a**: Test set up- with 6 DOF articulated robot hanging from gantry of 1 DOF, **b**: Robot end effector passing by side c: end-effector passing over object **d**:Susceptibility triggered view synthesis process as peripheral view in test runs. Shaded part VR module

senor grid set around a center voxel of SLA(i,j,k) (fig.7). It has 2 m + 1 layers of (2 m + 1) long, (2 m + 1) wide grid of cells. The 3D SLA grid is indexed by indices i,j,k varying in range as per factor 'm' as analogous to FZ. The arrangement ensures that each SLA cell is located in separate voxel if center location SZ(i = 0, j = 0, k = 0) is located at a VWS location. A Susceptibility zone sensing Array 'SZS' is also similarly formed but it holds 1 or 0 binary value. A '0' disables corresponding voxel from contributing in process.

Principal Approach Detection 'PAD': We define two 3D arrays, max filter $MF_{(i,j,k)}$ and minimum filter $LF_{(i,j,k)}$ as arrays of same size as $SSA_{(i,j,k)}$ for holding binary values 1/0. First neighbouring susceptibility data from WVA(x,y,z) around present ***R*** location x,y,z is buffered in IBA -

For(-m<=i>=m, -m<=j>=m, -m<=k>=m),

$SLA_{(i,j,k)} = WVA\ (x+i, y+j, z+k)$, $IBA_{(i,j,k)} = SLA_{(i,j,k)} * SSA_{(i,j,k)}$:

Then, within the shaped sensor zone the max value of β say $\boldsymbol{\beta_h}$ is found by scanning entire $IBA_{(i,j,k)}$. Occurrence of $\boldsymbol{\beta_h}$ in the SLA zone is marked in $MF_{(i,j,k)}$

For(-m<=i>=m, -m<=j>=m, -m<=k>=m),

if $IBA_{(i,j,k)} = \boldsymbol{\beta_h}$ then $MF_{(i,j,k)} = 1$ else = 0.

Centre location P_h of $\boldsymbol{\beta_h}$ zone, is found by –

For(-m<=i>=m, -m<=j>=m, -m<=k>=m),

If $MF_{(i,j,k)} = 1$ then $\boldsymbol{\Sigma i} = +i$, $\boldsymbol{\Sigma j} = +j$, $\boldsymbol{\Sigma k} = +k$, ***count*** = +1;

Finally using **ct** for count ; X. $P_h = \boldsymbol{\Sigma i}$ ***/ct***, Y. $P_h = \boldsymbol{\Sigma j}$ ***/ct*** and Z.$P_h = \boldsymbol{\Sigma k}$ ***/ct***

Similarly if we designate minimum value of β say $\boldsymbol{\beta_l}$ and form $LF_{(i,j,k)}$ just as above, then by replacing $\boldsymbol{\beta_h}$ by $\boldsymbol{\beta_l}$ and $MF_{(i,j,k)}$ by $LF_{(i,j,k)}$), use of same method gives X. P_l, Y. P_l, and Z. P_l. The vector connecting $\mathbf{P_l}$ to $\mathbf{P_h}$ is principal approach vector '**PAV**', *and* its direction is '**PAD**'

5.4 Impending Hit View Console 'IHVC'

A 3D graphics rendering object 'GRO' has been built using *Microsoft Direct –X ver.10*. [8]. It renders all the objects of workspace by treating them as '*in-situ*' CAD models[7]. It treats robot link objects through robot kinematics-model and joint value based modeller to form the suspended robot (fig.8) model. Its properties include 'camera position X_c,Y_c,Z_c' and 'look at point' X_L,Y_L,Z_L in virtual space. Other properties are lighting and magnification. The VR environment in our experiment comprises a multi-view window using multiple instances of 'GRO'. Normally the operator chooses 2 windows (fig 12a, 12b) for forming views of his choice mostly aiming at manipulation zone. By the current development a third window 'c' is introduced which is autonomous hit zone renderer.

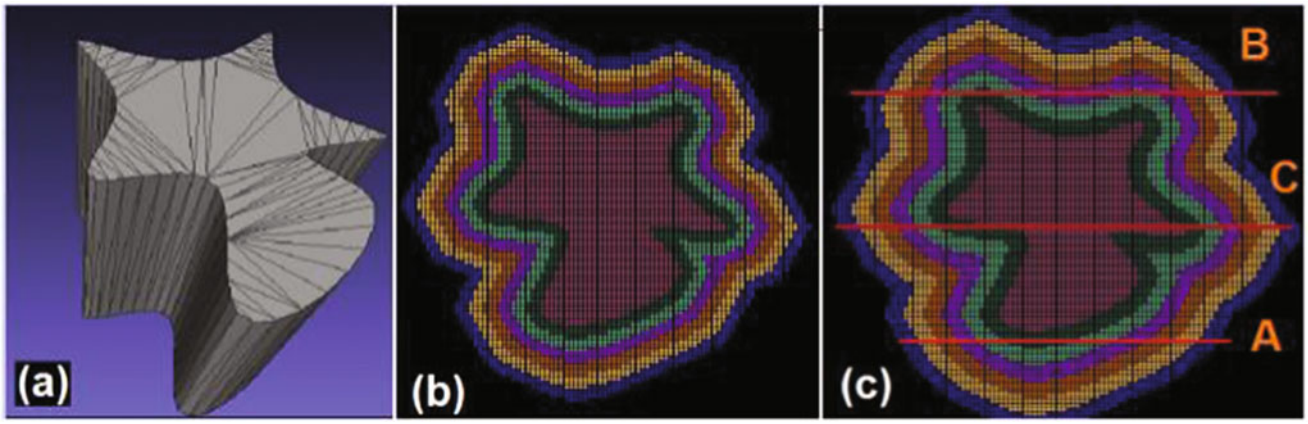


Figure 9 a:complex cylinder with concavities **b**:STS built with spherical forming element FZ of m = 3, and **c**: m = 4 sized FZ. Mark the better result in **c**.

6 Performance Of The developed Technique

The test set-up comprises an articulated arm (fig.8) Virtual sensors are attached on arm links (fig.5c). For brevity kinematics computation used in test is not included. Results are described for dynamic positions of S1 to S3(fig 5b). Voxelized workspace WVA is formed using 6 STS with $\boldsymbol{\beta}_{(n)}$ as 32,25,20,17,15,13 for n = 1 to 6 respectively. Effect of FZ size and SE on STS formation have been observed. Resulting 'susceptibility seen by probe on robot body '***R***' is also presented. Effectiveness of trigger is tested. Two aspects of the techniques are tested for performance a: trueness of STS formation around objects b: trigger generation for a moving '***R***' in workspace by them using sensor on robot body (fig5b).

6.1 Voxel conversion test:

Several objects in STL-CAD form were converted to voxel form. Asymmetric and concave nature usually cause errors and so a wedge (fig2e) and complex cylinder object (fig 8 c) with such nature are used for tests.

6.2 Correctness of STS forming algorithm:

Experimental results with spherical SE : Performance of the approach is shown on complex cylinder for q = 6 i.e. 6 STS. In $STS_{(1)}$, difference is not appreciable but for $STS_{(6)}$ improvement is substantial (fig.9b,9c). Results rendered in 3D form (fig. 10) prove the effectiveness of theSTS forming processes for complex shaped objects in true 3D space.

Susceptibilty trigger test on complex cylinder: Use of m = 3 and m = 4 i.e. sphere in FZ 7x7x7 and 9x9x9, improved trueness of STS shapes (fig.9). The β values in $STS_{(1)}$ to $STS_{(6)}$ are 32, 20 13,11,10,9, value in object body is set 100. For moves along paths A,B,C (fig.8b,9c). STH was assessed on-line along end-effector locus (S1). On path C above object and D (still higher) STH parameter 'β' is constant low(fig 11c,d), matching susceptibility behaviour (fig 10c).

Sensitivity to speed was tested by computing linear speed '*V*' of ***R*** using joint velocities and kinematics model of the robot. Lower STH (fig. 6a) at higher *V* changes susceptibility characteristic of an object (fig. 11a). A velocity based look-up table for STH effectively caused earlier trigger to VR system at higher speed. It enables user to see impending hit from relatively more distance.

Test of VR creation on impending hit viewing console IHVC Remote robot was interactively operated for grasping object using windows Fig12 a, b chosen by operator to visualise state around manipulated part, window 'd' is based on 'hit-susceptibility' and operates autonomously using STH trigger process It is very effective for showing any hit condition in the workspace for large robots.

6.3 Use of 'SPZ' and principal approach vector 'PAV' function:

For the locus 'A' the speed based STH trigger threshold was 25(fig 11a). The system invoked VR module to form graphic scenario (fig 13 b1 to b4). The cone points along **PAV** *viz a viz* the object (fig 13). The precision representation of a plane orthogonal to PAV approach vector and passing through trigger location is called '**POA**'. It, is formed using PAD as normal for POA. The POA virtual plane forms container plane for view vector in synthesizing lateral view of the near contact conditions at triggered location in

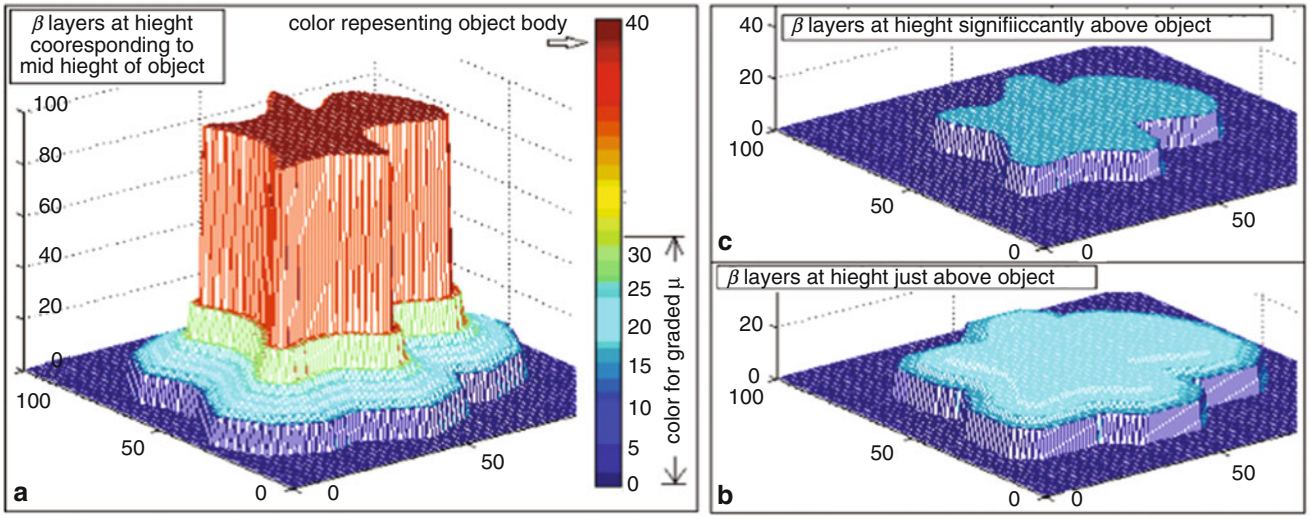


Figure 10 Susceptibility formation around the complex cylinder of figure 9a is tested by plotting the suceptibilty β of the STS layers with color coded plots after accessing them from WVA coded workspace. Red denotes complex cylinder object body. At intermediate 'Z' (e.g. fig.10a) susceptibility rise is clearly visible. Above object top (fig 10.b), middle order susceptible layer exits. Higher above object, shrunken lower order STS only exists (fig.10c).

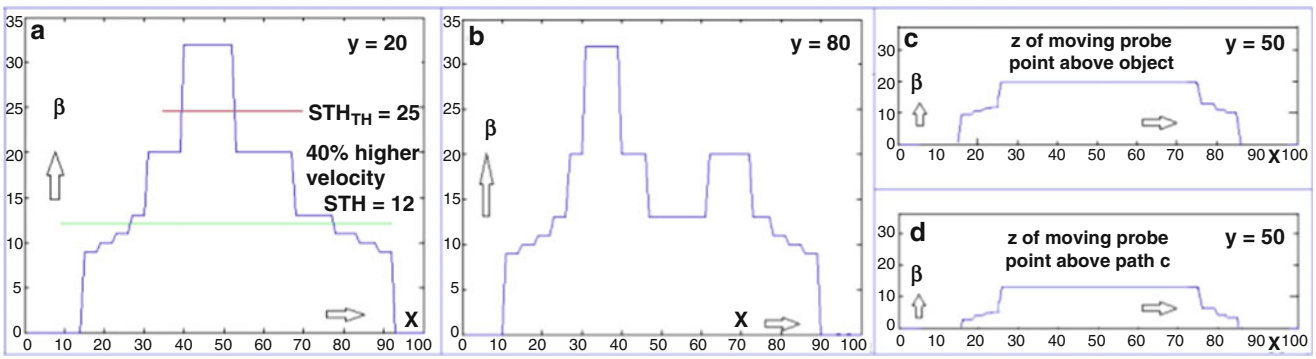


Figure 11 (**a**); Hit susceptibility seen by test robot move on gantry parallel to X axis of workspace along path 'A'(fig.8b, 9c) passes very close to object at (Y = 20). (**b**) Mark the two peaks in susceptibility seen by ***R*** moving along 'B' at Y = 80 (**c**) Susceptibility along 'C' slightly above part, (**d**): The β is lower owing more distance and spans on shorter part of the path owing to its reduction in 3D space around edges.

workspace. The orthogonal view development feature is an effective 'human machine interface' HMI feature and is a boon for remote working in densely populated workspace.

The rear extended upper arm is difficult to protect against hit but S3 (fig 5b) generated SPZ modality sensitizes operator (fig. 13 a1, a2) to imminent hit very effectively by presenting lateral side and top views synthesized using POA. The POA when graphically rendered with partial transparency attribute is very effective in perceiving immediate engulfing vicinity conditions in penetration space around object concavities (fig 13b1 to b4).

VI. Conclusion An autonomous view synthesis paradigm for compensating the lack of peripheral vision has been devised by intelligent 'See Pertinent Zone' function based on 'hit susceptibility' criteria. Though the synthesized visualization is not complete emulation of biological peripheral view, the approach unburdens the operator from 'look for hit avoidance' task, which is the main functionality of peripheral vision. It autonomously detects impending hit conditions and supports interpretation by rich visualization tools. The visible translucent POA plane is very valuable situation interpretation support offered by the innovative

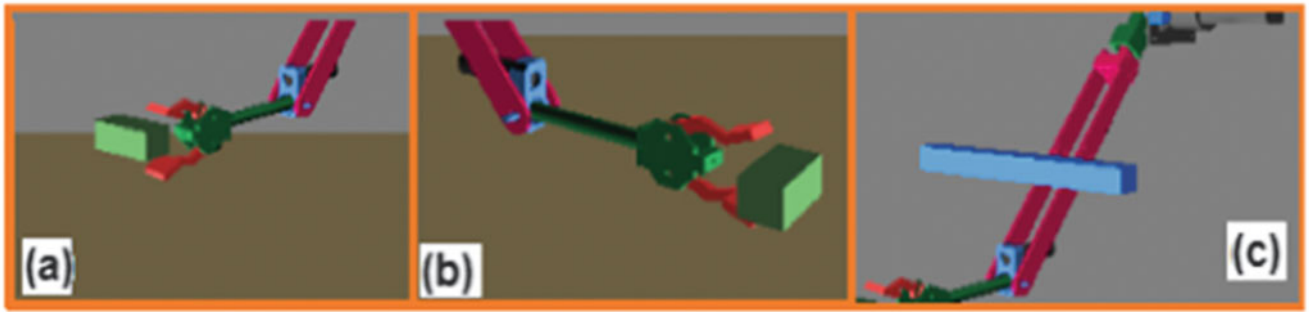


Figure 12 Console a and b are user chosen conventional mode synthesized view windows. (**c**): IHVC console shows zone around blue bar where near hit condition exists. Function of IHVC is autonomous and not chosen by operator.

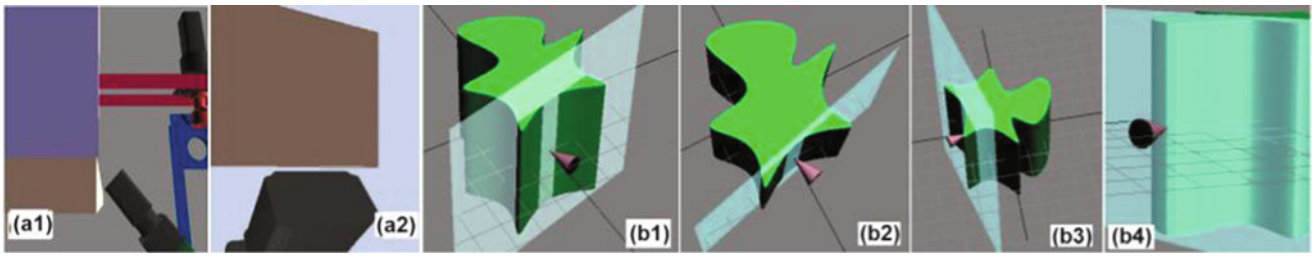


Figure 13 a1 and a2 Hit zone views autonomously created by rear probe S3 (fig. 5b) using intelligent viewer location in tangent plane to PAV. Views b1 to b4 are produced for sensor S1 and show the semitransparent POA for relative position in concavities

method. Modeling of precise shape of real workspace object and STS layers using accurate CAD models and automated STS formation from CAD models without any shape constraint are the key stones. The method is highly amenable to parallel execution as probe clusters on robot body can be assigned to separate processors for achieving real-time speed requirement for dense coverage of large robots by susceptibility sensors. By doing away with search of 'hit' the method achieves deterministic computation load that can be reliably managed in real time.

References

[1] J. K. Mukherjee et. al. "Remote managed Teleprobing for nuclear Applications" Proc.- Int. conf. PEACEFUL ATOM (nuclear instrument-pp154-169)Sept29-Oct.3, 2009 New Delhi

[2] J. K. Mukherjee et al, 'Machine Intelligence Tools for supporting metrics on active structural elements and on-line inspection in Fuel Cycle Facilities". International. Conference- SMiRT 21, New Delhi, Nov. 6-21 2011

[3] G. R. Hopkinson, A. Mohammadzadeh, "Radiation Effects In (CCD) Imagers& CMOS Active Pixel Sensors" J. High Speed Elect and Sys., Vol.14, Issue 2, June 04, ISSN: 0129-1564

[4] J. E. Speich, "Method for Simultaneously Increasing Transparency and Stability in Bilateral Telemanipulation". IEEE Int. Con. Robo. Auto., pp. 2671-76, April 2000.

[5] Rituparna mukherjee *et al* 'Intelligent Contextual Assistance for Visuals in Remote Campaign Operation' IEEE conf. Intelligent Human Comp.Interface IHCI 2012, Dec 2012 IIT Kharagpur.

[6] Oren Tropp *et al* 'A fast triangle to triangle intersection test for collision detection' Comp. Anim. Virtual Worlds, Wiley Inter Science (www.interscience.wiley.com). DOI: 10.1002/cav.115

[7] J.D. Foley *et al*. Computer Graphics: Principles and Practice (2nd Ed.), Addison–Wesley 1990 ISBN-10: 0201848406

[8] Peter J Kovach Inside Direct3D. Microsoft Press.(2000) ISBN-13: 978-0735606135

[9] 'Mental Vision: a Computer Graphics Platform for Virtual Reality, Science and Education' école Polytechnique Fédérale De Lausanne

[10] Spong, Huthinson, Vidysagar; 'Robot modeling and Control' Wiley, 2006. "Forward and inverse kinematics" page 65-100

[11] M. Lin *et al* Collision detection between geometric models: A survey. In Proc. of IMA Con.- Mathematics of Surfaces, 1998. http://www.cs.unc.edu/ dm/collision.html.

[12] Y. Zhou and S. Suri. Analysis of a bounding box heuristic for object intersection. Jour. ACM, 46 no. 6:833–857, Nov. 1999.
[13] Y. Koren, *et-al*, Potential Field Methods and Their Inherent Limitations for Mobile Robot Navigation Proc. IEEE Conf. on Robotics & Auto., California, April 7-12, 1991, pp. 1398-1404
[14] J.K. Mukherjee, "AI Based Tele-Operation Support Through Voxel Based Workspace Modeling and Automated 3D Robot Path Determination", IEEE Conf - Convergent Tech. for the Asia-Pacific Region, vol. 4 of 4, Oct.2003, pp.305-309.
[15] Kaufman, A., Shimony, E., "3D Scan-Conversion Algorithms for Voxel Based Graphics", Proceedings of the 1986 workshop on Interactive 3D graphics, pp. 45-75, 1987.
[16] H. Samel, "Design and analysis of spatial data structures in Computer graphics, Image processing and GIS", Addison Wesley, Reading M. A. 1990

Improving Multi-Panel Lamination Process Optimization using Response Surface Methodology and Neural Network

Wimalin Laosiritaworn

1 Introduction

Design of experiment (DoE) is a process of planned experiments and analyzing the result so that valid and objective conclusions are obtained [1]. DoE investigates the effect of independent variables on dependent variables through structured experiments in order to maximize information gain with fewer experiments. 2^k factorial design (k factor, each with two levels) is the most general design usually used as initial experiment design for factor screening [2].

Although experimenting at two levels (high and low) provides advantage of reducing number of experiments but it brings about the assumption of linearity in the factor effects. In two-level factorial experiment, first-order-model is fitted to the responses. In the case that the curvature presents in the response, second-order (or quadratic) response surface model has to be considered. One way to check the adequacy of the first-order-model is to use 2^k factorial design with center point. If the curvature is significant, other technique such as response surface methodology (RSM) has to be used, which means more experiments has to be carried out. Response surface methodology is the most popular optimization method [3] that utilizes simple empirical models such as low-degree polynomials to approximate the relationship between a response variable over a current region of interest [4]. However, running RSM after factorial experiments will results in more experiment, which means additional cost and potentially more process disruption.

Neural network (NN) is by definition "an interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal neuron. The processing ability of the network is stored in the inter-unit connection strengths or weights, obtained by a process of adaptation to, or learning from, a set of training patterns"[5]. NN has various applications in manufacturing and business [6, 7]. One application of NN is manufacturing process modeling where set of production process parameters (e.g., process setting) and process responses (e.g., quality characteristics) are used for NN training. There have been successful cases of using NN for manufacturing process modeling, for example superplastic forming process [8], coil baking process[9], grinding process [10], and plasma etching process[11].

Some literatures have been published regarding of NN and RSM. However most of them are comparative study. For instance, reference [12] compares NN and RSM in fermentation media optimization, while reference [13] compares both method in modeling of biochemical reaction. Both works have similar conclusion that NN is superior to RSM. Reference [14] compares NN and RSM in predicting surface roughness of molded parts. The result showed NN provide slightly more accurate model but computationally much more costly than RSM.

Even though, NN could provide better modeling results than RSM, but NN can only use for prediction not optimization. To obtain optimal result, it has to be used with other techniques such as grid search or genetic algorithm, which make the approach computationally expensive. This work proposed a method of using NN trained with factorial design experimental data and use it to predict response for RSM model. This approach, not only save time and cost of conducting experiment but also does not add too much complexity computationally. A case example of multi-panel process optimization is used to illustrate the proposed method.

The remaining of this paper is organized as follows; section 2 describes research methodology related to the use of NN for process modeling and RSM for process optimization. Section 3 provides detail results and discussion of multi-panel process optimization and finally section 4 is the conclusions.

W. Laosiritaworn (✉)
Department of Industrial Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
e-mail: wimalin@hotmail.com

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_33,

2 Research Methodology

This research started with training neural network to model multi-panel lamination process with data from 2^3 full factorial designs. After training, it was used to predict response for RSM using Central composite design (CCD). Therefore, research methodology is divided into 3 parts, which are multi-panel lamination process optimization with factorial design, neural network and response surface.

2.1 Multi-panel lamination process optimization with factorial design

Multi-panel lamination is a process used for manufacturing flexible print circuit (FPC) in a case study company, who is a manufacturer of FPC as a component for hard disk drive actuator. The company was experiencing high excessive adhesive squeeze out defect in their multi-panel lamination process and decided to use factorial experiment with center point to determine appropriate parameter setting that could bring the defect rate down. The detail and the result of these experiments can be found in reference [15]. To sum up, factorial design was used with three factors at 2 levels, 3 center points and 2 replicates. The three factors affecting defect rate are prebake time, lamination pressure and type of film. The first two factors are numeric; therefore their center points can be calculated. Type of film, on the other hand is categorical type and has no center point. The total number of experiment of the 2^3 design with center point was 22 runs. Input factors and their setting can be found in Table 1. The response of multi-panel lamination process is defect percentage. ANOVA results confirm the significant of curvature, therefore more experimental data has to be obtained to fit higher order model. Instead of conducting actual experiment neural network was used to construct lamination process model. The method of model construction is described in the next section.

2.2 Neural network

A Multi-layer neural network trained with back-propagation algorithm is the most extensively adopted network among many types of neural network[16] and is the type of NN used in this work. Neural network is constructed by connecting simple processing elements or neurons. These processing elements are connected together and the strength of the connection is quantified by 'weight'. By adapting a set of weight, NN can be used to learn relationship between set of inputs and outputs.

In multi-layer neural network, neurons are located in input layer, hidden layer, or output layer (Fig. 1). Input layer takes data from the outside of neural network. In this work, NN is use with data obtained from 2^3 Full factorial design (3 input factors), therefore, input layer has 3 nodes which stands for prebake time, lamination pressure and type of film. Hidden layer is where the actual calculation is done through 'weight' alteration. In this work, number of appropriate hidden layers and hidden nodes were obtained through exhaustive search. Output layer is where the prediction answer is sent back outside the network. There is only one response in this case, which is defect rate, therefore only one neuron in the output layer.

After NN architecture was set up, the next step is to train it with back-propagation algorithm [17]. Firstly, 'forward pass' was performed by fed inputs through the network and calculate weighted sum (Sj) for every neuron.

$$S_j = \sum_i a_i w_{ij} \tag{1}$$

Where a_i is the activation level of unit i, and w_{ij} is the weight from unit i to unit j. Then, the logistic transfer function, i.e. $g(x) = \frac{1}{1+e^{-x}}$ where $x = S_j$, were applied to the output. Then, $g(x = S_j)$ becomes the output of unit j, and the same procedure repeats for all neurons.

After that, the Back-Propagation performs a 'backward pass', where the error is calculated to update (adjust) the weight for each neuron for the output layer using the following equation.

Table 1 Factor setting for 2^3 Full factorial design of multi-panel lamination process

		Level setting		
Factor	Symbol	Low	Center point	High
Prebake time (minute)	A	0	20	40
Lamination pressure (Klb)	B	310	370	430
Type of film	C	Dahlar	None	Sekisui

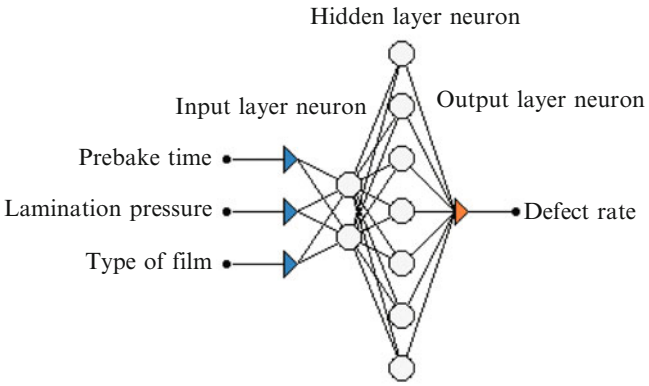


Fig. 1 Neural network architecture of multi-panel lamination process model

Table 2 Central composite design for muli-panel lamination process optimization

	Coded Variables		Natural Variables		Responses	
Standard Order	Prebake time (x_1)	Lamination pressure (x_2)	Prebake time (ξ_1)	Lamination pressure (ξ_2)	y_1(Dahlar)	y_2(Sekisui)
1	-1.00000	-1.00000	5.8579	327.574	17.1160	0.246720
2	1.00000	-1.00000	34.1421	327.574	3.9170	0.059984
3	-1.00000	1.00000	5.8579	412.426	26.8429	0.262879
4	1.00000	1.00000	34.1421	412.426	4.6086	0.124765
5	-1.41421	0.00000	0.0000	370.000	26.8490	0.262654
6	1.41421	0.00000	40.0000	370.000	3.9376	0.059925
7	0.00000	-1.41421	20.0000	310.000	4.5899	0.125474
8	0.00000	1.41421	20.0000	430.000	17.0066	0.246762
9	0.00000	0.00000	20.0000	370.000	7.2094	0.204501
10	0.00000	0.00000	20.0000	370.000	7.2094	0.204501
11	0.00000	0.00000	20.0000	370.000	7.2094	0.204501
12	0.00000	0.00000	20.0000	370.000	7.2094	0.204501
13	0.00000	0.00000	20.0000	370.000	7.2094	0.204501

$$\delta_j = (t_j - a_j)g'(S_j) \tag{2}$$

For hidden layer equation 3 is used

$$\delta_j = \left[\sum_k \delta_k w_{kj}\right] g'(S_j) \tag{3}$$

In these equations, t_j is the target value for unit j, a_j is the output value for unit j, $g'(x)$ is the derivative of the logistic function g and S_j is weighted sum of inputs to j. Then, the weight adjustment is calculated as $\Delta w_{ji} = \eta \delta_j a_i$ where η is the learning rate. These forward and backward processes repeat with new input vector until stopping criteria are met.

2.3 Response surface methodology

A central composite design (CCD) is used in designing response surface experiments. Two response surfaces were plot separately for each type of film (Dahlar and Sekisui) using the design in Table 2. Factor setting in Table 2 was fed to NN to obtain results for Dahlar and Sekisui film type. The response surface η is defined by

$$\eta = f(x_1, x_2) \tag{4}$$

Where x_1 is prebake time and x_2 is lamination pressure. As previous work suggested that there is curvature in the system[15], the second-order model is used.

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum \sum \beta_{ii} x_j x_i + \varepsilon \tag{5}$$

3 Results and discussion

3.1 Neural network modeling

NN modeling was done using 'Alyuda NeuroIntelligence' software. The modeling was divided into 4 phases which were data analysis, data preprocessing, NN architecture design and NN training.

Data analysis. Training data was partitioned into 3 groups for training, validation and testing. The training set was used in the training process for weight adjusting. The validation set was used to tune network parameters other than weight. For example, it was used to detect number of hidden units when the network performance became worse. This was to prevent 'overtraining' problem. Finally, the testing set was used to test the trained network on its performance with the 'unseen' data. In this works, the 22 available records were divided for training, validation and testing at 16, 3 and 3 records respectively.

Data preprocessing. Before feeding numerical data into NN, they have to be scaled into the same rage. Scaling was done to prevent the errors due to higher valued variable to have a greater effect than those with lower magnitude [18]. Input data were scaled to the range of [-1,1] using the following formula.

$$SF = \frac{(SR_{\max} - SR_{\min})}{(X_{\max} - X_{\min})} \tag{6}$$

$$X_p = SR_{\min} + (X - X_{\min}) \times SF \tag{7}$$

Table 3 Input and output setting and values for NN training

Factor	*Type*	*Format*	*Min*	*Max*	*Scaling range*	*Scaling Factor*
A	Input	Numeric	0	40	[-1,1]	0.005
B	Input	Numeric	310	430	[-1,1]	0.016667
C	Input	Category	n/a	n/a	[-1,1]	n/a
Y	Output	Numeric	0	30.56	[0,1]	0.032723

Table 4 Accuracy measurements of NN model

	Data set			
Measurement	Training	Validation	Testing	Overall
Mean absolute error	0.025693	0.009251	0.008032	0.021043
Correlation of coefficient	0.987177	0.988209	0.977017	0.987957

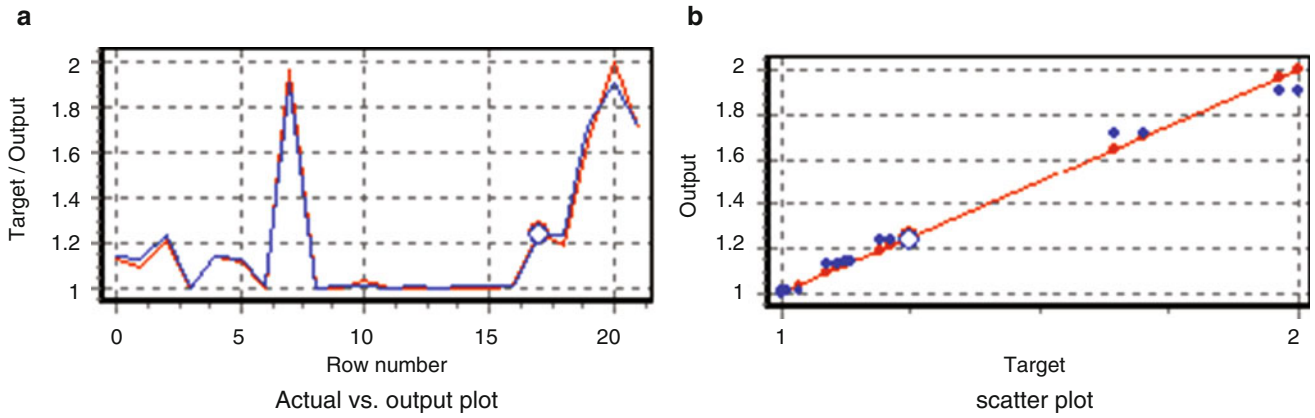


Fig. 2 NN training result

Where X is actual value of a considered numeric column, X_{min} is the minimum actual value of the column, X_{max} is maximum actual value of the column, SR_{min} is the lower scaling range limit, SR_{max} is the upper scaling range limit, SF is the scaling factor, and X_p is the preprocessed value. Scaling factor (SF) used was summarized in Table 3.

Output data was scaled to the range of [0,1] using logistic activation function.

$$F(x) = \frac{1}{(1+e^{-x})} \tag{8}$$

Neural network architecture design. Architecture design refers to number of hidden layer and hidden nodes. Exhaustive search was used to identify appropriate architecture. A search was carried out for up to 2 layers and up to 8 neurons in each layer. The best architecture found was the one with two hidden layers, 2 neurons in the first hidden layer and 7 neurons in the second hidden layer (Fig. 1).

Neural network training. After NN architecture was identified, NN training was implemented using quick propagation algorithm. Mean absolute error (MAE) and the correlation coefficient were used to measure modeling accuracy with the following formula

$$MAE = \frac{\sum_{i=1}^{n} |f_i - y_i|}{n} \tag{9}$$

$$r = \frac{n\sum_{i=1}^{n} f_i y_i - \left(\sum_{i=1}^{n} f_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\left(\sum_{i=1}^{n} f_i^2\right) - \left(\sum_{i=1}^{n} f_i\right)^2}\sqrt{n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2}} \tag{10}$$

where f_i is the prediction from neural network for record i, y_i is the actual value for record i, n is the total number of data. MAE and r of the trained network can be found in Table 4. As MAE achieved are small and r approaching 1, it can be concluded that the model has high accuracy.

In order to review model adequacy graphically, actual vs. output plot (Fig. 2 a) and scatter plot (Fig. 2 b) were drawn. Actual vs. output plot indicates good match between predicted results and actual results. Scatter plot forms straight line with little deviation from target, which confirms the adequacy of the model.

a) Actual vs. output plot b) scatter plot

3.2 Central composite design

The NN model obtains previously was used to predict responses of CCD design. The result is shown in Table 2. Minitab software was used for data analysis. ANOVA results of Dahlar and Sekisui film are shown in Table 5 and Table 6 respectively. For both film types, it can be concluded

that both factors and also their interactions are significant at 95 % confidence as they are all having P value less than 0.05.

Fig. 3 shows surface plot of defect rate for Dahlar film (Fig. 3 a) and Sekisui film (Fig. 3 b). Defect rate has the lower the better characteristic and Sekisui film has shown much better defect rate than Dahlar film. Minitab's response analyzer was used to find optimum setting of Sekisui film. Optimum setting obtained for Dahlar film results in 6.5876 % defect at prebake time 40 minutes and lamination pressure of 310 Klb while optimum setting for Sekisui results in 0.0009 % defect at prebake time 36.77 minutes and lamination pressure 310.51Klb. Therefore, Sekisui film type and their optimum setting were chosen. This new optimum solution was better than optimum solution found in reference[15] at 1.080 % defect percentage at prebake time 40 minutes and lamination pressure 430 Klb.

Table 5 Analysis of variance result for Dahlar film type

Analysis of Variance for dahlar

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	5	821.645	821.645	164.329	152.87	0.000
Linear	2	673.046	673.046	336.523	313.05	0.000
A	1	575.198	575.198	575.198	535.08	0.000
B	1	97.848	97.848	97.848	91.02	0.000
Square	2	128.189	128.189	64.095	59.62	0.000
A*A	1	105.631	116.843	116.843	108.69	0.000
B*B	1	22.559	22.559	22.559	20.99	0.003
Interaction	1	20.409	20.409	20.409	18.99	0.003
A*B	1	20.409	20.409	20.409	18.99	0.003
Residual Error	7	7.525	7.525	1.075		
Lack-of-Fit	3	7.525	7.525	2.508	*	*
Pure Error	4	0.000	0.000	0.000		
Total	12	829.169				

Table 6 Analysis of variance result for Sekisui film type

Analysis of Variance for Sekisui

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	5	0.058854	0.058854	0.011771	68.23	0.000
Linear	2	0.054717	0.054717	0.027358	158.58	0.000
A	1	0.046749	0.046749	0.046749	270.97	0.000
B	1	0.007967	0.007967	0.007967	46.18	0.000
Square	2	0.003546	0.003546	0.001773	10.28	0.008
A*A	1	0.002954	0.003256	0.003256	18.87	0.003
B*B	1	0.000591	0.000591	0.000591	3.43	0.107
Interaction	1	0.000591	0.000591	0.000591	3.43	0.107
A*B	1	0.000591	0.000591	0.000591	3.43	0.107
Residual Error	7	0.001208	0.001208	0.000173		
Lack-of-Fit	3	0.001208	0.001208	0.000403	*	*
Pure Error	4	0.000000	0.000000	0.000000		
Total	12	0.060062				

4 Conclusion

This paper proposes a new method of using neural network to predict response for response surface experiments. This method is particularly useful when the screening factorial experiment was trial and curvature is confirmed. In that case, more experiment has to be carried out as more data are needed to fit higher order model. Instead of running more experiments, NN was used to train with factorial experiment data to predict results needed to fit response surface. A case study of multi-panel lamination process was used to demonstrate the proposed method. The result confirms that the optimization result obtained from response surface using data from neural network achieved better solution.

The use of NN modeling for predicting experimental result not only provide better optimization solution, but also reduce cost, time and effort of conducting additional experiments. Further work is to incorporate NN model with

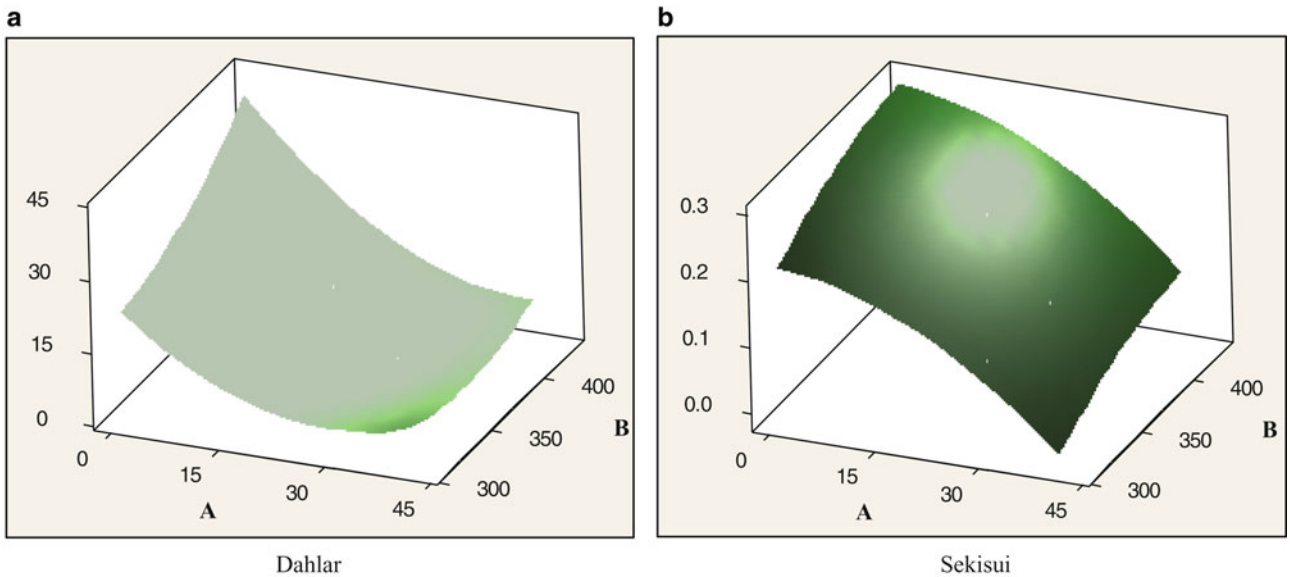


Fig. 3 Surface plot defect rate

other optimization techniques as NN model, once constructed, is not limited to predicting result for response surface, but could be used to predict response of any experimental design with at any number of levels. Also, other modeling technique for example, support vector machine, could be used instead of NN to investigate if modeling accuracy can further improve.

References

1. Montgomery, D.C.: Design and Analysis of Experiments. John Wiley & Sons, Inc., New Jersey, USA. (2009)
2. Hanrahan, G., Lu, K.: Application of Factorial and Response Surface Methodology in Modern Experimental Design and Optimization. Critical Reviews in Analytical Chemistry 36, 141-151 (2006)
3. Baş, D., Boyacı, İ.H.: Modeling and optimization I: Usability of response surface methodology. Journal of Food Engineering 78, 836-845 (2007)
4. Steinberg, D.M., Hunter, W.G.: Experimental Design: Review and Comment. Technometrics 26, 71-97 (1984)
5. Gurney, K.: An Introduction to Neural Networks. UCL Press Limited, London (1997)
6. Vellido, A., Lisboa, P.J.G., Vaughan, J.: Neural networks in business: a survey of applications (1992–1998). Expert Systems with Applications 17, 51-70 (1999)
7. Zhang, H.C., Huang, S.H.: Applications of neural networks in manufacturing: a state-of-the-art survey. International Journal of Production Research 33, 705-728 (1995)
8. Sukthomya, W., Tannock, J.: The training of neural networks to model manufacturing processes. Journal of Intelligent Manufacturing 16, 39-51 (2005)
9. Laosiritaworn, W.S.: Coil Baking Process Modeling with Neural Network. In: IEEE International Conference on Industrial Engineering and Engineering Management, pp. 1656-1660. (Year)
10. Liao, T.W., Chen, L.J.: Manufacturing Process Modeling and Optimization Based on Multi-Layer Perceptron Network. Journal of Manufacturing Science and Engineering 120, 109-119 (1998)
11. Rietman, E.A., Lory, E.R.: Use of neural networks in modeling semiconductor manufacturing processes: an example for plasma etch modeling. Semiconductor Manufacturing, IEEE Transactions on 6, 343-347 (1993)
12. Desai, K.M., Survase, S.A., Saudagar, P.S., Lele, S.S., Singhal, R. S.: Comparison of artificial neural network (ANN) and response surface methodology (RSM) in fermentation media optimization: Case study of fermentative production of scleroglucan. Biochemical Engineering Journal 41, 266-273 (2008)
13. Baş, D., Boyacı, İ.H.: Modeling and optimization II: Comparison of estimation capabilities of response surface methodology with artificial neural networks in a biochemical reaction. Journal of Food Engineering 78, 846-854 (2007)
14. Erzurumlu, T., Oktem, H.: Comparison of response surface model with neural network in determining the surface quality of moulded parts. Materials & Design 28, 459-465 (2007)
15. Laosiritaworn, W.S., Aoonchan, P.: Multi-Panel Lamination Process Optimization with Design of Experiment. In: World congress on Engineering and Computer Science, pp. 1383-1387. International Association of Engineers, (Year)
16. Nascimento, C.A.O., Giudici, R., Guardani, R.: Neural network based approach for optimization of industrial chemical processes. Computers & Chemical Engineering 24, 2303-2314 (2000)
17. Dayhoff, J.E.: Neural Network Architectures: An Introduction. Van Nostrand Reinhold, New York, USA (1990)
18. Swingler, K.: Applying Neural Network: A Practical Guide. Academic Press Limited, London, UK. (1996)

Selecting right questions with Restricted Boltzmann Machines

Maciej Zięba, Jakub M. Tomczak, and Krzysztof Brzostowski

1 Introduction

The problem of selecting the proper set of questions was initially considered as a part of 20 questions game [5]. In this game one person thinks about one word and the other has to discover the concept by asking as few questions as possible[1]. In the typical game we consider only yes/no answers, but there are some extensions that allows to specify some uncertainty level. There are couple of projects that enable to play the 20 questions game including *Akinator* [6] (the goal is to guess the name of a famous character), *20Q A.I.* [2] (refers to the general objects) or *Winston* [1] (aims at discovering animals).

The issue of selecting proper question basing on the previous answers plays an important role in the process of information retrieval on the Internet [4, 11]. While searching for some information the user is obliged to type the proper keywords in the browser. In the case of using imprecise keywords or forgetting some of them, results returned by a browser may be far from the desired. In such situation the set of few supporting questions may be asked to the user by the system to gather the missing data and return reasonable results. Let us consider the example described by authors of [11]. Imagine that the user saw a suricate in a ZOO and after returning home he wants to learn more about it on the Web. The problem is he cannot recall its name and any keywords he can think of the animal are only related to its appearance. He could spend much time adjusting those keywords and clicking result links forth and back, or scrolling image results until he comes across some information about suricates. By questioning the user about the features of the animal the proper name could be recalled faster.

The domain of eliciting business requirements is another application field in which the problem of asking right questions is considered [14]. The role of the business analyst is to gather relevant information about the product or service from a customer during the requirements specification. The sequence of the asked questions during the interview with customer is essential to understand the needs of business partner and to adjust the product or service to the stated requirements. The analyst interviews the customer according to the predefined procedure having under consideration numerous scenarios that may occur during the conversation. However, he is not prepared for all possible cases and usually is not familiar with all features of the potential product or service. Therefore, the system for supporting the analyst by recommending the questions to be asked could improve the process of eliciting business requirements. Let us consider the example in which the service represents the solution of optimization problem in the transportation domain. The goal of the analyst is to ask a sequence of questions and identify the optimization problem that occurred in the transportation company. For instance, the company may have problems with scheduling bus-drivers optimally [9]. Most of the optimization problems in the transportation domain are well defined and some solutions are proposed in the literature. Moreover, each of the optimization models can be described by the set of characteristic features. As a consequence, the analyst should ask the sequence of questions basing on the features of the models to identify the optimization problem that is observed in the transportation company or conclude that there are not suitable models that meet the requirements. In order to identify the proper optimization problem the analyst should have thorough knowledge about all the optimization models in the transportation domain and should have ability to distinguish them by asking relevant questions. If the analyst is supported by some recommended questions that should be asked the process of eliciting business requirements may be

[1] Usually up to 20 questions. For that reason the game is called 20 questions game.

M. Zięba (✉) • J.M. Tomczak • K. Brzostowski
Faculty of Computer Science and Management, Wroclaw University of Technology, Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_34,

performed almost automatically and it would be sufficient for the analyst to have only basic knowledge about the optimization models.

The presented applications illustrate how important is the problem of choosing the right questions. More formally, the problem can be described in the following way. Assume that there are some concepts (e.g. products, services, animals, famous people, etc.) and each of the concept is described by the vector of common binary features. If the value of the *i*-th feature equals 1 then we say that the object has this property, and 0 – otherwise. For instance, the cat has the feature "mammal" equal 1. The main issue of the problem is to select a sequence of the features values to be questioned about that will help to discover the right concept.[2] Additionally, total number of selected features should be minimized.

Several methods are presented in the literature to solve the problem of selecting proper features to be questioned about. One of them is to represent the problem as the classification problem with the individual class label for each of the concepts, in which a classifier is constructed, e.g., *C4.5* decision tree [10]. This group of classification methods has ability to put only relevant features in the nodes, while edges represent decision variants and class labels are stored in the leafs of the tree. As a consequence, the process of discovering features is performed by starting from the root that contains the most informative feature and moving downwards on the selected path until one of the leafs is reached.

A different approach to solve the stated problem is described in [11] which makes use of semantic network together with Concept Description Vectors (CDF) [13] to search for the right questions. Each feature in the network is represented by a node and each relation between two features is described with an edge. Each edge is described by a CDF that contains two values denoting support of the connection and the confidence of knowledge. The questions are selected basing on the CDFs of the features in relation with the actually considered feature.

In this work we present slightly different approach basing on a machine learning method that applies Restricted Boltzmann Machines (RBM) to approximate the distribution over the features describing concepts, which is further used to formulate proper sequence of questions. The RBMs are widely used in many applications including: features extraction [7], classification [8] and collaborative filtering [12]. In the considered application we exploit the reconstruction abilities of RBM to identify the most relevant features. We represent the concepts by the sparse binary vectors of the features and we train the RBM in the unsupervised mode. Next we construct an evolving random process to find the most probable features to be asked for. Finally, we identify the closest concept basing on the answers to the stated questions.

The paper is organized as follows. In Section 2 the problem is stated. In Section 3 a brief introduction to RBMs is presented. Section 4 contains the description of the algorithm for relevant features selection. In Section 5 experimental studies are discussed. The paper is summarized with conclusions in Section 6.

2 Problem Statement

The considered problem aims at asking questions in order to identify unknown concept. We are interested in formulating such sequence of questions which allows to find the solution using as few questions as possible.

Let us formalize the problem in the following manner. Assume each of considered objects (concepts) is described by a vector of binary features $\mathbf{x} \in \{0, 1\}^D$. Moreover, it is possible that one concept may be described by more than one binary vector. Practically, it means that couple of former game players may have perceived the concept in different way and, as a consequence, they have given different answers for same questions. Additionally, we label each object (concept) with y, where $y \in \{1, \ldots, K\}$.

Next, by $\mathbf{x} \in \{0, 1\}^D$ denote a vector of features describing considered objects (concepts) and by $\mathbf{q} \in \{0, 1\}^D$ – a vector of questions. If one asks about the k-th feature, we set $q_k = 1$. We can think of the process of asking questions as a random process which depends on previously asked question. Such random process evolves in time because at each step of asking questions the distribution of selecting a new inquiry changes due to already given answers (discovered values of the features).

At the beginning we assume that there exists the distribution of selecting the questions to be asked:

$$p(q) = Cat(q|\pi_0) \tag{1}$$

where Cat$(\cdot)$ denotes the categorical distribution, and π_0 is a vector of initial probabilities. Next, after asking n questions we obtain a sequence of inquires $\mathcal{S} = \{q^1, \ldots, q^n\}$. For convenience we introduce an auxiliary set of indexes of already asked features $\mathcal{I}$. Then, the probability of asking $(n + 1)$ question is the following:

$$p(q_k = 1 \mid \mathcal{S}) = \begin{cases} 0, & \text{if} k \in \mathcal{I} \\ \pi_{n+1}, & \text{otherwise} \end{cases} \tag{2}$$

We make two remarks about the given random process. First, the probability distribution $p(q \mid \mathcal{S}) = Cat(q \mid \pi_{n+1})$ has more zeros for increasing n. This fact means that the

[2] In this text we associate the issue of discovering the values of features with problem of asking binary questions about the properties of the concepts.

random process of asking question is not recurrent. It makes sense because we want to avoid asking about the same thing two times.[3]

Second, the crucial quantity in (2) is the vector of probabilities after n already asked questions, π_{n+1}. In order to determine π_{n+1} we use the following strategy known in the 20 questions game. At each step we should ask about the feature which allows to split the set of concepts roughly in half. Therefore, if we are able to discover the distribution over features, we could find the most probable features and use them in the questioning process. Following this line of thinking, we use a universal approximator over binary variables, known as *Restricted Boltzmann Machine*, to determine important features.

Apart from the process of asking questions, we need to have some procedure for determining the stopping point for questioning. This can be done if the probability or value of some similarity measure of some object is large enough, i.e., above fixed threshold.

3 Restricted Boltzmann Machines

Restricted Boltzmann Machine (RBM) is a bipartie Markov Random Field in which visible and hidden units can be distinguished. In RBM only connections between the units in different layers are allowed, i.e., visible to hidden units. The joint distribution of binary visible and hidden units is the Gibbs distribution:

$$p(x, h \mid \theta) = \frac{1}{Z} exp\big(- E(x, h \mid \theta). \tag{3}$$

with the following energy function:

$$E(x, h \mid \theta) = -x^\top W h - b^\top x - c^\top h, \tag{4}$$

where $\mathbf{x} \in \{0, 1\}^D$ are the visible units, $\mathbf{h} \in \{0, 1\}^M$ are the hidden units, Z is the normalizing constant dependent on θ, and $\theta = \{W, b, c\}$ is the set of parameters, namely, $W \in \mathbb{R}^{D \times M}$, $b \in \mathbb{R}^D$, $c \in \mathbb{R}^M$ are the weight matrix, visible and hidden bias vectors, respectively.

Since there are no connections among the units within the same layer, i.e., no visible to visible, or hidden to hidden connection, the visible units are conditionally independent given the hidden units and vice versa:

$$p(x_i = 1 \mid h, W, b) = sigm\big(W_{i\cdot} h + b_i. \tag{5}$$

$$p(h_j = 1 \mid x, W, c) = sigm\big((W_{\cdot j})^\top x + c_j. \tag{6}$$

where $sigm(a) = \frac{1}{1+exp(-a)}$ is the sigmoid function, $\mathbf{W}_{i\cdot}$ is the i-th row of the weight matrix, and $\mathbf{W}_{\cdot j}$ is the j-th column of the weight matrix.

Unfortunately, in order to learn parameters θ gradient-based optimization methods cannot be directly applied because exact gradient calculation is intractable. Fortunately, we can adopt *Constrastive Divergence* algorithm which approximates exact gradient using sampling methods [7].

4 The algorithm for relevant features selection

We have access to sequences of questions describing the concepts, i.e., there is the training data $\mathcal{Y}_N = \{x_n, y_n\}_{n=1}^N$, where $\mathbf{x}_n$ represents the vector of features for the n-th object in training data, y_n is the label that identifies one of the possible concepts. Let us denote the unlabelled data with $\mathcal{D}_N = \{x_n\}_{n=1}^N$ and the set of corresponding labels by $\mathcal{Y}_N = \{y_n\}_{n=1}^N$.

The procedure of identifying the relevant features is given in Algorithm 1. The procedure is initialized by creating empty sequence $\mathcal{I}$ of features indexes selected for questioning and vector $\mathbf{v}_q$ representing discovered values of the features (initially $\mathbf{v}_q$ contains zeros). In the first step values of the model parameters $\mathbf{W}$, $\mathbf{b}$ and $\mathbf{c}$ are calculated in the process of training RBM on unlabelled data $\mathcal{D}_N$. Next, we apply one step of Gibbs sampling to generate a random vector $\mathbf{h}_q$ of hidden values from distribution $p(h_m = 1 \mid \mathbf{v}_d, \mathbf{W}, \mathbf{c})$. Further, we construct vector $\boldsymbol{\pi}$ that contains probability values $p(v_i = 1 \mid \mathbf{h}_d, \mathbf{W}, \mathbf{b})$ in such fashion that each element π_i has assigned probability value of activating visible unit on i-th position. We take under consideration only the units that corresponds to the unasked features (the features that have not been included in $\mathcal{I}$ yet). For the asked features (current members of $\mathcal{I}$) the corresponding value of vector $\boldsymbol{\pi}$ is equal 0. We normalize the vector $\boldsymbol{\pi}$ by dividing each of the components by the sum of the elements.[4] Finally, we sample the position of the feature to be asked j_q from the categorical distribution with vector of parameters $\boldsymbol{\pi}$, so the probability of selecting j-th feature is equal π_j. Practically, it means that instead of sampling vector of visible units from distribution $p(v_j = 1 \mid \mathbf{h}_q, \mathbf{W}, \mathbf{b})$ we are interested in sampling among the vectors that has only one activated unit (j_q is the index of the unit for the activation). For all of the selected features the probability π_i is equal zero so there is no risk for selecting the

[3] In some cases asking about the same thing more than once might be necessary, e.g., when a question is ambiguous and the answer cannot be correctly answered. However, we leave this issue for further research.

[4] Each of the elements of vector $\boldsymbol{\pi}$ is greater or equal 0.

same feature twice. The sequence $\mathcal{I}$ of selected features is updated as long as the vector $\mathbf{v}_q$ is properly modified with the discovered value.

The issue of selecting the right concept basing on the given answers is performed by application of *K-means* method. In general, the procedure can be described as follows:

1. Calculate the average vector $\widehat{x}_k$ for each of the concepts ($k \in \{1, \ldots, K\}$) using data $\mathcal{D}_N$.
2. Calculate the distances between each vector $\widehat{x}_k$ and $\mathbf{v}_q$ considering only features with indexes from $\mathcal{I}$.
3. Select concept k with the smallest distance between $\mathbf{v}_q$ and $\widehat{x}_k$.

5 The experimental studies

In this section we aim at evaluating empirically the performance of the proposed method of selecting relevant features. First, we take under consideration well-known *Zoo* dataset available in UCI Machine Learning Repository [3]. Each of the instances in the considered dataset represents one of the 101 animal species. Each animal is described by a vector of 28 binary features.[5] The dataset was initially used to solve the classification problem of animal type assignment. In this section we formulate the problem of identifying the animal using as few questions about the animal as possible. Therefore, we apply our RBM-based approach to solve the stated issue.

The results of the experiment are presented in Figure 1. We present the accuracy values of correctly detected animals for the number of discovered features from 1 to 20. It is worth noticing that the accuracy of detection for all of the considered 101 animal species is equal 58% when all of the 28 features are discovered.[6] Further, it can be noticed that RBM-based approach gained the accuracy at the comparable level after discovering only 18 features. For the random selection approach (that was used as a baseline method in the experiment) the accuracy of identifying the animal after discovering 18 features was slightly over 40%. The RBM-based approach performed better than the baseline method (see Figure 1).

The second of the considered problems in the experimental studies is combined with searching for the e-mails by discovering the words that may occur in the subject or in the body of the message. Assume that you are searching for the e-mail in your mail box but you do not have any idea what keywords should be used. The intelligent system should give the sequence of the questions about the words that can be observed in the e-mail and basing on the responses should return the correct text. In this part of the experiment we make use of *DBWorld e-mails* [3] dataset

Algorithm 1: The procedure of selecting relevant features.

Input : $\mathcal{X}_N = \{\mathbf{x}_n\}_{n=1}^N$: the set of vectors of features, L: number of questions to be asked.
Output: $\mathcal{I}$: sequence of L features indexes selected for questioning, $\mathbf{v}_q$: vector of the discovered values of the features.

$\mathcal{I} \longleftarrow \emptyset$;
Estimate the parameters $\mathbf{W}$, $\mathbf{b}$ and $\mathbf{c}$ by training RBM on $\mathcal{X}_N$;
Create D-dimensional vector $\mathbf{v}_q$ that contains zeros;
for $l = 1 \rightarrow L$ **do**
  Sample $\mathbf{h}_q$ from $p(h_m = 1|\mathbf{v}_q, \mathbf{W}, \mathbf{c})$;
  Create D-dimensional vector $\tilde{\boldsymbol{\pi}}$ that contains zeros;
  for $i = 1 \rightarrow M$ **do**
    if $i \notin \mathcal{I}$ **then**
      $\tilde{\pi}_i \longleftarrow p(v_i = 1|\mathbf{h}_q, \mathbf{W}, \mathbf{b})$;
    end
  end
  $\boldsymbol{\pi} \longleftarrow \tilde{\boldsymbol{\pi}} \cdot (\sum_i \tilde{\pi}_i)^{-1}$;
  Sample $\mathbf{q} \sim \mathrm{Cat}(\mathbf{q}|\boldsymbol{\pi})$;
  Discover the value v for the j_q feature included in sampled $\mathbf{q}$, i.e., $q_{j_q} = 1$;
  Assign the value v on the j_q position of vector $\mathbf{v}_q$: $v_{q,j_q} \longleftarrow v$;
  $\mathcal{I} \longleftarrow \mathcal{I} \cup \{j_q\}$;
end

[5] Two nominal features from the initial dataset were transformed to binary attributes.

[6] It was impossible to correctly distinguish all of the species with the given set of features.

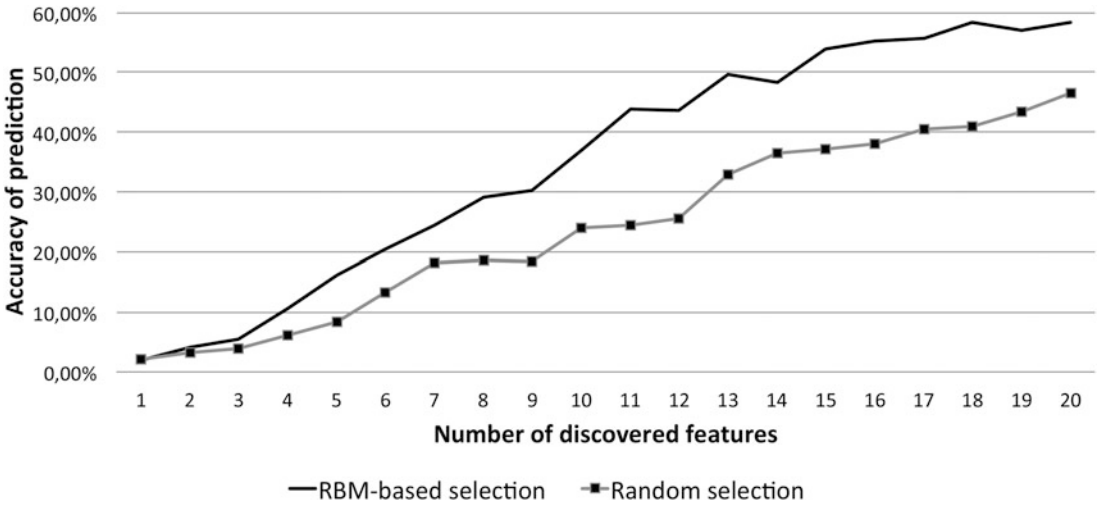


Fig. 1 The chart presents the accuracy of identifying animals for various numbers of discovered features (*Zoo* dataset). The results are presented for RBM-based approach and random selector.

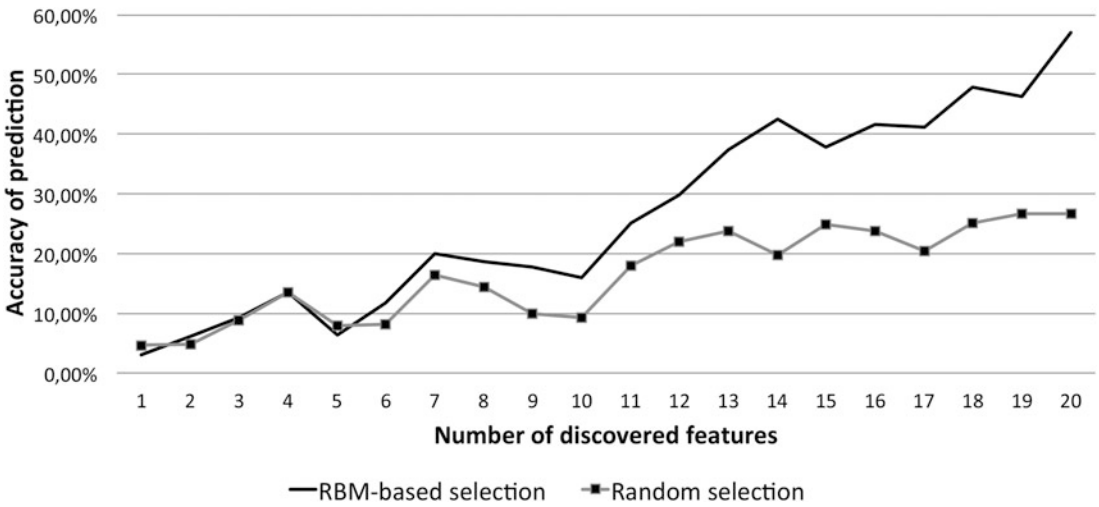


Fig. 2 The chart presents the accuracy of identifying the e-mails basing on the words taken from the subjects for various numbers of discovered features (*DBWorld e-mails* dataset). The results are presented for RBM-based approach and random selector.

which is composed of 64 e-mails, each described by 230 features that represents words occurrences in the subjects and 3722 features that represent the words included in the bodies.[7] The results for correct e-mail detection are presented in Figure 2, where we consider only 'subject' features, and in Figure 3, where only 'body' features are used. In both cases the accuracy of detecting proper e-mails by applying RBM-based question selection performed better than the random-based approach. In the case of 'body' features the accuracy of detecting the proper message was equal 90% using only 12 from 3722 features. For the comparison, the accuracy gained by random selection after discovering 12 features was slightly below 26%.

6 Conclusions

In this work we present RBM-based approach for selecting relevant features to be asked about in order to identify the considered concept. The approach is based on reconstruction capabilities of the RBM and making use of the proposed construction of the evolving random process find the most suitable question to be asked. We make two experiments to evaluate the proposed approach. The results are promising comparing to the outcomes gained by the baseline method.

[7] The dataset was initially used to detect e-mails with conference announcements.

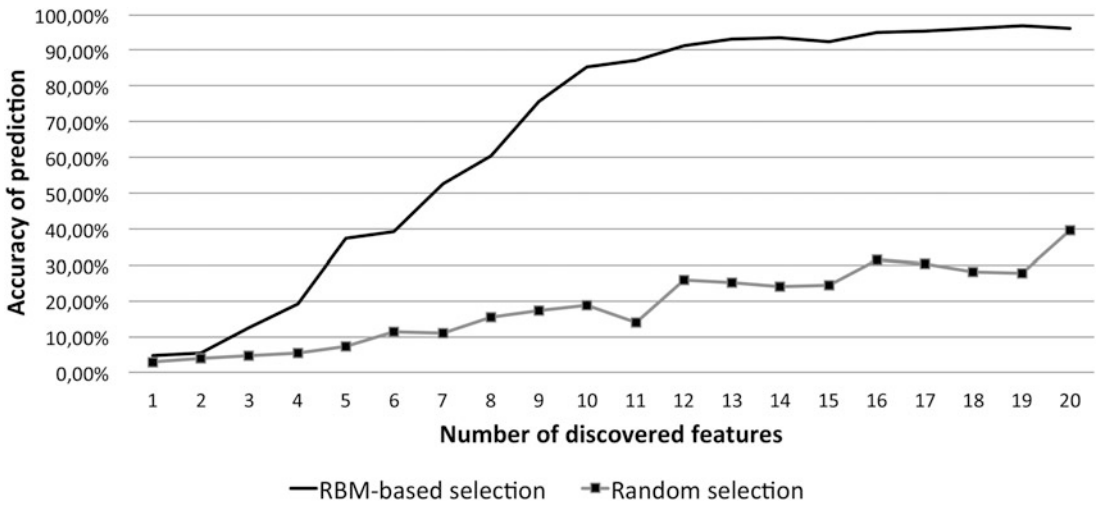


Fig. 3 The chart presents the accuracy of identifying the e-mails basing on the words taken from the body for various numbers of discovered features (*DBWorld e-mails* dataset). The results are presented for RBM-based approach and random selector

Moreover, our experiments indicate the necessity of proper sequence of questions in order to identify unknown concept.

Acknowledgments The research presented in this paper was partially supported by the European Union within the European Regional Development Fund program Number POIG.01.03.01-02-079/12.

References

1. Winston. http://kask.eti.pg.gda.pl/winston/20q.
2. 20Q.net Inc. Q20 A. I. URL http://www.20q.net/.
3. K. Bache and M. Lichman. UCI machine learning repository, 2013.
4. Włodzisław Duch and Julian Szymański. Semantic web: Asking the right questions. In *Proceedings of the 7 international conference on information and management sciences*, pages 1–8, 2008.
5. Włodzisław Duch, Julian Szymański, and Tomasz Sarnatowicz. Concept description vectors and the 20 question game. In *Intelligent Information Processing and Web Mining*, pages 41–50. Springer, 2005.
6. Elokence. Akinator. URL http://en.akinator.com/.
7. Geoffrey E Hinton. A practical guide to training restricted boltzmann machines. In *Neural Networks: Tricks of the Trade*, pages 599–619. Springer, 2012.
8. Hugo Larochelle and Yoshua Bengio. Classification using discriminative restricted boltzmann machines. In *Proceedings of the 25th international conference on Machine learning*, pages 536–543. ACM, 2008.
9. Helena R Lourenço, José P Paixão, and Rita Portugal. Multiobjective metaheuristics for the bus driver scheduling problem. *Transportation Science*, 35(3):331–343, 2001.
10. J. R. Quinlan. C4.5: Programs for machine learning. *Machine Learning*, 16:235–240, 1994. 10.1007/BF00993309.
11. Jacek Rzeniewicz, Julian Szymański, and Włodzisław Duch. Adaptive algorithm for interactive question-based search. In *Intelligent Information Processing VI*, pages 186–195. Springer, 2012.
12. Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798. ACM, 2007.
13. Julian Szymański and Włodzisław Duch. Information retrieval with semantic memory model. *Cognitive Systems Research*, 14 (1):84–100, 2012.
14. Min Wang and Yong Zeng. Asking the right questions to elicit product requirements. *International Journal of Computer Integrated Manufacturing*, 22(4):283–298, 2009.

A formal approach for identifying assurance deficits in unmanned aerial vehicle software

Adrian Groza, Ioan Alfred Letia, Anca Goron, and Sergiu Zaporojan

1 Introduction

Assuring safety in complex technical systems is a crucial issue [6] in several critical applications like air traffic control or medical devices. Safety assurance and compliance to safety standards such as DO-178B [10] may prove to be a real challenge when we deal with adaptive systems, which we consider with continuous changes and without a strict behavioral model. Traditional methods, which are mainly based on previous experiences and lessons learned from other systems are not effective in this case. Argument-based safety cases offer a plausible alternative basis for certification in these fast-moving fields [10].

Goal Structuring Notation (GSN) is a graphical notation for structured arguments used in safety applications [7]. GSN diagrams depict how individual goals are supported by specific claims and how these claims or sub-goals are supported by evidence. A GSN diagram consists of the following nodes: achieved goals, not achieved goals, context, strategy, justification, assumption, validated evidence and not validated evidence. The nodes are connected by different supporting links like: has-inference or has-evidence. To support automatic reasoning on safety cases, we formalise the GSN standard in DL.

Our solution exploits *reasoning in description logic* to identify assurance deficits in the GSN model. The identified flaws are given to a *hybrid logic-based model checker* to be validated in a given Kripke structure. All formulas were verified using the Hybrid Logic Model Checker (HLMC) [5] extended to include Next, Future and Until operators, while the reasoning in Description Logic (DL) was performed on RacerPro [8].

2 System Architecture

The solution is based on three technical instrumentations: (i) the $\mathcal{SHI}$ version of DL, (ii) the GSN standard, and (iii) hybrid logics (HLs). For the syntax, the semantics and explanation about families of description logics, the reader is referred to [2]. For the GSN graphical notation the minimum elements are introduced in section 3, while for a complete description the reader is referred to [7]. We assume also that the reader is familiar with model checking in temporal logic. However, in the following we provide specific details about HLs.

Hybrid logics extend temporal logics with special symbols that name individual states and access states by name [1]. With nominal symbols $\mathcal{N} = i_1, i_2, \ldots$ called *nominals* and $\mathcal{S}_{\text{var}} = x_1, x_2, \ldots$ called *state variables* the

A. Groza (✉) • I.A. Letia • A. Goron
Department of Computer Science, Technical University of Cluj-Napoca, Cluj-Napoca, Romania
e-mail: Adrian.Groza@mail.utm.md; Letia@mail.utm.md; Anca.Goron@mail.utm.md

S. Zaporojan
Department of Computer Science, Technical University of Moldova, Chisinau, Moldova
e-mail: zaporojan@mail.utm.md

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_35,

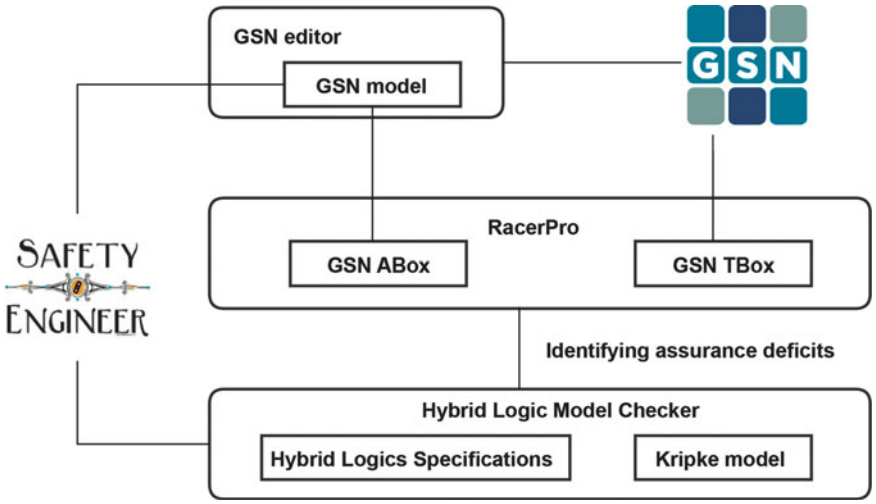


Fig. 1. System architecture

syntax of hybrid logics is $\varphi := TL \mid i \mid x \mid @_t\varphi \mid \downarrow x.\varphi \mid \exists x.\varphi$. With $i \in \mathcal{N}$, $x \in \mathcal{W}_{\text{var}}$, $t \in \mathcal{N} \cup \mathcal{W}_{\text{sym}}$, the set of *state symbols* $\mathcal{W}_{\text{sym}} = \mathcal{N} \cup \mathcal{W}_{\text{var}}$, the set of *atomic letters* $\mathcal{A}_{\text{let}} = \mathcal{P} \cup \mathcal{N}$, and the set of *atoms* $\mathcal{A} = \mathcal{P} \cup \mathcal{N} \cup \mathcal{W}_{\text{var}}$, the operators $@, \downarrow, \exists$ are called *hybrid operators*. The semantics of hybrid logic is formalized by the following statements:

$$
\begin{array}{rlll}
\mathcal{M}, g, m \vDash a & \text{iff} & m \in [V, g](a), a \in \mathcal{A} & \\
\mathcal{M}, g, m \vDash @_t\varphi & \text{iff} & \mathcal{M}, g, m' \vDash \varphi, \text{where} & [V, g](t) = m', t \in W_{sym} \\
\mathcal{M}, g, w \vDash \downarrow x.\varphi & \text{iff} & \mathcal{M}, g^x_m, w \vDash \varphi & \\
\mathcal{M}, g, m \vDash \exists x.\varphi & \text{iff} & \text{there is } m' \in \mathcal{M} & \text{such that } \mathcal{M}, g^x_{m'}, w \vDash \varphi
\end{array}
$$

The semantics, where $\mathcal{M} = \langle M, R, V \rangle$ is a Kripke structure, $m \in M$, and g is an assignment, specifies the roles of the @ operator (shifts evaluation to the state named by nominal t), the downarrow binder $\downarrow$, respectively the existential binder $\exists$, binding the state variable x to the current state, respectively to some state in the model [5]. A hybrid Kripke structure $\mathcal{M}$ consists of an infinite sequence of states $m_1, m_2, ..., R$ a family of binary accessibility relations on $\mathcal{M}$ and a valuation function L that maps ordinary propositions and nominals to the set of states in which they hold, i.e. $\mathcal{M} = \langle \langle m_1, m_2 ... \rangle, R, L \rangle$ [4]. In the graph oriented representation of $\mathcal{M}$, the nodes correspond to the sequence of states brought about by different modalities represented as links between states. Each state is labeled by a different nominal, while links are labeled by the relation connecting two states.

Running scenario. The illustrative scenario regards the safe insertion of a UAV into the civil air traffic as shown in [3]. The presented Unmanned Aircraft System consists of the UAV itself equipped with an autonomous control system, a ground station and the Air Traffic Management, which provides the required coordinates for the UAV. The goal is to prove that an UAV can complete safely its mission inside the civil air traffic and that all the major implied risks (e.g. collision with other objects or UAVs, loss of critical functions) are mitigated. For space considerations, we will restrict ourselves to those safety cases related to collision risks. In this specific context, an autonomous decision making system must consider at all times the set of safety regulations elaborated to deal with collision detection and avoidance imposed during a mission [11].

The corresponding hybrid Kripke structure is illustrated in Fig. 2. Its states correspond to the basic functions of the system (table 1): path following along the established corridor (*PathFollowing*), detection of possible obstacles (*DetectObstacles*), avoidance maneuver (*AvoidObstacles*), fault control (*FaultControl*) and landing (*Land*). The transition from one state to another is triggered by an event that leads to a change in the system's parameters: *obs* (signals presence of obstacles), *d* (returns distance between UAV and obstacle), *errObs* (signals an error in the *Detect* function) and *errAvoid* (signals an error in the *Avoid* function). For example, the signaling of an approaching obstacle when in the *DetectObstacles* state leads to the transition of the system in the *AvoidObstacles* state. The signaling of an error in any of the *DetectObstacles* or *AvoidObstacles* functions,

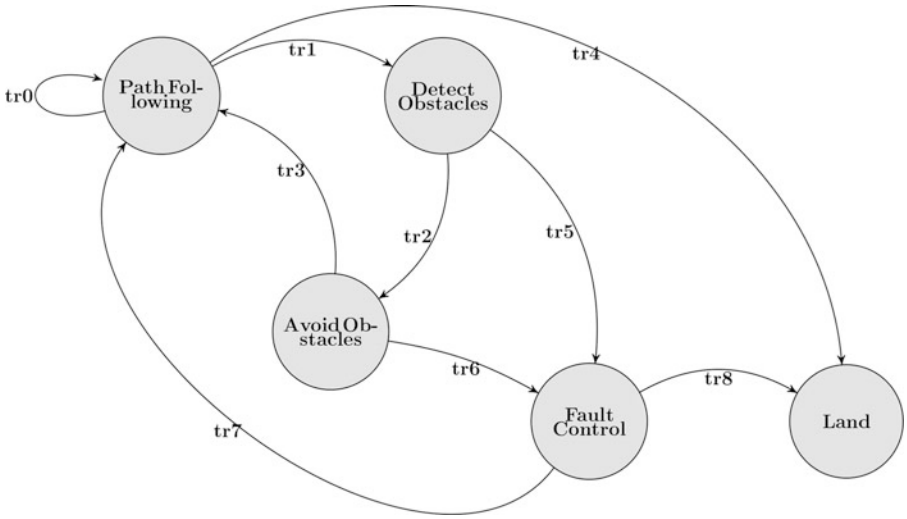


Fig. 2. Kripke model for the UAV.

Table 1. Set of states of the UAS

State	Parameters	Specification
PathFollowing	$\neg obs$	UAV follows the path on the given corridor
DetectObstacles	*obs*	Obstacles are signaled by sensors
AvoidObstacles	$obs \wedge d$	UAV performs an avoidance maneuver
FaultControl	$errObs \vee errAvoid$	Error signaled by *Detect* or *Avoid*
Land	$\neg obs \vee errObs \vee errAvoid$	UAV performs the landing procedure

leads the system in the *FaultControl* state. If the system recovers from the error state, it returns to the *PathFollowing* state. Otherwise, the mission is aborted by landing to a designated location, hence entering in the *Land* state.

3 A Formal Model for the GSN Standard

3.1 The GSN Safety Case for the UAV Scenario

In this section we build the safety case for our scenario as a GSN diagram. The top level goal states that all risks of collisions are managed (Fig. 3). This claim is refined into two more specific sub-goals, each capturing a different possible context: with another UAV present in the same corridor space or with another object. We further decompose the sub-claim referring to the exceptional case of collision with another UAV, arguing that the avoidance must be ensured by a specific emergency procedure (based on the so-called Detect&Avoid function) and by mitigating all the risks in case of function loss. We continue the refinement process until we come to elementary claims that may be ensured by evidences. Considering the sub-claim referring to the situation in which no other UAV is present in the corridor space, we argue that the ATM transmits the correct coordinates for the right path following and that the UAV acknowledges correctly the commands received and sets its trajectory based on them. Moreover, we argue that the Detect&Avoid function is correct and therefore it must ensure that an obstacle is identified within a certain distance which allows the safe application of avoidance maneuver and that the UAV performs the indicated safety avoidance commands. Additionally, a safety claim is associated to the Detect&Avoid function stating that the risks of loss for this function are acceptable. The risks are calculated as specified in [3] using the Functional Hazard Assessment (FHA) for identifying a severity for each failure and flight mode(automatic or manual) of the UAV, and the Preliminary System Safety Assessment (PSSA) for deriving the safety requirements. These analyses are taken as evidences in validating as acceptable the risks of loss for the Detect&Avoid function.

3.2 Modeling the Goal Structuring Notation in DL

The relationship *supportedBy*, allows inferential or evidential relationships to be documented. The allowed connections for the *supportedBy* relationship are: goal-to-goal, goal-to-strategy, goal-to-solution, strategy to goal.

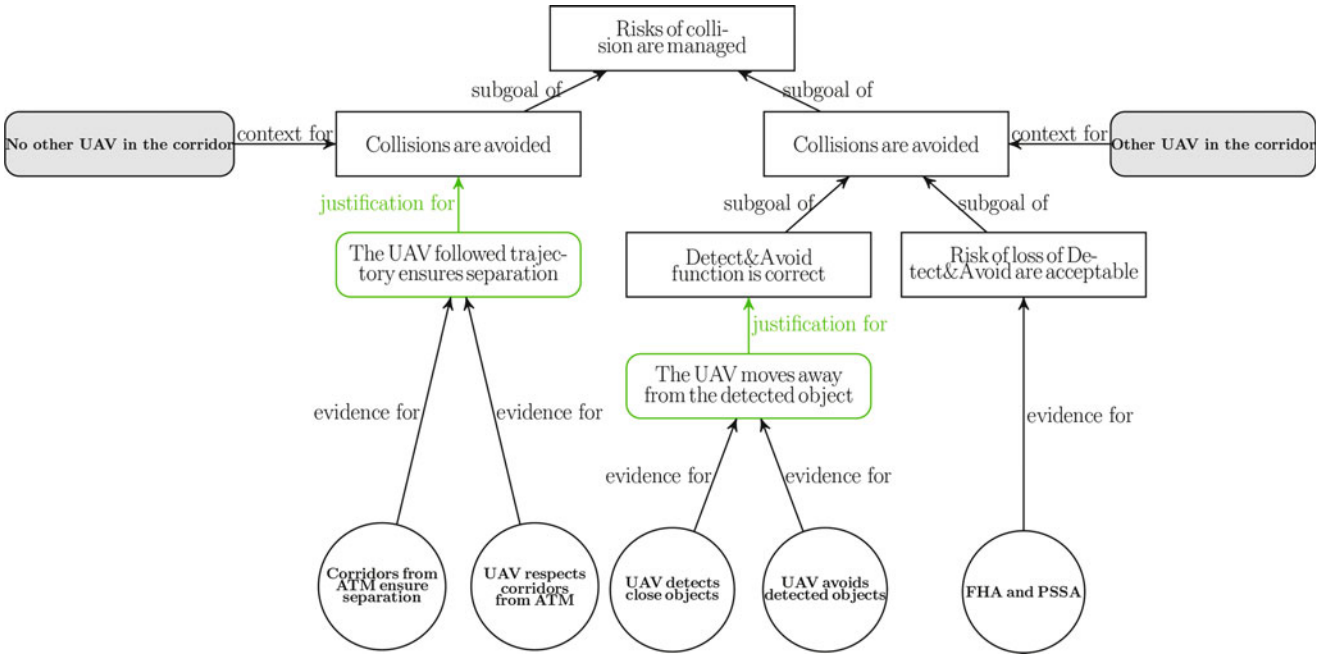


Fig. 3. Goal Structuring Notation

Axiom A_1 specifies the range for the role *supportedBy*, axiom A_2 the range, axiom A_3 introduces the inverse role *supports*, and A_4 constraints the role *supportedBy* to be transitive.

(A_1) $\top \sqsubseteq \forall supportedBy.(Goal \bigsqcup Strategy \bigsqcup Solution)$
(A_2) $\exists supportedBy \top \sqsubseteq Goal \bigsqcup Strategy$
(A_3) $supportedBy^{-} \equiv supports$
(A_4) $supportedBy \sqsubseteq supportedBy$

Inferential relationships declare that there is an inference between goals in the argument. Evidential relationships specify the link between a goal and the evidence used to support it. Axioms A_5 and A_8 specify the range of the roles *hasInference*, respectively *hasEvidence*, while A_6 and A_9 the domain of the same roles. Definitions A_7 and A_{10} say that the *supportedBy* is the parent role of both *hasInference* and *hasEvidence*, thus inheriting its constraints.

(A_5) $\top \sqsubseteq \forall hasInference.Goal$
(A_6) $\exists hasInference \top \sqsubseteq$ Goal
(A_7) $hasInference \sqsubseteq supportedBy$
(A_8) $\top \sqsubseteq \forall hasEvidence.Evidence$
(A_9) $\exists hasEvidence \top \sqsubseteq$ Goal
(A_{10}) $hasEvidence \sqsubseteq supportedBy$

Goals and sub-goals are propositions that we wish to be true that can be quantified as quantified or qualitative, provable or uncertainty.

(A_{11}) $QuantitativeGoal \sqsubseteq Goal$
(A_{12}) $QualitativeGoal \sqsubseteq Goal$
(A_{13}) $ProvableGoal \sqsubseteq Goal$
(A_{14}) $UncertaintyGoal \sqsubseteq Goal$

A sub-goal supports other high level goals. Each safety case has a top level *Goal*, which does not support other goals.

(A_{15}) $SupportGoal \equiv Goal \sqcap \exists$supports.$\top$
(A_{16}) $TopLevelGoal \equiv Goal \sqcap (\neg$SupportGoal$)$

For each safety argument, the elements is instantiated and a textual description is attached to that individual by enacting the attribute *hasText*:

(A_{17}) $\top \sqsubseteq \forall hasText.String$),
(f_1) $gt : TopLevelGoal$,
(f_3) $gp : ProvableGoal$,
(f_5) $gu : UncertaintyGoal$,
(A_{18}) $\exists hasText.Statement \sqsubseteq \top$
(f_2) $(gt$,"The system meets requirements"$) : hasText$
(f_4) $(gp$,"Quick release are used"$) : hasText$
(f_6) $(gu$,"Item has a reliability of 95%"$) : hasText$

Intermediate explanatory steps between goals and the evidence include statements, references, justifications and assumptions.

(A_{20})
$Explanation \sqsubseteq Statement \bigsqcup Reference \bigsqcup Justification \bigsqcup Assumption$
(A_{21}) $Statement \equiv \neg$Reference, (A_{22}) $Statement \equiv \neg$Justification
(A_{23}) $Statement \equiv \neg$Assumption, (A_{24}) $Reference \equiv \neg$Justification,
(A_{25}) $Reference \equiv \neg$Assumption, (A_{26}) $Justification \equiv \neg$Assumption,

The evidences or solutions form the foundation of the argument and will typically include specific analysis or test results that provide evidence of an attribute of the system. In our approach, the evidence consists in model checking the verification for a specification of the system.

(A_{27}) $Evidence \sqsubseteq$ $\exists$ *has Formula.Formula* $\sqcap$ $\exists$ *has Specifiacation.Statement* $\exists hasModel.KripkeModel \sqcap \exists hasTestResult.\top$

Given the above formalization for GSN, our scenario depicted in Fig. 3 is formally represented in Fig. 4.

(A_{28}) $NotVerifiedGoal \equiv Goal \sqcap \exists hasEvidence.NotVerifiedEvidence$
(A_{29}) $NotVerifiedEvidence \equiv Evidence \sqcap \exists hasTestResult.False$

Table 2. Retrieving information about the GSN model.

Query	RacerPro query	RacerPro answer
Top level goal	*(concept − instances TopLevelGoal)*	g_1
Support goals	*(concept − instances SupportGoal)*	g_2, g_3, g_4, g_5
Evidence supporting goal g_1	*(individual − fillers $g_1$hasEvidence)*	e_1, e_2
Evidence verified against the model m_1	*(individual − fillers m_1) (inverse hasModel))*	e_1, e_2, e_3, e_4, e_5
Evidence not verified	*(concept − instances(and Evidence (some hasTestResult False))*	e_1, e_2, e_3, e_4, e_5
Goals supported by not verified evidence	*(concept − instances NotVerifiedGoals)*	g_1, g_2, g_3, g_4, g_5

Fig. 4. The Abox of the UAV scenario.

$$
\begin{array}{l}
g_1 : Goal,\ g_2 : Goal,\ g_3 : Goal,\ g_4 : Goal,\ g_5 : Goal \\
e_1 : Evidence,\ e_2 : Evidence,\ e_3 : Evidence,\ e_4 : Evidence,\ e_5 : Evidence \\
(g_2, g_1) : supports,\ (g_3, g_1) : supports,\ (g_4, g_3) : supports,\ (g_5, g_3) : supports \\
(g_2, e_1) : hasEvidence,\ (g_2, e_2) : hasEvidence,\ (g_4, e_3) : hasEvidence \\
(g_4, e_4) : hasEvidence,\ (g_5, e_5) : hasEvidence \\
(g_1, \text{"}Risks\ of\ collision\ are\ managed.\text{"}) : hasText \\
(g_2, \text{"}Collisions\ are\ avoided\ -\ No\ UAV.\text{"}) : hasText \\
(g_3, \text{"}Collisions\ are\ avoided\ -\ UAV.\text{"}) : hasText \\
(g_4, \text{"}Detect\&Avoid\ function\ is\ correct.\text{"}) : hasText \\
(g_5, \text{"}Risk\ of\ loss\ of\ Detect\&Avoid\ is\ acceptable.\text{"}) : hasText \\
(e_1, \text{"}Corridors\ from\ ATM\ ensure\ separation.\text{"}) : hasSpecification \\
(e_2, \text{"}UAV\ respects\ corridors\ from\ ATM.\text{"}) : hasSpecification \\
(e_3, \text{"}UAV\ detects\ close\ objects.\text{"}) : hasSpecification \\
(e_4, \text{"}UAV\ avoids\ detected\ objects.\text{"}) : hasSpecification \\
(e_5, \text{"}FHA\ and\ PSSA.\text{"}) : hasSpecification \\
c_1 : Context,\ (c_1, \text{"}No\ other\ UAV\ in\ the\ corridor.\text{"}) : hasText \\
c_2 : Context,\ (c_2, \text{"}Other\ UAV\ in\ the\ corridor.\text{"}) : hasText \\
m_1 : KripkeModel,\ (e_1, m_1) : hasModel,\ (e_2, m_1) : hasModel \\
(e_3, m_1) : hasModel,\ (e_4, m_1) : hasModel,\ (e_5, m_1) : hasModel
\end{array}
$$

4 Interleaving Reasoning with HL and DL for Identifying Assurance Deficits

Our method interleaves two steps: First, we check with hybrid logic if the evidence nodes from the GSN representation have their corresponding formulas validated against the Kripke model. Second, by reasoning in DL, we identify which goals in the GSN model are not supported by verified evidence.

4.1 Validating Evidence with Model Checking

For the given scenario, we start by verifying the first two pieces of evidence e_1 and e_2 in the model M_1. The verification uses three parameters: (i) the minimum distance d_{min} allowed between the UAV and another object without risk of collision; (ii) the reported coordinates c_{uav} by the UAV; and (iii) the given coordinates c_{ATM} by the ATM. Formula f_1 attached to evidence e_1 through the assertion (*related* e_1 f_1 *hasFormula*) in DL is expressed in HL as:

$$f_1 = \downarrow i(c_{ATM}) \rightarrow @_i[F](c_{ATM} > d_{min}) \quad (1)$$

f_1 states that if the ATM starts transmitting coordinates at a state i, then for all future states the coordinates will be transmitted such that to ensure that the minimum safe distance is preserved between the UAV and other objects.

The formula corresponding to the evidence e_2 is:

$$f_2 = \downarrow i(c_{ATM}) \rightarrow @_i[Next](c_{uav} = c_{ATM}) \quad (2)$$

According to f_2, if the ATM starts transmitting coordinates at a state i, then in the next state the UAV should acknowledge the newly received coordinates by reporting the exact coordinates as the ones transmitted in the previous state. The justification j_2 of the sub-goal g_2 supported by e_1 and e_2 is expressed as:

$$j_2 = \downarrow i(c_{uav} = c_{ATM}) \rightarrow @_i[F](c_{uav} > d_{min}) \quad (3)$$

The implication $f_1 \wedge f_2 \rightarrow j_2$ is true (the acknowledgment and following of the coordinates from the ATM ensures the required minimum safe distance).

Fig. 5 Updating the Abox for the GSN model with the newly validated evidences.

$$
\begin{array}{ll}
(e_1, f_1) : hasFormula, & (e_1, "true") : hasTestResult \\
(e_2, f_2) : hasFormula, & (e_2, "true") : hasTestResult \\
(g_2, j_2) : hasJustification, & (f - g_2, "true") : hasTestResult \\
(e_3, f_3) : hasFormula, & (e_3, "true") : hasTestResult \\
(e_4, f_4) : hasFormula, & (e_4, "true") : hasTestResult \\
(g_4, j_4) : hasJustification, & (f - g_4, "true") : hasTestResult
\end{array}
$$

Evidences e_3 and e_4 are used to validate the sub-goal g_4 about the correctness of the Detect&Avoid function. To check the supporting evidences, two parameters are required: (i) the reported distance d_{obs} between the UAV and another approaching UAV; and (ii) the minimum distance d_{min} allowed without any risk of collision. The justification j_4 for the sub-goal g_4 is formalized as:

$$j_4 = \downarrow i(d_{obs} < d_{min}) \rightarrow @_i[F]((d_{obs} \neq 0)U(d_{obs} > d_{min})) \quad (4)$$

Justification j_4 states that if we bind to i the state in which the reported distance between the UAV and another approaching UAV is less than the minimum one then for all future states the reported distance must be kept higher then 0, increasing it, at the same time, until no danger of collision ($d_{obj} > d_{min}$).

Evidence e_3 (*UAV detects close objects*) is formally expressed as:

$$f_3 = \downarrow i(d_{obs}) \wedge @_i(d_{obs} < d_{min}) \rightarrow \downarrow i(obs) \quad (5)$$

According to f_3, if in the current state named by nominal i, the distance to a possible obstacle is transmitted to the UAV and the distance is less than the minimum allowed one, the presence of an obstacle is reported by the sensors to the UAV signaling a risk for collision.

Evidence e_4 (*UAV avoids detected objects*) is formally expressed as:

$$f_4 = \downarrow i(obs) \rightarrow @_i((d_{obs} \neq 0)U(d_{obs} > d_{min})) \quad (6)$$

Equation 6 states that if we bind to nominal i the state in which an obstacle is signaled by the sensors to the UAV, then the reported distance to the obstacle must be maintained different than 0 until the increase of distance between the UAV and the obstacle becomes higher then the minimum established threshold, indicating that the avoidance maneuver was performed.

To complete the validation of g_4, we have to prove the formula $f_3 \wedge f_4 \rightarrow j_4$, which is true (the presence of an obstacle indicated by an observed distance, which is less than the minimum accepted one will entail an avoidance maneuver).

4.2 Identifying Assurance Deficits

At this time check, the formal GSN model is updated with the assertions in Fig. 5. Given the new information, the GSN model can be interogated to retrieve goals and evidence which are not validated yet. Querying the RacerPro engine for the *NotVerifiedGoal*s, we obtain g_1, g_3, g_5, while the concept *NotVerifiedEvidence* includes only one instance, the evidence e_5. The RacerPro system is able to provide explanations why a specific goal belongs to a specific concept. In this way, the safety engineer can figure that the goal g_3 is not validated because of g_5, which relies on the piece of evidence e_5 whose formula was not checked in the given kripke model M_1. This reasoning mechanism is particularly useful in real application where a GSN model has hundreds of nodes.

Given the above knowledge, the safety engineer is aware that the *SupportGoal* g_5 should be validated. In the given scenario, the validation is based on the analysis results of the FHA and PSSA considering the reported error parameters *errObs* and *errA*. The maximum acceptable degree of risk will be referred as r_a. If in the FaultControl state the parameters *errObs* and *errA* will lead to a risk result r_{err} which is higher than the maximum degree of allowed risks, then the emergency landing is performed. Formally:

$$f_7 = \downarrow i(FaultControl) \wedge @_i(r_{err} < r_a) \rightarrow ([Next]i \rightarrow Land) \quad (7)$$

One can observe from the Kripke structure in Fig. 2 that there is a valid transition from state *FaultControl* to state *Land* in case that the risk is higher than the acceptable limit, but also to *PathFollowing* in case that the returned result is a positive one and it allows the UAV to continue its mission safely. Therefore, formula f_7 proves as true. With this new information sent to the RacerPro engine, the concept *NotVerifiedGoals* will contain no instances, which formally validates the safety case from the GSN model.

5 Discussion and Related Work

While both argumentation [7, 6, 10] and model checking [11] have been applied for certification of safety systems, we aimed to demonstrate that combining the two methods might bring about additional advantages such as preliminary validation of argumentation schemes constructed to support safety

cases, ensuring in advance that the stability of the system will not be affected by the available choices and, at the same time, foreseeing possible impediments in selecting one option over another. Considering the benefits of abstractization by combining DL with model checking [9], we complemented the graphical GSN standard with a formalized model. We argue that this joint approach increase the transparency and trust when certifying critical safety systems.

6 Conclusion

The contributions of the paper are: 1) integrating hybrid logic with argumentation theory, and 2) providing a formal model of the GSN standard in description logic. While the GSN graphical argumentation language structures safety cases and facilitates understanding for the human agent, the hybrid logic is able to validate the evidence nodes of the diagram. Description logic was used as a middleware language to lightly integrate GSN and model checking. DL's reasoning capabilities are used to analyze the status of the arguments and their supporting evidence. In our view, the proposed method is a step towards a formal model for the GSN standard. Currently, we are investigating the feasibility of our solution against large-scale safety cases.

Acknowledgments This work was supported by the Romania-Moldova Bilateral Agreement entitled "ASDEC: Structural Argumentation for Decision Support with Normative Constraints", from the National Research Council of the Romanian Ministry of Education and Research and Moldova Ministry of Education.

References

1. Areces, C., ten Cate, B.: Hybrid logics. In: Blackburn, P., Van Benthem, J., Wolter, F. (eds.) Handbook of Modal Logic, pp. 821–868. Elsevier Amsterdam (2007)
2. Baader, F.: The description logic handbook: theory, implementation, and applications. Cambridge university press (2003)
3. Brunel, J., Cazin, J.: Formal methods for the certification of autonomous unmanned aircraft systems. In: Formal Verification of a Safety Argumentation and Application to a Complex UAV System. pp. 307–318. SAFECOMP'11, Springer-Verlag, Berlin, Heidelberg (2012)
4. Cranefield, S., Winikoff, M.: Verifying social expectations by model checking truncated paths. Journal of Logic and Computation 21(6), 1217–1256 (2011)
5. Franceschet, M., de Rijke, M.: Model checking hybrid logics (with an application to semistructured data). Journal of Applied Logic 4, 279–304 (2006)
6. Graydon, P., Habli, I., Hawkins, R., Kelly, T., Knight, J.: Arguing conformance. Software, IEEE 29(3), 50–57 (2012)
7. Graydon, P., Kelly, T.P.: Using argumentation to evaluate software assurance standards. Information and Software Technology 55(9), 1551–1562 (2013)
8. Haarslev, V., Hidde, K., Möller, R., Wessel, M.: The racerpro knowledge representation and reasoning system. Semantic Web 3 (3), 267–277 (2012)
9. Letia, I.A., Groza, A.: Compliance checking of integrated business processes. Data Knowl. Eng. 87, 1–18 (2013)
10. Rushby, J.: A safety-case approach for certifying adaptive systems. In: AIAA Infotech@Aerospace Conference, American Inst. of Aeronautics and Astronautics (2009)
11. Webster, M., Fisher, M., Cameron, N., Jump, M.: Formal methods for the certification of autonomous unmanned aircraft systems. In: Proceedings of the 30th International Conference on Computer Safety, Reliability, and Security. pp. 228–242. SAFECOMP'11, Springer-Verlag, Berlin, Heidelberg (2011)

A Load Optimization Considering Reverse Synergy that May Occur with Mixed Load

Yongmin Kim, Munhwan Kim, and Hongchul Lee

1 Introduction

1.1 Background

Today the industry comes to the fore about concept of the Supply Chain Management, so reducing cost became a very big part in awareness of logistics. It is not only concept of carrying out the product but also has great influence to business competitive power. Actually, this influence can be absolute to the future. Many companies have already established a SCM department to gain the upper hand in business management and studying about reducing logistics costs has been also performed actively in the industry and academia (Lee et al, 2003).

1.2 Existing research

In this study, we focus on bin-packing problem which is in large field of logistics. Bin-packing problem is the distribution problem to load different volume of cargos in limited load containers efficiently. This is the kind of problem that geometric objects should be placed optimally(Kim et al, 2000). We can set an example of load container as a truck, van, container etc and if a container has constant load capacity, we call it "bin". The bin-packing problem belongs to the category of NP-hard problems and even though this problem with the basic mathematical models and algorithms has been proposed previously, many researchers are still trying to improve these problems(Tarantilis et al, 2009).

Kim proposed a formulation and heuristic algorithms of bin packing problems based on the fuzzy set theory (Kim et al, 2000) and Park proposed the two-dimensional bin-packing optimization techniques for mother plate design (Park et al, 2006).

Y. Kim (✉) • M. Kim • H. Lee
Industrial Management Engineering, Korea University, Seoul, Korea
e-mail: iamkym@korea.ac.kr; munhwan@korea.ac.kr; hclee@korea.ac.kr

1.3 Direction of research

However, up to the present time, any researches have not considered reverse synergy that occurs with mixed load. Mixed load is a load container which loads different kind of cargos together and it can occurs reverse synergy depending on characteristic of cargos at this time. For example, if we load mixed cargos which are very damageable, fragile and overweight, we will have penalties like paying for additional protective device to prevent damage to cargos at process of distribution. Also, we should consider reverse synergy even if we load two types of cargos which have the opposite inclination like one must be maintain high temperature and the other one must be maintain low temperature. Therefore, in this study, we consider the characteristics of cargos and provide a pairwise matrix of reverse synergy for each pair of cargos. After that, we improve mathematical model to reflect it.

2 Mathematical model of bin-packing problem

Prior to explaining about reverse synergy, we describe the existing basic formulation of bin-packing problem briefly.

2.1 Index and Inputs

I : A set of containers.
J : A set of cargos.
w_i : The load capacity of container i.
t_j : The load capacity consumption of cargo j.

2.2 Decision variables

y_i : The binary variable representing whether the container is used.

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_36,

$$\begin{cases} 1, & \textit{if the container } i \textit{ is used}, \\ 0, & \textit{otherwise}. \end{cases}$$

x_{ij}: The binary variable representing whether the cargo is loaded the container.

$$\begin{cases} 1, & \textit{if the cargo } j \textit{ is loaded the container } i, \\ 0, & \textit{otherwise}. \end{cases}$$

2.3 Formulation

$$\begin{aligned} \textit{Minimize} \quad & \sum_{i \in I} c_i y_i \\ s.t \quad & \sum_{j \in J} t_j x_{ij} \le w_i y_i, \quad \forall\, i \in I, \\ & \sum_{i \in I} x_{ij} = 1, \qquad \forall\, j \in J, \\ & \textit{All } x_{ij}, y_i \textit{ are binary}. \end{aligned}$$

The objective function $\sum_{i \in I} c_i y_i$ means the cost of using containers and it should be minimized in this problem. The first constraint $\sum_{j \in J} t_j x_{ij} \le w_i y_i, \quad \forall \quad i \in I$ means that the cargos loaded each container cannot exceed the load capacity of each container w_i. The second constraint $\sum_{i \in I} x_{ij} = 1$, $\forall\, j \in J$ means each cargo must be loaded a container. The last one represents decision variable x_{ij} and y_i should have the value of 0 or 1.

3 Mathematical model considering reverse synergy

We suggest the mathematical model considering reverse synergy that has not been considered previously. We propose a pairwise matrix in respect of reverse synergy first. And then, we improve mathematical model to reflect our suggestion as below.

3.1 A pairwise matrix in respect of reverse synergy

We already pointed out a few easy examples for a reverse synergy at the chapter 1.3. Depending on the characteristic of each cargo for carrying, it is decided whether reverse synergy exists or not. In addition, it decides the cost of reverse synergy for each pair of cargos. In this chapter, we reflect this concept by using a pairwise matrix A. A is defined by $n \times n$ square matrix as the number of items n.

$$A = \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ a_{21} & \ddots & & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & 0 \end{bmatrix}$$

Each ingredient of matrix A, a_{kl} represents the cost of reverse synergy between item k and l. Therefore, a_{kl} and a_{lk} should have same value, and matrix A is symmetric. If the cost of reverse synergy a_{kl} is large, it means the penalty will be also large when we put both items to same container. Furthermore, all the diagonal ingredients of matrix A are 0 because reverse synergy does not affect when we load same items.

3.2 Improvement of the mathematical model

In this chapter, we modify the existing mathematical model to reflect the proposed concept of reverse synergy.

3.2.1 Additional Index and Inputs

n : The number of items.

N : A set of items, $N = \{1,2,\ldots,n\}$.

E_m : A set of cargos that belong to the item m.

a_{kl} : The cost of reverse synergy between item k and l.

3.2.2 Additional Decision variables

s_{ikl}: The binary variable representing whether the reverse synergy between item and affects. It is same with the one representing whether both item k and l are loaded the same container i.

$$\begin{cases} 1, & \textit{if the reverse synergy between item } k \textit{ and } l \textit{ affects}, \\ 0, & \textit{otherwise}. \end{cases}$$

3.2.3 Improved formulation

$$\begin{aligned} \textit{Minimize} \quad & \sum_{i \in I} c_i y_i + \sum_{i \in I} \sum_{k \in N} \sum_{l \in N} a_{kl} s_{ikl} \\ s.t \quad & \sum_{j \in J} t_j x_{ij} \le w_i y_i, \quad \forall\, i \in I, \\ & \sum_{i \in I} x_{ij} = 1, \qquad \forall\, j \in J, \\ & s_{ikl} \ge x_{it} + x_{iu} - 1, \\ & \forall i \in I,\ \forall k, l \in N,\ \forall t \in E_k, \forall u \in E_l, \\ & \textit{All } x_{ij}, y_i, s_{ikl} \textit{ are binary}. \end{aligned}$$

The term for cost of reverse synergy $\sum_{i \in I} \sum_{k \in N} \sum_{l \in N} a_{kl} s_{ikl}$ is added to the existing objective function. It means the total of cost incurred as each pair of item is loaded a same container.

The newly added constraint $S_{ikl} \geq x_{it} + x_{iu} - 1$ plays a role in making s_{ikl} have a proper value.

4 Conclusions

In this study, we proposed the mathematical model to solve an optimal solution considering reverse synergy that may occur with mixed load. This goes further to the existing method which only aims to minimize the cost for load container and we put meaning on considering the cost of reverse synergy between the load items. The details of our study are still insufficient but we expanded our goal from existing study which seek only to minimize the cost of using load containers. In other words, this study makes it possible to consider a number of factors simultaneously by the proposed model and we hope it will become a foundation of relative future research.

Acknowledgement This work was supported by the BK21 Plus(Big Data in Manufacturing and Logistics Systems, Korea University).

References

Casazza, M., Ceselli, A. (2014), Mathematical programming algorithms for bin packing problems with item fragmentation, *Computers & Operations Research*, 46, 1-11.

Kim, J. K. (2000), A formulation and heuristic algorithms of bin packing problems based on the fuzzy set theory, *Kaist*.

Lee, Y. D. (2003), A case study of container bin-packing problem, *Journal of Management*, 10(1), 17-25.

Park, S. H., & Jang, S. Y. (2006), Two-dimensional bin packing optimization model for mother plate design, *The Korean Institute of Industrial Engineers*, 5, 137-142.

Tarantilis, C. D., Zachariadis, E. E., Kiranoudis, C. T. (2009), A Hybrid Metaheuristic Algorithm for the Integrated Vehicle Routing and Three-Dimensional Container-Loading Problem, *Intelligent Transportation Systems, IEEE Transactions on*, 10(2), 225-271.

Predictability of Firm Financial Sustainability Using Artificial Neural Networks: The Case of Qatar Exchange

Farzaneh Amani and Adam Fadlalla

1 Introduction

Investors, mangers and many other stakeholders use a firm's publically available financial information published in financial statements in assessing the current financial position of the firm, making decisions about their investments and predicting future firm financial sustainability. Predicting future financial sustainability accurately is a challenging task because many variables that affect the firm are beyond the control of the stakeholders [1]. Different techniques were used in predicting future earnings such as artificial neural networks (ANNs) [2], support vector machines (SVM) [3, 4], fuzzy systems, [5, 6], genetic algorithms [7, 8] and other sophisticated techniques. Many researchers used only the accounting-based data reported in the financial statements in constructing their predictive models such as [9] and [10], while others such as [11] used market-based data. Recent literature such as the work of [12] and [13] showed a trend towards building hybrid models that use both accounting-based and market-based data.

Many researchers assessed the predictive power of accounting-based models versus market-based models with contradictory results. For example, whereas [14] found that the market-based model of Merton [15] outperforms the well-known accounting-based models of Altman [16] and Ohlson [17] in the assessment of default risk and therefore the firm non-sustainability. In addition, [18] reported that market-based models outperformed accounting-based models because market-based data includes additional information, such as macroeconomic factors and investors' expectations regarding the future of the company, which is not reflected in the accounting-based data. On the other hand, [19] found that accounting-based data is more relevant in bankruptcy (non-sustainability) prediction even without the inclusion of market-based data. Also [19] reported that accounting-based and market-based data possess complementary information in the prediction of default (non-sustainability).

The purpose of this paper is to provide evidence in the context of Qatar Exchange (QE) with the respect to ability of a hybrid model of accounting and market based measures to predict firm future financial sustainability using multilayer perceptron neural network (MLP-NN). More specifically the paper examined (1) the impact of market-based measures on predicting firm future financial sustainability, (2) the impact of accounting-based measures on predicting firm future financial sustainability, and (3) ability of a hybrid model of accounting and market based measures to predict firm future financial sustainability. Future financial sustainability is measured by (1) future firm earnings using return on assets and (2) future sales growth rate. The rest of the paper is organized as follows: section two provides a background about QE and the measures used to predict financial sustainability; section three describes the research methodology, section four presents and discusses the research results, and section five offers conclusions.

2 Background

ANNs have been used extensively in predicating future financial performance, for example stock price performance [20, 21, 22, 23], and stock closing price [24].

QE is the main stock market of State of Qatar and it was established in 1997. Is one of the main Middle East stock markets (Saudi, Dubai, Abu Dhabi, Bahrain, Kuwait, Oman and Qatar). As of April 2014, QE has market capitalization of $183.14B, total volume of $43.19M and 43 listed companies [25].

Authors used many of accounting and market based measures to predict future performance for example [19] and [26]. Examples of market-based measures are earning per share (EPS), price earning (PE) ratio, price to cash flow ratio, and price to book ratio and examples of accounting-

F. Amani (✉) • A. Fadlalla
Qatar University, Doha, Qatar

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_37,

Table 1 Descriptive Statistics for Accounting-Based Measures.

	N	Range	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
BS_TOT_ASSET	98	79967.7	70.1	80037.8	17253.165	2182.5078	21605.7253
CF_CASH_FROM_OPER	98	13867.0	-5918.2	7948.8	678.565	154.2815	1527.3087
LT_DEBT_TO_TOT_ASSET	98	57.8	.0	57.8	7.269	1.1183	11.0709
RETENTION_RATIO	98	97.1	2.9	100.0	43.710	2.8736	28.4474
Valid N (listwise)	98						

Table 2 Descriptive Statistics for Market-Based Measures.

	N	Range	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
IS_EPS	98	12.0	.4	12.4	4.346	.2744	2.7163
PE_RATIO	98	52.5	6.0	58.5	14.117	.7758	7.6797
PX_TO_BOOK_RATIO	98	5.3	.6	5.9	2.030	.1119	1.1075
PX_TO_CASH_FLOW	98	281.9	.9	282.8	20.907	3.9345	38.9497
Valid N (listwise)	98						

based measures are total assets, cash flow from operations, long term debt to total asset and retention ratio [27]. To the best of our knowledge none of the finding of the financial sustainability has been validated in the context of QE. Thus it is the aim of this paper to close this gap.

3 Methodology

We obtained the historical financial data for companies listed in QE from Bloomberg for the period 2000 to 2013. We downloaded for market-based measures: EPS, PE ratio, price to cash flow ratio, and price to book ratio. For accounting-based measures we downloaded: total assets, cash flow from operations, long term debt to total asset and retention ratio. Only companies with complete data for specific years were maintained. Most of the data were missing and only total of 98 company-years have complete data. This represents the sample of empirical analysis. Tables 1 and 2 provide summary statistics for these variables.

We developed MLP-NN with 70% of the data as training set and 30% of the data as testing set. The MLP-NN had an input layer, a hidden layer, an output layer and used backpropagation algorithm. Table 3 and Figure 1 summarize the characteristics of the hybrid neural network model.

4 Results and Discussion

4.1 Market-based Model

We developed an MLP-NN using the four market-based measures; namely, EPS, PE ratio, price to cash flow ratio, and price to book ratio to predict firm future earnings. The results of the MLP-NN indicated that PE ratio (a proxy for equity valuation) and price to book value ratio (a proxy for bankruptcy tolerance) were the most significant market-based measures in predicting firm future earnings. The MLP-NN achieved a relative error of 0.70 and a mean absolute error of 4.24.

On the other hand, we developed an MLP-NN using the same four market-based variables to predict firm future sales growth rate, price to book ratio and price to cash flow ratio were the most significant predictors. MLP-NN achieved a relative error of 0.53 and a mean absolute error of 4.10. It is clear that market-based measures are more effective predictors for sustainable sales growth rate than they are for predicting firm future earnings.

4.2 Accounting-based Model

An accounting-based neural network using total assets, cash flow from operations, long term debt to total asset and retention ratio to predict firm future earnings achieved a relative error of 0.63 and a mean absolute error of 3.18. Total assets (a proxy for firm size) and long-term debt to total assets (a proxy for solvency) were the most important predictors.

In addition, we developed an MLP-NN using the same accounting-based measures to predict firm future sales growth rate. This time retention ratio and cash flow from operation were the most significant predictors. The model achieved a relative error of 0.92 and a mean absolute error of 4.94.

Our empirical results indicate that accounting-based measures are more powerful predictors of firm future earnings compared to market-based measures. This result is

Table 3 Neural Network Information.

Input Layer	Covariates	1	BS_TOT_ASSET
		2	CF_CASH_FROM_OPER
		3	IS_EPS
		4	LT_DEBT_TO_TOT_ASSET
		5	PE_RATIO
		6	PX_TO_BOOK_RATIO
		7	PX_TO_CASH_FLOW
		8	RETENTION_RATIO
	Number of Units[a]		8
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		5
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	ANN_RETURN_ON_ASSET_NY
	Number of Units		1
	Rescaling Method for Scale Dependents		Standardized
	Activation Function		Identity
	Error Function		Sum of Squares

a. Excluding the bias unit

consistent with the finding related to the power of accounting-based measures reported in [19]. However, when predicting firm future sales growth rate market-based measures took the lead, this result is similar to what was reported in [14] and [18].

4.3 Hybrid Model:

To test the ability of the combined accounting and market based measures in predicting firm future financial sustainability we developed a hybrid NN model using this combined set of variables. The MLP-NN results demonstrated improvement in predicting both measures of financial sustainability. The hybrid model achieved the best relative errors of 0.33 and 0.73 and mean absolute errors of 2.42 and 4.14 for predicting firm future earnings and firm future sales growth rate, respectively. The hybrid model was more effective in predicting firm future earnings than firm future sales growth rate. We believe this is because of the market noise embedded in the market measures. In this case, our results are consistent with [19].

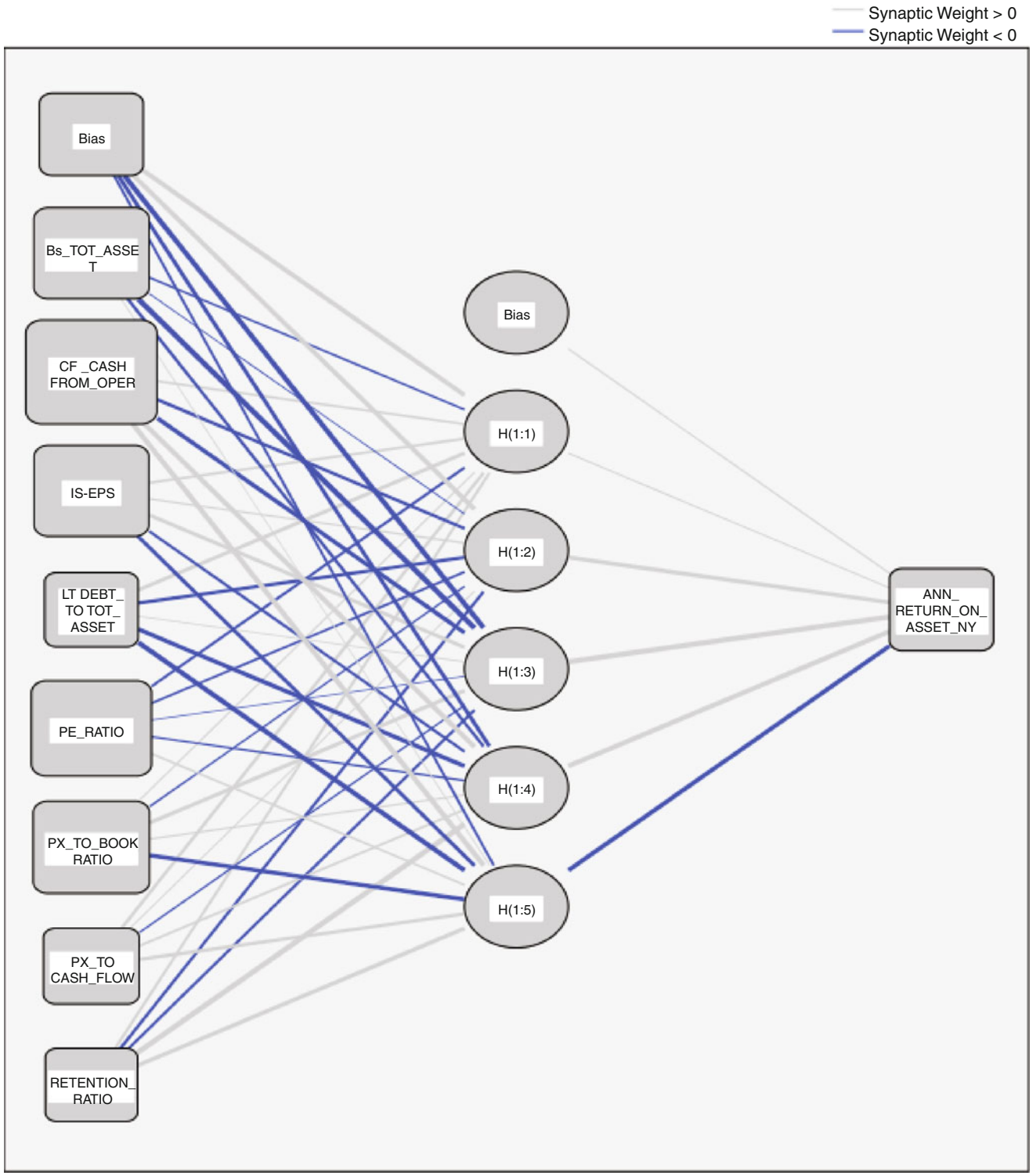


Fig. 1 Neural Network Structure

5 Conclusions

We used neural networks to assess the predictive separate and combined power of accounting and market measures in predicting firm future financial sustainability for companies listed in the stock market of Qatar. Our results indicate that whereas accounting-based measures are better predictors for firm future earnings, market-based measures are more powerful predictors for future sales growth rate. Combining the accounting and market measures into a hybrid model improved the prediction of both measures of the firm future financial sustainability. However, the hybrid model achieved much better results for predicting firm earnings in comparison firm sales growth rate. We believed, the last finding is due to the noise embedded in the market variables, by their very nature.

References

1. Kuo, R. J., Chen, C. H., Hwang, Y. C.: An Intelligent Stock Trading Decision Support System through Integration of Genetic Algorithm Based Fuzzy Neural Network and Artificial Neural Network. Fuzzy Sets and Systems. 118, 21-45 (2001)
2. Yoon, Y., Swales, G.: Predicting Stock Price Performance: A Neural Network Approach. In System Sciences, 1991. In: 24th Annual Hawaii International Conference, vol. 4, pp. 156-162. IEEE (1991)
3. Bao, Y., Lu, Y., Zhang, J.: Forecasting Stock Price by SVMs Regression. In Artificial Intelligence: Methodology, Systems, and Applications, pp. 295-303. Springer, Berlin Heidelberg (2004)
4. Pai, P. F., Lin, C. S.: A Hybrid ARIMA and Support Vector Machines Model in Stock Price Forecasting. Omega. 33, 497-505 (2005)
5. Wang, Y. F.: Predicting Stock Price Using Fuzzy Grey Prediction System. Expert Systems with Applications. 22, 33-38 (2002)
6. Hadavandi, E., Shavandi, H., Ghanbari, A.: Integration of Genetic Fuzzy Systems and Artificial Neural Networks for Stock Price Forecasting. Knowledge-Based Systems. 23, 800-808 (2010)
7. Kim, K. J., Han, I.: Genetic Algorithms Approach to Feature Discretization in Artificial Neural Networks for the Prediction of Stock Price Index. Expert Systems with Applications. 19, 125-132 (2000)
8. Cheng, C. H., Chen, T. L., Wei, L. Y.: A Hybrid Model Based on Rough Sets Theory and Genetic Algorithms for Stock Price Forecasting. Information Sciences. 180, 1610-1629 (2010)
9. Olson, D., Mossman, C.: Neural Network Forecasts of Canadian Stock Returns using Accounting Ratios. International Journal of Forecasting. 19, 453-465 (2003)
10. Easton, P. D., Monahan, S. J.: An Evaluation of Accounting-Based Measures of Expected Returns. The Accounting Review. 80, 501-538 (2005)
11. Reisz, A. S., Perlich, C. (2007). A Market-Based Framework for Bankruptcy Prediction. Journal of Financial Stability, 3(2), 85-131.
12. Agarwal, V., Taffler, R.: Comparing the performance of market-based and Accounting-Based Bankruptcy Prediction Models. Journal of Banking and Finance. 32, 1541-1551 (2008)
13. Li, M. Y. L., Miu, P.: A Hybrid Bankruptcy Prediction Model with Dynamic Loadings on Accounting-Ratio-Based and Market-Based Information: A binary Quantile Regression Approach. Journal of Empirical Finance. 17, 818-833 (2010)
14. Hillegeist, S. A., Keating, E. K., Cram, D. P., Lundstedt, K. G.: Assessing the Probability of Bankruptcy. Review of Accounting Studies. 9, 5-34 (2004)
15. Merton, R. C.: On the Pricing of Corporate Debt: The risk structure of interest rates*. The Journal of Finance. 29, 449-470 (1974)
16. Altman, E. I.: Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance. 23, 589-609 (1968)
17. Ohlson, J. A.: Financial Ratios and the Probabilistic Prediction of Bankruptcy," Journal of Accounting Research. 109-131 (1980)
18. Baixauli, J. S., Alvarez, S., Módica, A.: Combining Structural Models and Accounting-Based Models for Measuring Credit Risk in Real Estate Companies. International Journal of Managerial Finance. 8, 73-95 (2012)
19. Das, S. R., Hanouna, P., Sarin, A.: Accounting-based versus Market-Based Cross-Sectional Models of CDS Spreads. Journal of Banking and Finance. 33, 719-730 (2009)
20. Cao, Q., Parry, M. E., Leggio, K. B.: The Three-Factor Model and Artificial Neural Networks: Predicting Stock Price Movement in China. Annals of Operations Research. 185, 25-44 (2011)
21. Baba, N., Kozaki, M.: An Intelligent Forecasting System of Stock Price Using Neural Networks. In: Neural Networks, IJCNN, International Joint Conference, vol. 1, pp. 371-377. IEEE, (1992)
22. Chang, T. S.: A Comparative Study of Artificial Neural Networks, and Decision Trees for Digital Game Content Stocks Price Prediction. Expert Systems with Applications. 38, 14846-14851 (2011)
23. Di, X., Dajun, M., Yuanxi, L.: The Application of Neural Networks to Stock Price Forecasting. Systems Engineering-Theory and Practice. 11, (1998)
24. Dutta, G., Jha, P., Laha, A. K., Mohan, N.: Artificial neural network models for Forecasting Stock Price Index in the Bombay Stock Exchange. Journal of Emerging Market Finance. 5, 283-295 (2006)
25. Qatar Exchange, http://qatarexchange.net/
26. Johnson, R., Soenen, L.: Indicators of Successful Companies. European Management Journal. 21, 364-369 (2003)
27. Investopedia, http://www.investopedia.com/

The Periodic Signal Filtration Using the Robust Digital Filter Order Calculation Optimized by Approximation

Alexey Sergeev-Horchynskyi and Valeriy Rogoza

1 Introduction

Various technical problems require filtration of noisy signals. One of signal filtration options is low-frequency signal filtration using a lowpass filter which is designed to separate out the low-frequency component of a signal represented by a series of discrete noisy samples.

For the best extraction of the low-frequency signal component, it is important to choose the correct order of the digital filter. There exist a class of problems in which there is no information about the original noise-free signal; in this case, filter adjustment is executed by the use of unsupervised ("blind") learning [1].

In [2], the authors used optimized order calculation of Simple Moving Average (SMA) filter to filter periodic noisy signals retrieved from accelerometer. After applying the calculated orders of SMA filter, periodic oscillations corresponding to test human walking steps were separated.

In this work, we have developed software library to simulate signal generation and digital processing in order to assess the results of applying the optimized SMA filter order calculation based on approximation. The original periodic signals of different shapes were generated, and their modified copies were obtained by adding a random component (noise) to the original signal.

In *Filtered and Approximated Values Calculation* section, the calculations of filtered and approximated values for noised signal are described.

In *Optimal Filter Order Calculation* section, the calculation of the optimal SMA filter order based on minimal MAE with respect to the original noise free signal is described.

In *Optimized Filter Order Calculation* section, the calculation of the optimal SMA filter order based on minimal MAE with respect to the approximated signal is described.

In *Filtration of Periodic Signals* section, the results of experiments that were performed to compare the values of orders obtained by means of the optimized calculation and the optimal calculation are described.

In *Conclusions* section, the authors generalize the results obtained in *Filtration of Periodic Signals* section.

2 Creating a test environment

During the experiments, the program library for generating discrete periodic signals of different characteristics and processing the generated signals has been developed. The library structure is shown in Fig. 1.

The program library module objectives are as follows. The "Signal generating module" is intended to generate periodic signals and includes generating a periodic signal and a noise signal components. The "Module of a signal generation according to a function" implements a program generation of signals of the following shapes: sinusoidal, triangular, rectangular, sawtooth. The software implementation provides the possibility to adjust the amplitude, frequency and phase of the signal.

The "Module for generating a noise component" fits a signal waveform to one occurring during a real transmission. The module implements the technique for generating random numbers of the normal distribution and provides an ability to set the values of a noise component expectation and standard deviation.

To minimize the noise, the "Signal processing module" has been developed. The module implements the following algorithms: the simple moving average filtration (SMA); the repeated median fitting (approximation); the optimized discrete filter order calculation that minimizes the MAE value between the approximated and filtered signals; the optimal

A. Sergeev-Horchynskyi (✉)
NTUU KPI, ESC Institute of Applied System Analysis,
Kyiv, Ukraine
e-mail: alexey.sergeev@ymail.com

V. Rogoza
West Pomeranian University of Technology, Szczecin, Poland
e-mail: wrogoza@wi.zut.edu.pl

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_38,

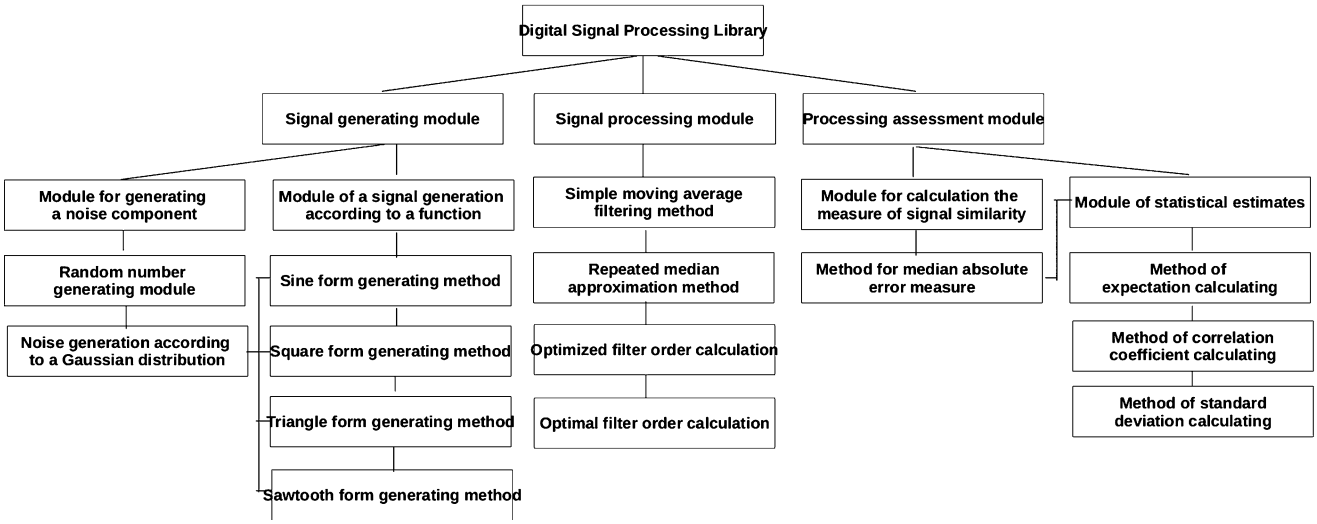


Fig. 1. The structure of the developed discrete signal processing library

discrete filter order calculation that minimizes the MAE value between the filtered and original (noise-free) signals.

The "Processing assessment module" implements the following software techniques: the calculation of the measure of signals similarity (in our case MAE); calculation of an expectation; calculation of the correlation coefficient; calculation of the standard deviation.

The developed software library was used to generate a noisy discrete signals and subsequently process them. A series of experiments was performed to assess the possibility of using the optimized by approximation method for the purpose of calculating a close to optimal LPF order.

3 Filtered and approximated values calculation

In this paper, the average values of a specified number of sampling point values are considered as the filtered ones, which are calculated by means of the SMA calculation algorithm.

The calculation of the filtered value by means of the SMA can be expressed as follows [3]:

$$y_{f\_m}[k] = \frac{1}{2 \cdot m + 1} \cdot \sum_{i=-m}^{m} y[k+i] \tag{1}$$

where k is the sampling point $(k = \{m, m+1, m+2, \ldots, N-m\})$, m is the SMA filter order $(m \in Z)$, $y[k+i]$ is the value of a noisy signal at the sampling point $k+i$, $y_{f\_m}(k)$ is the filtered signal value at the sampling point k.

In this work in order to approximate signal samples, the empirical regression equation is used. Calculation of approximated values (conditional expectations) can be expressed by the following relation (2) [4]:

$$y_{a\_n}[i] = b_0 + b_1 \cdot i \tag{2}$$

$$b_1 = \underset{i=0\ldots n-1}{med} \; \underset{j=0\ldots n-1, j \neq i}{med} \left(\frac{y[i] - y[j]}{i - j} \right) \tag{3}$$

$$b_0 = \underset{i=0\ldots n-1}{med} \left(y[i] - b_1 \cdot i \right) \tag{4}$$

where $i = \{0, 1, 2, 3, \ldots, n-1\}$ is a sampling point number, n is an approximation order (a quantity of samples (values) of a noisy signal $n = \{2, 3, 4, \ldots\}$, i.e., size of approximation interval), $y[i]$ is the value of a noisy signal at the sampling point i, $y_{a\_n}[i]$ is the signal value after approximation.

4 The optimal filter order calculation

If the values of the original (noise-free) signal are known, we could calculate the optimal value of the SMA filter order to bring nearer the filtered noise signal to the original (noise-free) one as much as possible.

The optimal SMA filter order m can be calculated by minimizing the MAE value when comparing the original $y_{sig}[k]$ and filtered signal $y_{f\_m\_optimal}[k]$, i.e., using the modified for two signals relation [5]:

$$M_{MAE}\left(Y_{sig}, Y_{f\_m\_optimal}\right) = med_k \left| y_{sig}[k] - y_{f\_m\_optimal}[k] \right| \tag{5}$$

where k is the sampling time point, $y_{sig}[k]$ is the value of the original noise-free signal at the sampling point $k = \{m, m+1, m+2, m+3, \ldots, N-m\}$, $y_{f\_m\_optimal}[k]$ is the value of the filtered signal at the sampling point k, m – order of optimal SMA filter, N is the total number of sampling points.

5 The optimized filter order calculation

If you do not know the values of the original noise-free signal, it is necessary to use methods of "blind" filtration. One of them is the method of the digital filter order calculation optimized by approximation. The calculation of the SMA filter order optimized by approximation involves the following steps:

1. Choosing the order m of SMA filter and estimation of filtered values for sampling time points, in accordance with (1).
2. Choosing the approximation interval n of sampling time points and calculation of the approximated sample values for the time point intervals, in accordance with (2).
3. Calculating the MAE values between the approximated and filtered signals is determined by the following relation:

$$M_{MAE}\left(Y_{app}, Y_{f\_m\_optimiz}\right) = med_k\left|y_{app}[k] - y_{f\_m\_optimiz}[k]\right| \tag{6}$$

where k is the sampling time point, $y_{app}[k]$ is the value of an approximated signal at the sampling point $k = \{m, m+1, m+2, m+3, \ldots, N-m\}$, $y_{f\_m\_optimiz}$ is the value of the filtered signal at the sampling point k, m – order of the optimized SMA filter, N is the total number of sampling points of the original noisy signal.

Next, consider the results of experiments which aimed at comparing the values of the optimal and optimized orders of SMA filter.

6 Filtration of periodic signals

To assess degree of closeness of the SMA filter order to the optimal one obtained using the proposed method, a series of experiments with periodic signals of different shapes has been conducted. For experimental purposes, signals of four shapes (sine, triangle, square, sawtooth) were generated.

The generated signals characteristics (see Table 1) were brought nearer to the characteristics of the signals received from accelerometer Kionix KXTF9-1026 [6] during human walking.

Fig. 2 shows the values of SMA filter order for sinusoidal signal at different signal-to-noise ratios of [-4.0, -1.0] dB interval for the optimal order (by original noise-free signal) and the optimized order (by approximated signal).

From Fig. 2, one can see that the maximal difference between the optimal and optimized SMA filter orders is less than 3, i.e., the two calculation methods enable to obtain similar results.

Table 1. Characteristics of test signals

Signal amplitude, m/s^2	5
Signal frequency, Hz	1
Noise component distribution	Gaussian
Sampling frequency, Hz	120
Signal detection time, s	15

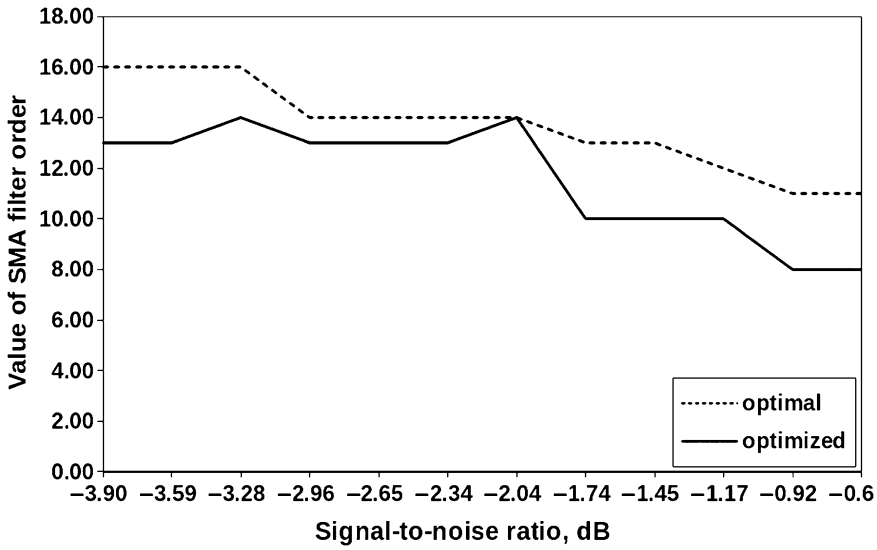


Fig. 2. The values of the optimal and the optimized orders of the SMA filter for sinusoidal signal

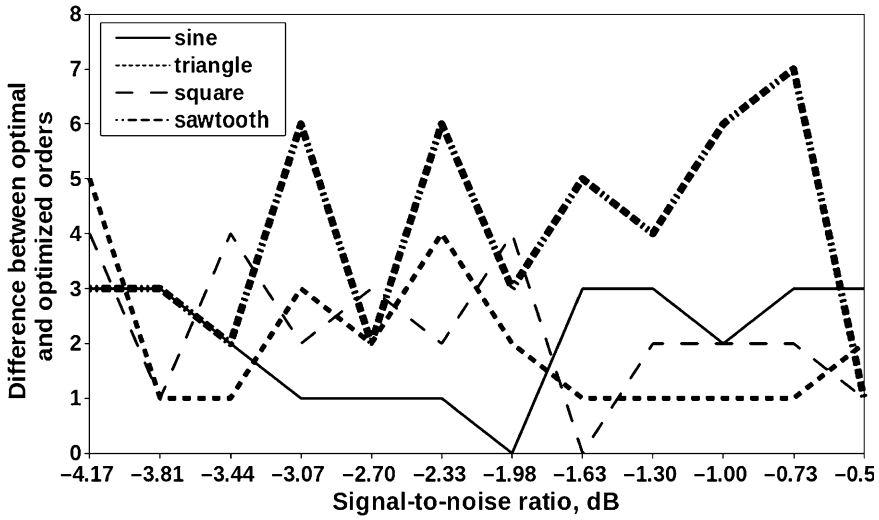


Fig. 3. The difference between the values of optimal and optimized orders of SMA filter for noisy signals of various shapes

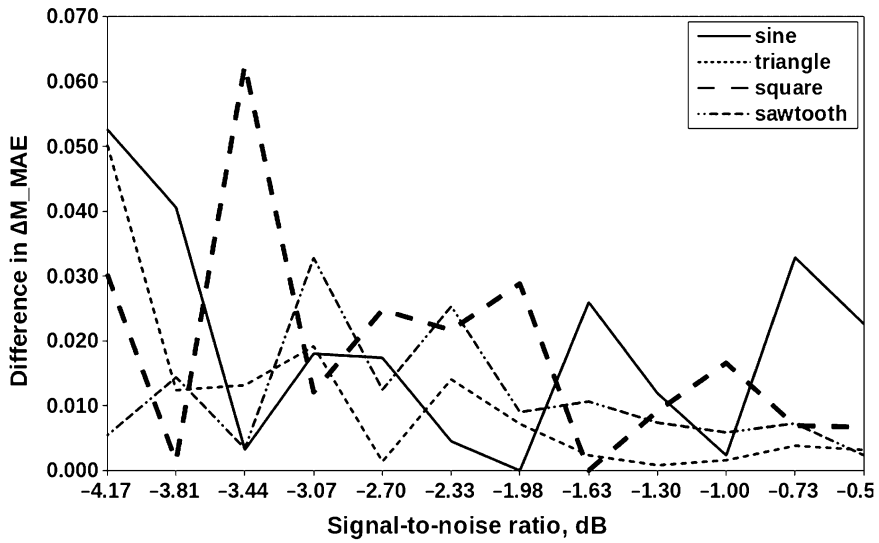


Fig. 4. The difference Δ_{MAE} between the optimal and optimized orders of SMA filter

Next, it was necessary to test the hypothesis that the signals of other shapes (triangular, square, and sawtooth) also allow us to calculate near-optimal values of a SMA filter order.

Fig. 3 shows the difference between the values of the optimal and optimized SMA filter orders for noisy signals of different shapes and various signal-to-noise ratios of the range [-4.0, -1.0] dB.

From Fig. 2 and Fig. 3, it follows that for optimized calculation of SMA order most of the calculated values differ from the optimal order in the range [0, 6].

Next, it was necessary to verify the assumption that the optimized SMA filter output values are near to the original noise-free signal values, i.e., the difference between MAE values is minimal.

The formula for calculation of the difference between the median absolute errors Δ_{MAE} has the following form:

$$\Delta M_{MAE} = M_{MAE}\left(Z, Y_{f\_optimiz}\right) - M_{MAE}\left(Z, Y_{f\_optimal}\right) \quad (7)$$

$$M_{MAE}\left(Z, Y_{f\_optimiz}\right) = med_k\left|z[k] - y_{f\_m\_optimiz}[k]\right| \quad (8)$$

$$M_{MAE}\left(Z, Y_{f\_optimal}\right) = med_k\left|z[k] - y_{f\_m\_optimal}[k]\right| \quad (9)$$

where k is the sampling point, $z[k]$ is the value of the original noise-free signal at the sampling point $k = \{m, m+1, m+2, m+3, \ldots, N-m\}$, $y[k]$ is the value of processed signal at the sampling point k, N is the total number of sampling points for Z (original noise-free signal) and Y (processed noisy signal).

Fig. 4 shows the difference Δ_{MAE} between the MAE values for different shape signals as a function of signal-to-noise ratio [-4.0, -1.0] dB.

Based on the data shown in Fig. 4, one can conclude that, for noisy signals of sine and triangular shapes with the value of signal-to-noise ratio greater than -3 dB, the use of optimized calculation of digital filter order is efficient and allows to calculate SMA filter order near to the optimal one (with MAE values within range [0.0, 0.03]).

7 Conclusions

Filtration of noisy signals is needed when solving technical problems, including processing of signals received from inertial sensors (such as accelerometer). Challenges of choosing the optimal order of the filter separating the useful signal from noise are caused by that sensitivity of sensors producing raw signals as well as the characteristics of these signals (including noise level) can vary within quite wide limits and are beforehand indeterminate.

When constructing a noisy signal filter, the method of low-pass filter order calculation optimized by approximation can be used.

In order to assess the possibility of applying the method for calculation of SMA filter order and for subsequent filtration of periodic signals with a signal-to-noise ratio [-4.0, -1.0] dB, the mathematical library composed of software modules was developed.

The developed library allowed to conduct a series of experiments in which noisy periodic signals of various shapes (sine, triangle, square, and sawtooth) were generated, corresponding optimal and optimized SMA filter orders (for given signal-to-noise ratios) were calculated, filtration with corresponding orders was executed, and estimated MAE values were compared.

After the difference of the values of the optimal and optimized SMA filter orders were estimated, it has been shown that these orders differ on average by values of interval [0, 6]. After all noisy signals were filtered, the mean difference of MAE values (with respect to original noise-free signals) belongs to [0.0, 0.03] interval.

The optimized SMA filter order calculation and subsequent filtration of square shaped signals with SNR less than -3.0 dB have been demonstrated the worse results with respect to sine, triangle and sawtooth shaped signals. The discussed method is recommended to filter signals that are similar to harmonic.

The developed library can be used for experimental processing of digital signals of various technical problems.

References

1. Sergienko A. Digital Signal Processing, 3-rd edition, BHV-Petersburg, St. Petersburg, 2011. — 592-593 pp.
2. Rogoza V., Sergeev A. The Comparison of the Stochastic Algorithms for the Filter Parameters Calculation, Advances in Systems Science, Vol. 240, Springer, Switzerland, 2014. — 244-245 pp.
3. Oppenheim A., Schafer R. Discrete-Time Signal Processing, 3-rd edition, Prentice-Hall, USA, 2010. — 17-18 pp.
4. Bernholt T., Fried R. Computing the update of the repeated median regression line in linear time, Information Processing Letters 88, Elsevier, Netherlands, 2003. — 111-112 pp.
5. Rousseeuw P., Croux C. Alternatives to the Median Absolute Deviation, Journal of ASA, Vol. 88, ASA, USA, 1993. — 1273-1274 pp.
6. KXTF9 Series: Accelerometers and Inclinometers, Kionix Inc., USA, 2011. — 1-2 pp.

A Reasoning System for Predicting Study Level based on User's Watching Behaviors

Jeonghyeok Kim, Jaemin Hwang, Sanggil Kang, and Nojeong Heo

1 Introduction

E - learning is one of important tools, which can help to improve self-directed learning habits of students [4]. Thus, it can play a critical role on the academic achievement of each student [1]. However, current E – learning systems are not satisfied by students yet because identical study materials are provided to students regardless of their intellectual levels. In addition, they provide lecture videos to students with predetermined academic level of students as the students choose lecture videos. In this case, the efficiency of education is getting less. This problem is same to teachers too. Teachers cannot check the reaction of students on E – learning, so do not recognize the academic level of each student in their class [2].

As a conventional method for inferring user's academic level, there is the case-based learning method [5]. It first collects the information about students such profile of students and their interest courses or levels from their teachers' survey. Based on the information, it recommends appropriate lecture videos and provided to the students. Similarly, there is the item-based learning method [8] using students' evaluation for various tests for each lecture. For those methods, we cannot infer their academic level without explicit students' information. The requirement of the explicit information can be unpleased by students [3].

To solve this problem, we need to infer the academic level based on user's implicit behaviors during watching lecture videos such skipping and rewinding frames of videos and playback time. It is because the user's behaviors can relate to the level of difficulty for the lecture. In this paper, we develop a study level reasoning system to estimate the student's academic level using video watching behavior and a statistical analysis of the history of watching.

The remainder of this paper is organized as follows. Section 2 presents the related work about online learning and reasoning method of student's academic level. Section 3 describes a main algorithm how to reason student's academic level in regular. sequence. Experimental results in Section 4 demonstrate study level reasoning system and we explain its experiment environment. Finally, conclusions are given in Section 5.

2 Our Reasoning System of Students' Study Level

2.1 Architecture of Our Reasoning System

Fig. 1 is the architecture of our reasoning system developed based on user's implicit behaviors which is consisted of two phase such as the extraction and reasoning. In the extraction phase, we calculate the relative time for students and the overall learning time for each lecture video to analyze the learning patterns of students based on watching behaviors (watching time, skipping, and rewinding). In the reasoning phase, we consider relative viewer ratings to fit distribution map, analyze their learning behaviors, and reflect students' analysis data. Finally, we make a recommendation list for providing a personalized lecture video. After students watched the lecture video, they can rate a satisfaction level of the lecture video for the improvement of the system.

2.2 Reasoning Algorithm

The procedure of our reasoning algorithm for predicting students' study level based on his/her implicit behaviors during watching lecture videos can be depicted as seen in

J. Kim (✉) • J. Hwang • S. Kang
Department of Computer Engineering, Inha University, Incheon, The Netherlands
e-mail: sspwiz@inha.edu; nulpis1@gmail.com; sgkang@inha.ac.kr

N. Heo
Department of Information and Communication Engineering, Dongyang University, Yeongju, South Korea
e-mail: nheo@dyu.ac.kr

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_39,

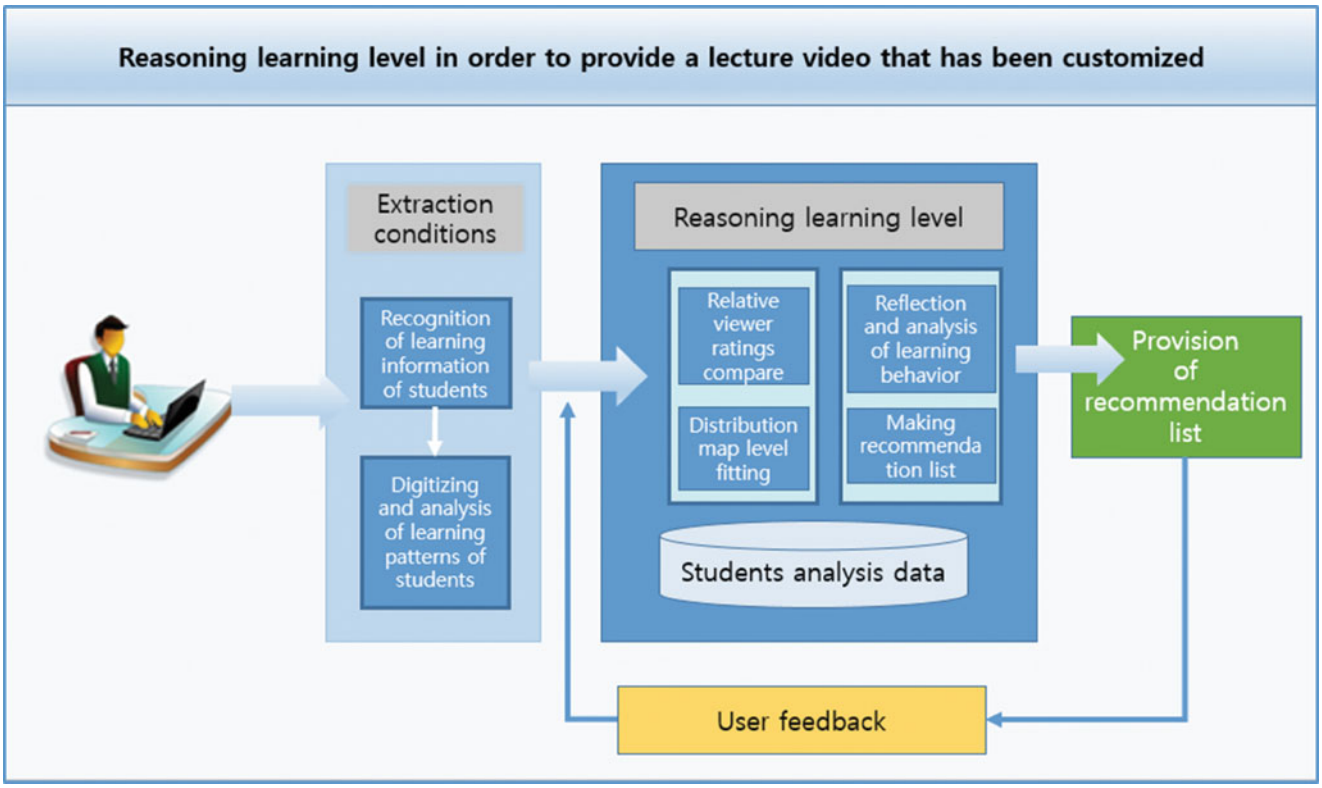


Fig. 1 Architecture of our reasoning system

Fig. 2. First, we have to find the relative ratio of one student's playback time of a lecture video to the total summation of the lecture video for every student, as seen the equation (1).

$$\frac{t}{T} \tag{1}$$

where t is one student's playback time of a lecture video, and T is all students playback time of the video lecture. For the equation (1), it does not take into account the learning behavior of the students own when watching video lectures. We consider the average lecture video playback time for each student from the lecture video watching history. To do that, the equation (1) is modified as seen in the equation (2).

$$\frac{t}{T\mu} \tag{2}$$

where μ means a ratio of watching time duration of each student to that of all students, which is computed as seen in the equation (3)

$$\mu = \frac{\Sigma_i^n t_i}{\Sigma_i^n T_i} \tag{3}$$

where t^{i} is the playback time of the i^{th} lecture video, T^{i} is the sum of playback time of all students who watched the i^{th} lecture video. In addition, n is the number of lecture videos watched by the student.

As mentioned in the previous section, students' implicit behaviors during watching the videos such as rewinding and skipping play an import role to estimate the interest of students for the videos. Thus, we apply the students' behaviors to predict students' study level as follows:

We calculate the standard deviation (σ) and average ($\tilde{\mu}$) of students' watching duration for all students who watched the video. If the equation (2) lies in the interval $[\tilde{\mu} - \sigma \leq \frac{t}{T_\mu} \leq \tilde{\mu} + \sigma]$, then we determine that the lecture video is appropriate to the student. In the case that the lecture video is determined as not appropriate level to the student, then we determine the student's study level differently according to the student's watch behavior.

For $\frac{t}{T_\mu} \leq \tilde{\mu} - \sigma$ case, if the student took skip actions frequently, then we can consider that the video is easier than the student's study level. Otherwise, we can consider the student has hard time to keep watching the video. It is difficult. Criterion for determining skip action is a case of occurring skip count of the lecture video more than average skip count of the students who watched the videos. About $\frac{t}{T_\mu} \geq \tilde{\mu} - \sigma$ case, if rewind count is increased, it is determined that lecture video is difficult.

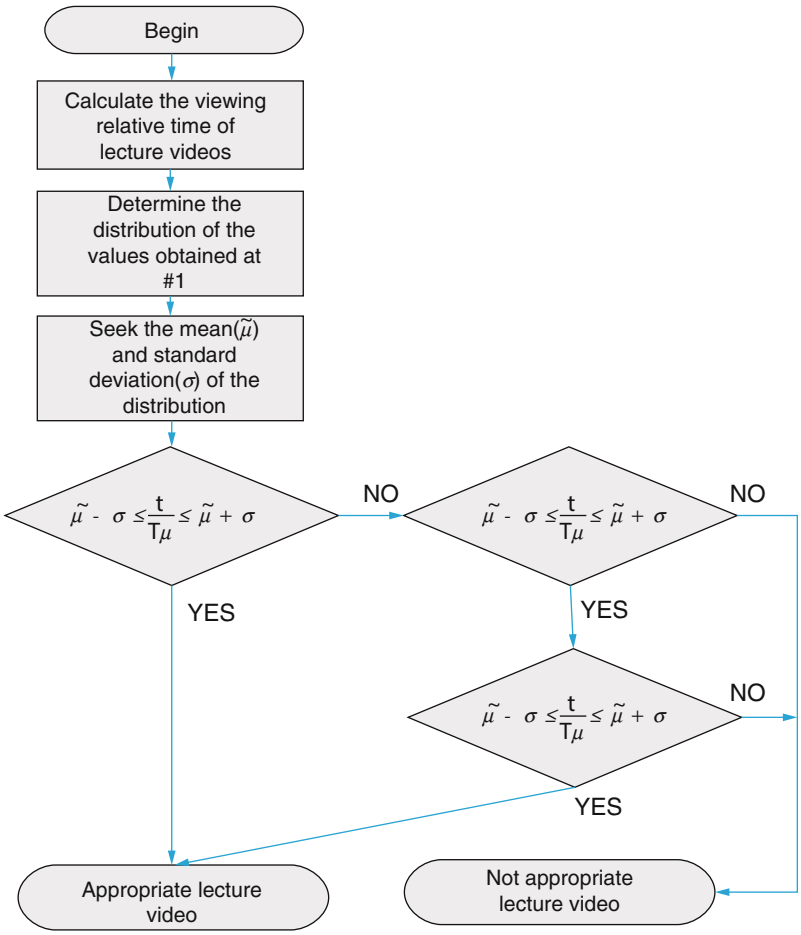


Fig. 2 Flow chart of our reasoning algorithm for predicting students' study level

2.3 Feedback from users

We generate the recommendation list based on result that is determined on the lecture time distribution and learning behavior between one student and all students. However, sometimes, the learning participation rate can differ each other because it depends on the relation rate of school curriculum or test. If the students evaluate the importance of the videos which they watched, we reflect their rates to the next recommendation process in order to adjust the level of lecture videos.

3 Demonstration and Evaluation

Fig 3 is the prototype of the reasoning system for predicting study level based on user's watching behaviors, which consists of two panels. The panel on the left shows the information of a selected video with lecture difficulty. The panel on the right is its lecture video. According to the degree of difficulty, we made different colored button, e.g., the red, yellow, and green means a high, easy, and normal level, respectively.

Using 100 high school students' video watching history, we developed the prototype and evaluated it. The watching history includes two courses (social and math) divided into three levels, respectively by beginner, intermediate, and advanced, with total 30 videos. To show the feasibility of our system, we evaluate our system based on students' satisfaction survey about the cases that our system is applied and it is not applied. To do that, we first have students watch three lecture videos during 40 minutes, which they want to watch. Then, we collected students' watching behaviors. Fig. 4 is the survey results. As seen in the figure, the number of students who answered 'good' for the case our system is applied and the other case our system is not applied is 45 students and 77 students, respectively. From the results, our system is better than the system without applying our reasoning algorithm.

4 Conclusion

In this paper, we proposed the reasoning system for predicting students' study level based on the students' watching behavior history. The contribution of our method

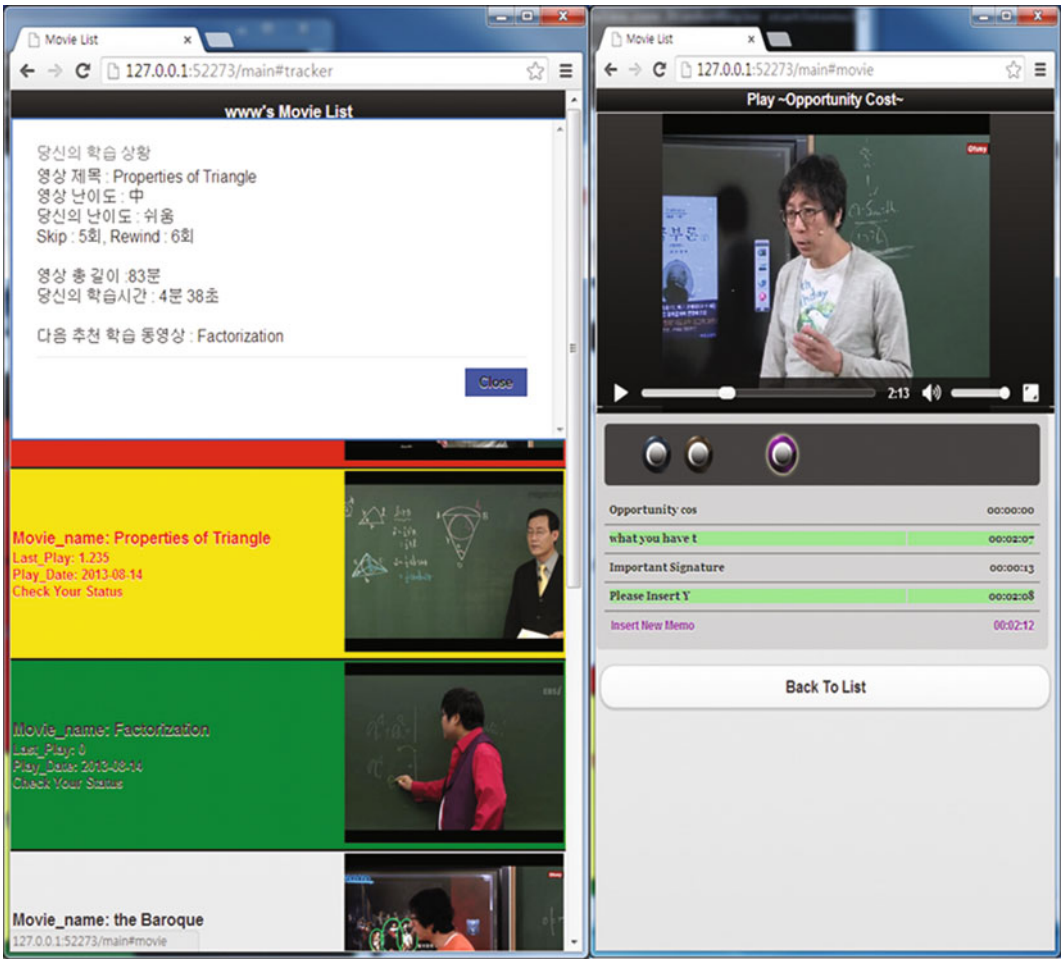


Fig. 3 Prototype of our system

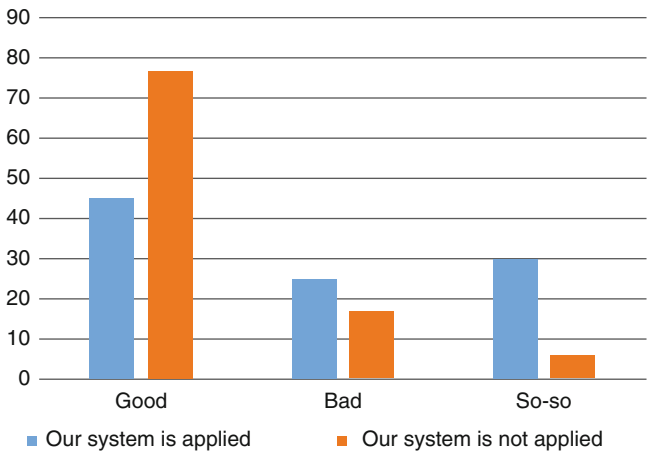


Fig. 4 Survey results

can be summarized as follows. First, we developed a prototype for our system and showed the feasibility of our system based on the positive responses from students as shown in the experimental section. Also, our system can release students' burden because our system do not need explicit students' information for reasoning students' study level.

Our system was developed based on the students whose learning environments are similar and the degree of difficulty of the contents of lectures. Therefore, we need to test our system using various students' watching history as a further work.

This research was supported by the MSIP(Ministry of Science, ICT & Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2014-H0301-14-1019) supervised by the NIPA (National IT Industry Promotion Agency)"

Reference

[1] J. Y. Namgoong, Y. J. Kim, Y. B. Kim, High School comprehensive analysis study, level analysis and the reality of school education (2012.12), Korea Education Development Institute

[2] Dianne H.B. Welsh, Mariana Dragusin, The New Generation of Massive Open Online Course (MOOCS) and Entrepreneurship

Education, Small Business Institute Journal, 2013, Vol. 9, No. 1, 51–65
[3] S.Y. Choi, H.J. Yang, E-learning ontology-based system to support their leading learning, Korea Computer Education Journal (2010.9) pp. 29-33
[4] Steve Cooper, Mehran Sahami, Reflections on Stanford's MOOCs, Communications of the ACM, 2013, Vol. 56, Issue 2, pp 28–30
[5] S. G. Kim, Y. H. Kim, H. J. Yoon, Experimentation and evaluation studies of case-based learning for the improvement of learning outcomes, Korea Computer Education Journal (2011.11), pp. 53–64
[6] Kai Yu, Anton Schwaighofer, Volker Tresp, Xiaowei Xu and Hans-peter Kriegel, "Probabilistic Memory-Based Collaborative Filtering", IEEE Transactions on knowledge and data Engineering, vol. no.1 January 2004
[7] Zan Huang, Daniel zeng, Hsinchun Chen. "A Comparison of Collaborative-Filtering Recommendation Algorithms for E-commerce", IEEE Intelligent Systems, p. 69–78, 2007
[8] M. Claypool, A, Gokhale, T. Miranda, P. Mumikov, D. Netes, and M. Sarti, "Combining Content-Based and Collaborative Filters in an Online Newspaper," ACM SIGIR Workshop on Recommender Systems, Berkeley, CA, 1999

Temporal Constraints and Sub-Dimensional Clustering for Fast Similarity Search over Time Series Data. Application to Information Retrieval Tasks.

Sidahmed Benabderrahmane

1 Introduction

In the two last decades, online monitoring of streaming data has become more and more important, as the number of available sources of information increased dramatically in different domains (eg: Electronic health records, Network monitoring, Financial application, Meteorology,...) [1]. Data represented as time series become ubiquitous, and recently there has been an explosion of interest in time series data mining in various application domains [2]. To cope with the lack of efficiency, the data dimensionality is sometimes reduced by transforming the original time series into multi-dimensional symbolic sequences [3]. Consequently, retrieval and search tasks into time series databases is still important, nowadays. Similarity search over time series is one of the most important tasks in streaming data mining. In fact, it is still a challenge to discover pertinent pattern-sequences that are recurrent over the overall sequences, and to find temporal associations among these frequently occurring patterns. Previous works on similarity search have been performed in different contexts such as diagnostic or failure detection of industrial materials [4, 5]. In whole query matching, a time series given as query is entirely compared to every time series of a database. The series should have same length, and a similarity measure is used to retrieve either a most similar time series or the top-k ranked time series [6]. However, theses methods suffer from a lack of flexibility of the used similarity measures, a lack of scalability of the representation model, and a penalizing runtime to retrieve the information. Moreover, in some real world applications, one can be interested in retrieving specific interesting subsequences that are frequently present at different instants. For example, if we consider the problem of network analysis, the user may be interested, in a reasonable runtime, by identifying a collection of repeated events which are similar to a suspicious behavior given as an input query. Furthermore, the user may also be interested in discovering hidden temporal relationships between the discovered patterns either in the same dimension or in multiple dimensions (e.g. overlapping or successive events,..).

Motivated by these observations, we have designed a framework tackling the query by content problem on streaming data, ensuring (i) fast response time, (ii) multi-level information representation, and (iii) representing temporal associations between extracted patterns. During the preparation step, all the multi-valued time series present in the database are transformed into a multi-resolution symbolic representation thus ensuring a lower dimensionality. Then, to accelerate the similarity search, our model creates an index over recurrent patterns in the time series collection. These patterns can be generated by different techniques. We have tested a segmentation using sliding windows, and the extraction of substrings from a generalized suffix trees. Finally, the extracted patterns are grouped by clustering and the resulting clusters are indexed in a table within their centroids. The retrieval step, consists (i) in symbolizing the time series given as input query and (ii) in retrieving the cluster whose centroid is the most similar to the symbolic query. The patterns present in this cluster are returned as candidate motifs. We expect the retrieval step to be very fast since it compares only a limited number of centroids to the input symbolic query rather than a brute force which would compare all subsequences in the database to a numerical input query.

2 Related Work

2.1 Time Series Data Mining:

Time series mining have risen many problems in the two last decades [2]. Authors such as *Chakrabarti et al*. [7] considered the indexing problem, on whole series. Given a time series

S. Benabderrahmane (✉)
Paris 8 University, IUT Montreuil. 2 Rue de la Liberte, 93200 Paris, France
e-mail: sidahmed.benabderrahmane@gmail.com

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_40,

query T_q, the goal is to retrieve the most similar series from a database with respect to a distance function $Dist(T_i, T_q)$. The results can be computed in two ways: top-k and nearest query. In the case of top-k queries, series are ranked regarding their similarity with the query and only the k best time series are kept. On the other hand, given a threshold δ, and a query T_q, the nearest query return the series T_j such that $Dist(T_j, T_q) \leq \delta$. In such applications, problems rise when choosing the similarity measure and the way of representing the time series itself. For example, in the case of real valued time series, many similarity measures have been used such as the Euclidean Distance and Dynamic Time Warping. However, while some of these measures generate good results for short sequences, they fail for long time series database [8].

Time series clustering is an important task in mining from data streams. It aims to generate natural grouping of the series under certain hypothesis [9]. Real time anomaly detection over streaming data represents another important challenge. Having a set of time series T_q, supposed to be normal, and an another unlabeled time series T_i, decide if T_i contains abnormal or interesting occurrences [10].

Besides the problems related to the complexity of the used similarity measures in all these mining methods, runtime is still being a major problem. Indeed, with the emergence of Big Data models, data are stored in huge DB, hence the need for fast information retrieval techniques.

2.2 Symbolic Representation:

Many discretization methods have been proposed in the literature to encode time series in symbolic strings [11],[7], [3],[12]. Among these methods, SAX (Symbolic Aggregate Approximation) [3] is very popular and has been widely used for similarity search and clustering. SAX maps a time series $T = \{t_1, t_2, \ldots, t_l\}$ to a sequence of symbols from an alphabet of size $s = |\Sigma|$. SAX divides the input series into m successive windows [3] of equal size (l), and associates a symbol to each mean value of a window position. An improved version of SAX called 1D SAX has been proposed for representing the time series as a binary sequence [13].

Recently *Megalooikonomou* et al. introduced MVQ (Multi Resolution Vector Quantization) [12],[14]. They showed that MVQ is more efficient than SAX for classification problems. The MVQ approximation represents a time series in a hierarchical model at multiple resolutions. MVQ begins by extracting the set of subseries of length l from a time series learning set. Then, these subseries are clustered and representatives of some selected clusters are retained as *codewords* (*cw*) and a symbol is associated to each codeword gathered in a so-called *codebook c b* [12]. In order to keep both local and global information and improve the accuracy, MVQ is able to produce multiple representations at different resolutions.

3 Query By Content Approach: i-StreamMiner

3.1 Overview:

The objective of our work is to tackle the query by content problem of subsequences over streaming data. Our line of reasoning consists of four major steps fixed in the road map:

Symbolic encoding: time series are transformed from the domain of real numbers to lower dimensionality symbolic representation. This step is performed with multi resolutions, i.e. different size of the alphabet Σ, thus ensuring multi level of abstraction. For instance, SAX or MVQ can achieve such transformation.

Patterns extraction: patterns are extracted from symbolic series either by segmentation with a sliding window, which will generate subsequences with same length, or by extracting substrings using suffix trees to generate patterns with difference length.

Patterns indexation: The generated patterns are grouped into clusters. An index is created for the centroids of all clusters. This index allows fast retrieval of similar patterns of a given query, by identifying the most similar centroid and matching patterns that it aggregate to the input query. This method avoids sequential visiting of all patterns present in the database, and hence ensures fast access. Many different similarity measures can be used for the clustering. We have tested many similarity measures described hereafter. Mainly, we aim to evaluate to which extent indexing temporal patterns in an index structure can yield faster retrieval rather than performing entire scan in the *DB* (BruteForce). Firstly, we will compare our *i-StreamMiner* method to the *BruteForce* approach in term of queries time processing. Subsequently, we will also assess whether the used encoding methods, the similarity measures, and the patterns extraction methods impact the accuracy of the querying process. We detail these steps in the following subsections.

3.2 Multi Resolution Encoding:

Multi resolution encoding aims to choose for both SAX or MVQ, different features of codewords (length of window) and codebooks (size of the alphabet). Long discretization windows associated with small alphabet size yield low resolutions, and a high level abstraction of the values recorded in the original time series. Conversely, small windows and big alphabet yields to high resolutions and encoding with less loss.

3.3 Pattern Sequences Generation:

Once time series are encoded into strings, subsequences can be generated for using theme in the stage of indexes construction. Since we performed multi resolutions encoding, we run the following algorithms at each resolution level, thus ensuring the extraction of relevant subsequences at different levels. Consequently, the resulting set P will contain for each dataset and for a each selected resolution level, the list of all temporal patterns extracted with the two techniques we want to evaluate and which are detailed below.

Pattern Generation with Segmentation of Time Series: The segmentation function aims to extract fixed length subsequences from the symbolic time series. A window of length w is slided along the strings of the input series, and the subsequence in the same position of the sliding windows is retrieved and added to the set of all patterns P for further processing. Then, the windows is moved over along the sequence using a moving step α given as parameter. If $1 \leq \alpha < |w|$ consecutive windows overlap.

Pattern Generation using Suffix Trees: A suffix tree is a data structure used to represent all substrings of a string as a tree [15]. Each internal node should have at least two descendant, and each edge is labeled by a substring. Moreover, different edges starting from the same node can not have same prefix. Constructing such a tree for any string S takes time and space proportional to the length of S. Once constructed, several operations can be performed quickly on the tree, for instance locating a substring in the sequence S. Suffix trees are very often used for searching patterns in a sequences. It allows the identification of a motif of length l in a time linearly dependent to its length $O(l)$. To that aim, the search starts from the root node, and follows the branch having as beginning label the first symbol of the motif. Then, by moving along the path of this branch, the search traverses new nodes by reiterating the process for the symbols remaining in the motif. The stopping condition is reached when there is no suffix tree node remains for the current motif or this symbol cannot match any edge label. Leaves of the tree represent the occurrences of a motif in a string.

3.4 Index Temporal Patterns by Subdimensional Clustering:

In the information retrieval field, search engines collect, regroup and index documents on the basis on their semantic similarity. The main goal of creating an index is to optimize the performance and speed of finding relevant motifs for a search query. Relying on an appropriate organization of patterns (data), it facilitates fast and accurate information retrieval. Without an index, searching a subsequence would necessitate to scan sequentially every time series in the database (Brute Force), which would require considerable time and processing power. In our case, the extracted patterns are grouped, using a clustering algorithm and a similarity measure *Sim*. The entry keys in the index are the centroids of the obtained clusters C. Each centroid gives an access to all the patterns present in the cluster labeled by this centroid. If the patterns in a cluster are highly similar, so they could be frequent subdimensional motifs. We can represent this model by a *Voronoi* diagram, where the search space is divided in different regions. The center of each region gives an access to elements of this region. This modeling results in a fast retrieval runtime. Indeed, instead of performing a sequential scan on the entire *DB* to find the most similar patterns, this method determines which centroid is the nearest to the input query. Then, it compares the query to the elements of the cluster associated with the selected centroid using similarity *Sim*.

Algorithm 1 Patterns index construction

Require: P: the set of patterns from a database DB, Sim: a similarity measure defined on P.
Ensure: C=$\{c_1, c2, .., ck\}$: the set of generated clusters,
$Index$: the set of the medoids of clusters c_i.
//Calculate the similarity matrix.
for Each pair of patterns $p_i, p_j \in P$ **do**
$SimMat[i][j]$=Sim(p_i, p_j)
end for
//Perform a clustering using $SimMat$, and optimize it with the Kelley method to determine k, the optimal number of clusters.
$C \leftarrow$GetClusters($SimMat$)
Optimize(C)
//Save the centroids of the produced clusters in Index.
$Index \leftarrow$GetCentroids(C)
return $Index$

The index computation method is detailed in Algorithm 1. The major steps of this algorithm are:

Calculating similarity between subsequences: Algorithm 1 requires as input a set P of all extracted subsequences from the database DB. Firstly, we must choose a function Sim to compute the similarity between pairs of patterns $p_i, p_j \in P$. This yields a similarity matrix $SimMat$. Since elements of P are symbolic strings, a text-based similarity measure can be chosen.

Similarity analysis between symbolic sequences is a key problem in Natural Language Processing (NLP) and bioinformatics [16]. According to [8] two kinds of similarity measures can be applied to time series, namely structure-based and shape-based similarity measures. This categorization is based on how features are extracted and how similarity is computed. While shape-based methods determine the similarity between two series by comparing their individual points (e.g. Euclidean, DTW), the structure-based approaches look at higher level structures (e.g. TF-IDF) [8].

In the present work, we have experimented both shape and structure-based measures. Concerning the first category, the *Jaccard* index computes the relatedness between two given subsequences. It is based on set-comparison theory, and it is defined as the size of the intersection divided by the size of the union of the sample sets. Given two patterns p_i, $p_j \in P$, $SimJac(p_i, p_j) = \frac{|p_i \bigcap p_j|}{|p_i \bigcup p_j|}$, where $p_i \bigcap p_j$ designates the set of symbols shared by these two subsequences, and $p_i \bigcup p_j$ is their union. The advantage of this measure is that it is normalized, very simple and easy to implement. The second measure tested in our experiments is based on a representation widely used in text mining problems. The Vector Space Model (VSM) proposed by Salton et al. [17] suggests to represent each document as a vector. Each dimension of the vector is a term in the global vocabulary. A value is associated to each term reflecting its importance in the document collection. In our case, an encoded time series could be considered as a document, and the symbol t_i (codeword) in it as a document term. This bag of words representation is widely accepted for documents [8]. The weighting scheme that we used for associating a value to a symbol, mainly relies on the information content (IC) of the different symbols. The information content of a codeword is based on its frequency, or probability, of occurrence in a sequence of a given time series. It reflects a semantic quantification of a codeword. We defined it as the negative logarithm of the probability of a codeword, $IC(t_i) = -Log(p(t_i))$. After having calculated the IC for each symbol in each encoded time series, we propose a new semantic similarity measure $SimSemSeq$ which is defined on temporal patterns of the set P: $SimSemSeq\ (p_i, p_j) = \frac{\Sigma_{t \in p_i \cap p_j} IC(t)}{\Sigma_{t \in p_i \cup p_j} IC(t)}$.

The third measure that we propose to use in our work is shape-based one, and it is defined through *Kappa* statistics [18]. Kappa measures the percentage of data values in the main diagonal of a contingency table and then adjusts these values for the amount of agreement that could be expected due to chance alone. It is a chance-corrected measure of co-occurrence between two sets of categorized data. Since the 1D-SAX encoding approach [13] generates binary representation of time series, thus *kappa* statistics is more suitable for such binary categorical scale than the Pearson correlation, which is typically used for continuous, non-categorical data. Other shape-based (elastic) similarity measures such as Levenshtein (*SimLev*) [8] and the longest common sub string (*SimLCSS*) [19, 20] defined for string comparison and signal processing, are also used in this work for the evaluation of time series data clustering.

Having defined the metrics, we can now generate the pairwise similarity matrices *SimMat* between all pairs of patterns in each set P, for the clustering purposes.

Sub dimensional clustering: Several clustering methods are described in the literature. In our case, we used Partitions Around Medoids algorithm (*PAM*) which is a classical partitioning technique that clusters the data set of n objects into k clusters given a priori, and gives for each cluster its representative medoid.

4 Experiments

4.1 Datasets and Experimental Settings:

We aim to study the impact of the choice of the encoding method (SAX, MVQ, 1D SAX), the similarity measure (*SimJac*, *SimSemSeq*, *SimKappa*, *SimLev* and *SimLCSS*) and the pattern extraction method (segmentation or suffix tree extraction) on runtime and accuracy at different resolutions. C++ and the R programming languages have been used to implement the methods. A duo core machine with 2 GHz and 4 GB RAM has been used during the evaluations.

In our experiments, we used a collection of four datasets containing multivariate time series downloaded from the UCR repository [21]. Table 1 gives the number of time series (Nb TS) and time variables (Nb Var) in each selected dataset. This two parameters change from a dataset to another to ensure diversity in the database. At the end of this stage, we save all the real valued time series T_i in the database DB. Concerning the encoding step, we combined both high and small resolution representations of times series to ensure a multi level mining processing. As illustrated in Table 1, for each dataset, time series are encoded using SAX, 1D SAX and MVQ, from small to high resolution. To facilitate the interpretation of future

Table 1: The four collections of time series datasets forming the database *DB*. Symbolic representations of each datasets by MVQ and SAX are given with different resolutions. For 1D SAX approach, we kept the top four resolutions only since it was useless to go beyond.

	Datasets							
	YOGA	ECG	BEEF	GUN	YOGA	ECG	BEEF	GUN
Dimensions (Nb TS, Nb Var)	(300,426)	(100,96)	(30,470)	(50,150)	(300,426)	(100,96)	(30,470)	(50,150)
	Resolutions using MVQ (cw,cb)				Resolutions using SAX (l,s)			
resolution 1	(10,5)	(10,5)	(10,3)	(10,5)	(32,4)	(16,4)	(64,4)	(32,4)
resolution 2	(10,10)	(5,5)	(10,5)	(10,10)	(32,16)	(16,32)	(64,64)	(16,8)
resolution 3	(10,15)	(5,10)	(10,10)	(5,5)	(16,4)	(8,8)	(32,64)	(16,64)
resolution 4	(10,20)	(5,15)	(10,15)	(5,10)	(16,16)	(8,16)	(16,16)	(8,8)
resolution 5	(10,30)	(5,20)	(10,20)	(5,15)	(16,32)	(4,4)	(16,32)	(8,64)
resolution 6	(10,40)	(5,25)	(10,40)	(5,20)	(8,4)	(4,8)	(16,64)	(4,4)
resolution 7	(10,50)	(5,30)	(5,30)	(5,25)	(8,16)	(4,32)	(4,16)	(4,32)
resolution 8	(5,10)		(5,40)		(8,64)		(4,32)	
resolution 9	(5,20)				(8,32)			
resolution 10	(5,30)				(4,4)			
resolution 11	(5,40)				(4,16)			
resolution 12	(5,50)				(4,32)			

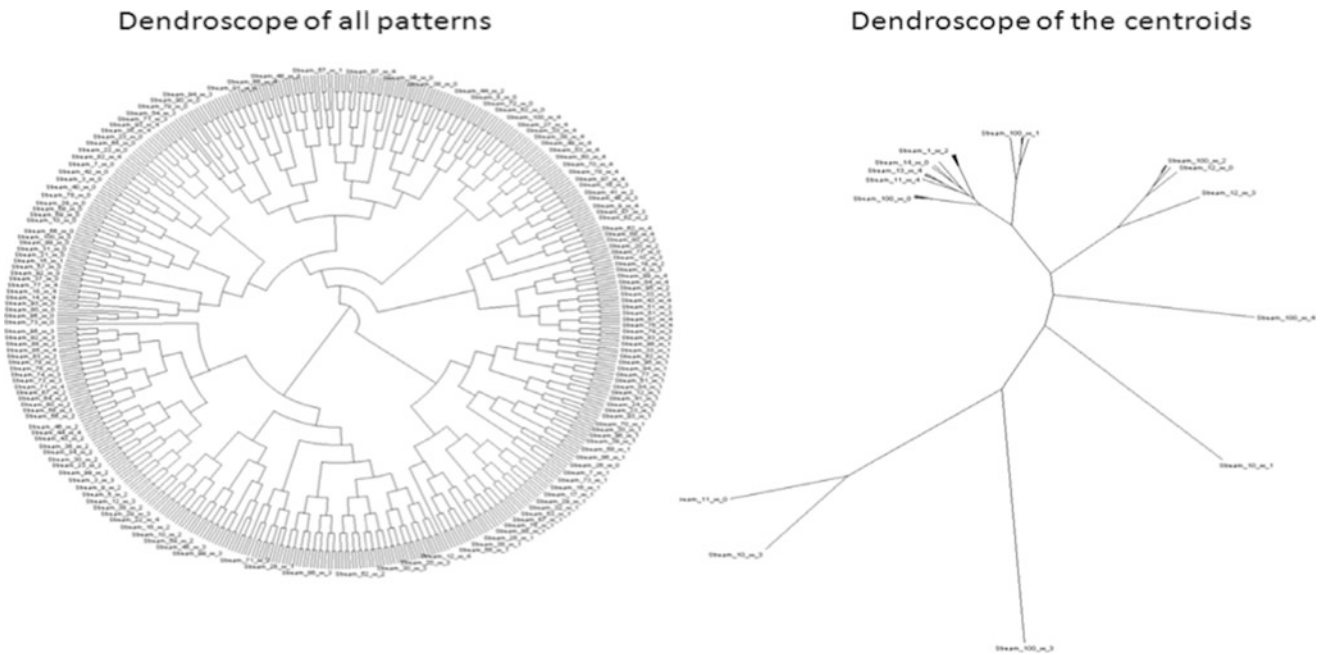


Fig. 1: Dendroscope visualization of the clustering results (Left). Kelley approach would suggest to prune the tree, keep coherent clusters, and determine their medoids. An index of patterns (Right) is generated through the medoids of the obtained clusters.

results, we will use the term *configuration* to mean the resolution of a dataset using an encoding method. Overall, we defined 68 configurations. At the end of the encoding step, we obtain the encoded time series $\widehat{T}$.

4.2 Sub Dimensional Clustering Results:

From 68 configurations, 2 patterns extraction methods, and 5 similarity measures, we generated 680 similarity matrices *SimMat*. Each matrix is used as an input for the clustering algorithm. To determine the optimal number of clusters, we used the *Kelley* penalty method [22]. It can help to decide where to prune a hierarchical cluster tree. At any level of the clustering tree in figure (1, left), the mean of the dissimilarity measure across all clusters, and the mean within each clusters is calculated. Then a ratio is calculated between these two values, and the minimum corresponds to the suggested pruning size (best value of k).

4.3 Indexes Generation Results:

After the pruning step, we obtain the collection C of clusters and their medoids (see Algorithm 1). Each medoid is representative of the patterns of its cluster. Figure ((1, right)) gives an example of an index visualization obtained by retaining medoids of the precedent clustering results.

We can now evaluate the gain in term of computation time regarding the patterns searching using an entire scan (Brute Force) and the index scan (*i-StreamMiner*).

4.4 Querying Evaluation Results:

The performances of *i-StreamMiner* are investigated through several evaluations, executed on the datasets presented above. The target task is to search in a database, and retrieve the time series related to a query formulated as key subsequences. Sets of queries were generated as patterns p_i, for each dataset and resolution. The runtime taken to answer the query as well as the result accuracy are then assessed. Figure 2 illustrates the results obtained using different configurations. Each experience has been performed at least 3 times, and the results are averaged. These experiences have been executed by combining entire scan (*BruteForce*) and *i-StreamMiner* (*Indexed* in the figure), with patterns extracted by segmentation (*Seg*) or Suffix Trees (*Suf*), using MVQ, SAX or 1D SAX, and over 5 similarity measures presented above.

Evaluating Time Processing: Values on Y-axis of figure 2 represent time processing (log) consumed to retrieve a query for a configuration. We can observe in most cases, significant differences between brute force and *i-StreamMiner* time processing. It is worth-noting that in every configurations, *i-StreamMiner* outperforms *BruteForce* (about 100 times) with respect to runtime processing. Similar performances were obtained when varying the datasets, the encoding methods, similarity measures, or patterns extraction approaches. Note that in general cases, the elastic measures (*SimLCSS* and *SimLev*) are time consuming since they search the best alignment between the compared sequences. The results validate our hypothesis that suggests the creation of an index to speed up time series searching tasks.

Evaluating The Accuracy of i-StreamMiner: The accuracy of *i-StreamMiner* represents it's ability to increase true positives and decrease false positives during the querying process. It is mostly related to the evaluation of used similarity measure, since they are used to generate clusters, and compare the queries with the centroids. Note that this evaluation does not consider *BruteForce* since it is meaningless to evaluate the accuracy of an exhaustive scan process. The accuracy here is calculated through the number of visited centroids for a given query. Firstly, we randomly generate a query with its known true centroid (called TC). After that, the index is interrogated with this query and its most similar centroid is extracted. Ideally, the cluster of this retrieved centroid, should contain this query (TC). In this case, the accuracy equals 1. Otherwise other centroids are visited and compared to TC and each new visit decreases the accuracy. This process is repeated with other randomly generated queries and a mean accuracy value is calculated. Figure 3 shows results obtained on the selected datasets. The first observation that we can make concerns the difference between the encoding methods. Clearly, MVQ is more accurate at encoding strings than SAX, since for MVQ the accuracy values range in [0.8, 1], whereas for SAX the minimum is 0.05. How ever, these two encoding approaches generate good results compared to 1D SAX that is used with Kappa similarity. These results confirm precedent results asserting that MVQ gives more efficient clustering than the other available methods [12]. The comparison of the similarity measures revealed that *SimJac* and *SimSemSeq* were competitive to each other, with a slight better accuracy for *SimSemSeq*. Values fluctuated from a dataset to an other. Best results were obtained with elastics measures (*SimLev* and *SimLCSSS*), even though they are time consuming. Worst results were obtained Kappa similarity. Overall, user can make a deal between the accuracy and time processing for information retrieval from time series data base using the proposed method. Elastic measures extract patterns with high accuracy but with important time processing, whereas the semantic similarity extract patterns with acceptable accuracy values and fast time processing. A difference was also observed between the values using Suffix Trees (Suf) and Segmentation (Seg). Pattern extraction based on Suffix Tree induced better accuracy than segmentation. This could be explained by the fact that suffixes as queries with different length were well matched with clustered patterns, hence decreasing false positives.

5 Conclusion

In this paper we have presented *i-StreamMiner*, a novel approach for fast similarity search over time series data. The advantages were illustrated by comparing *i-StreamMiner* to a BruteForce retrieval approach. The results show that using an index on patterns yields a fast similarity search over streaming data (about 100 time faster than Brute Force). We have also discussed the impact of different encoding methods, patterns extraction approaches, and similarity measures, on the retrieval accuracy of *i-StreamMiner*. Concerning the used similarity measures, it appears that elastic measures are more efficient but time consuming. The semantic similarity *SimSemSeq* is

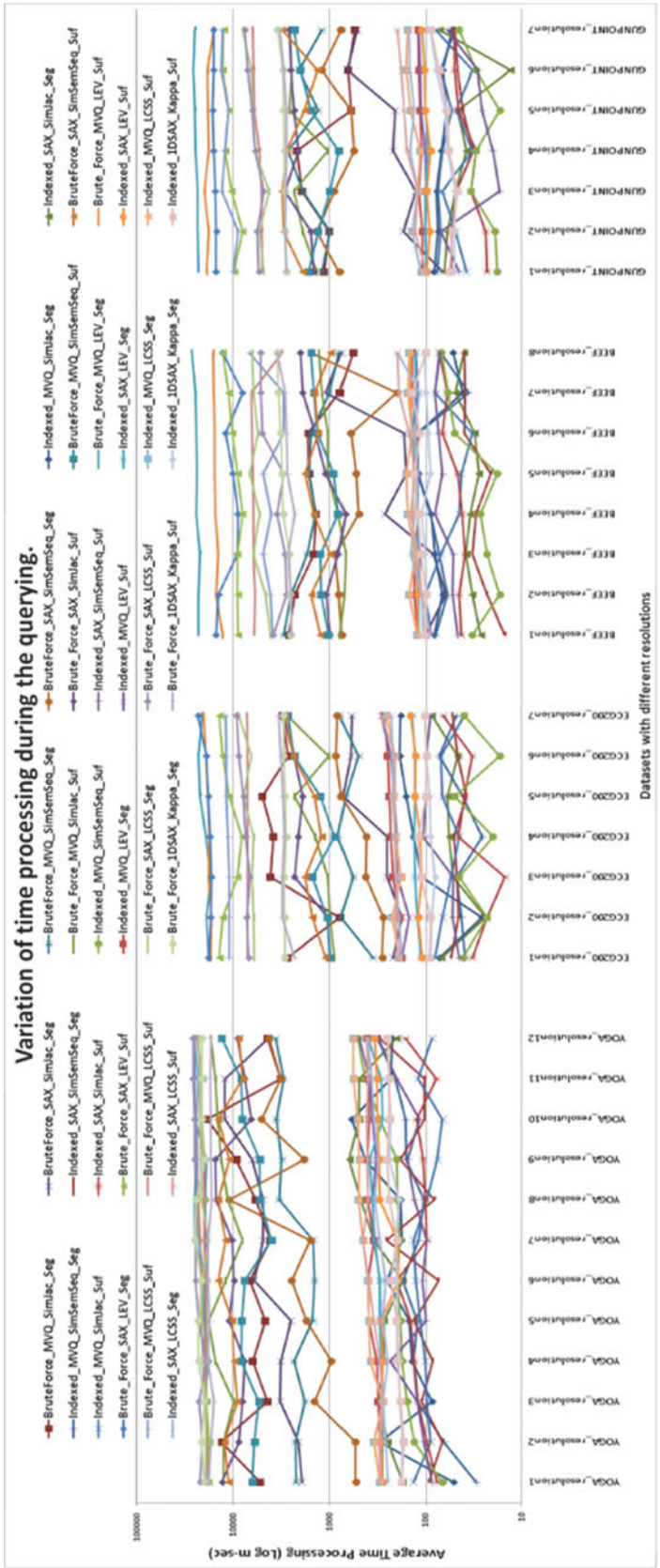


Fig. 2: Evaluation of time processing for BruteForce and i-StreamMiner.

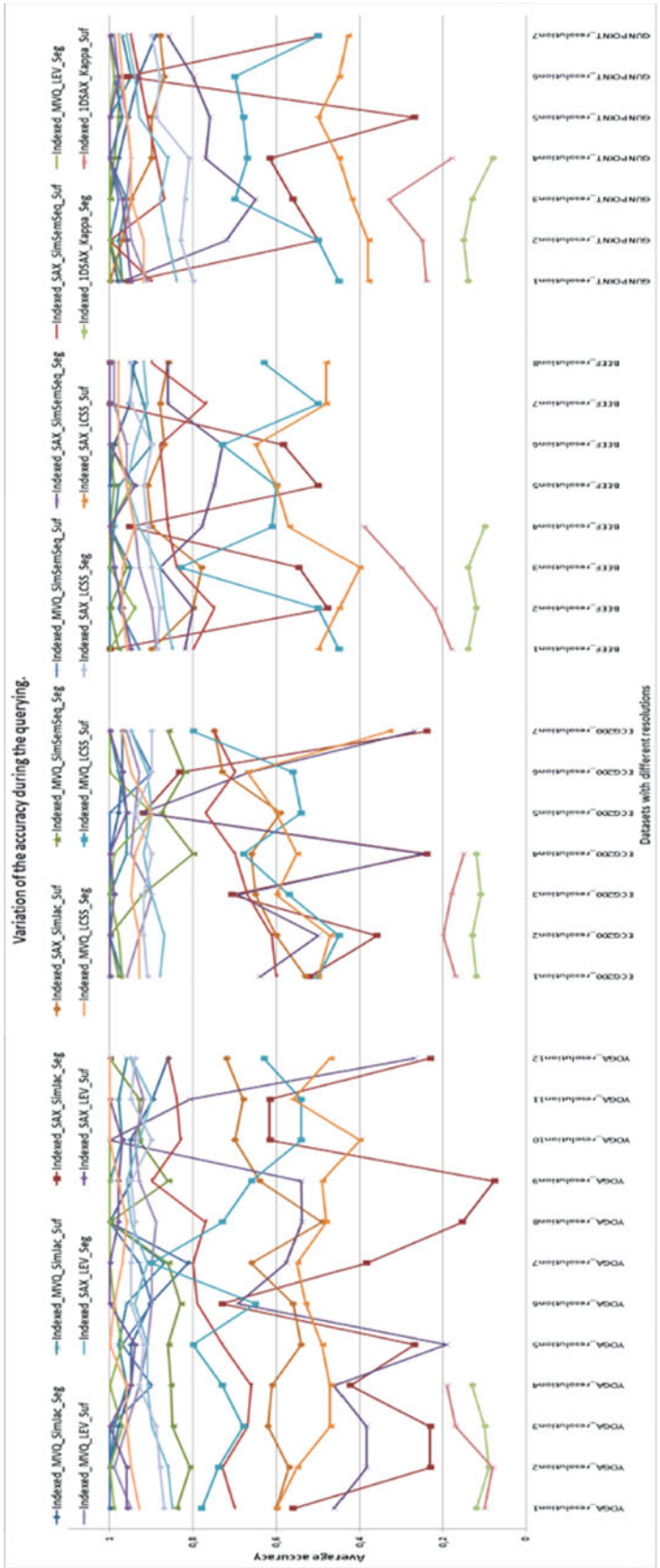


Fig. 3: Evaluation of the accuracy of i-StreamMiner during the querying.

competitive in accuracy and has fast time processing. How ever, regarding the patterns extraction methods, suffix trees provided a better accuracy than segmentation.

In future work, we want to test other rudimentary similarity measures, the test the method on real life applications such as satellite images mining.

References

1. Gaber Mohamed et al. Mining data streams: a review. *SIGMOD Rec.*, 34(2):18–26, 2005.
2. Ralanamahatana et al. Mining time series data. In *DMKD'05*, pages 1069–1103.
3. Lin et al. A symbolic representation of time series, with implications for streaming algorithms. In *ACM SIGMOD RIDMKD Workshop 03*, pages 2–11.
4. Wei et al. Efficient query filtering for streaming times series. ICDM '05, pages 490–497.
5. Capitani et al. Warping the time on data streams. *DKE*, 62 (3):438–458, September 2007.
6. Rafiei et al. Similarity-based queries for time series data. SIGMOD'97, pages 13–25.
7. Chakrabarti et al. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM TDS.*, 27(2):188–228, June 2002.
8. Lin Jessica et al. Rotation-invariant similarity in time series using bag-of-patterns representation. *JIIS*, 39(2):287–315, 2012.
9. Dhiral et al. Distance measures for effective clustering of arima time-series. In *ICDM'01*.
10. Guralnik et al. Event detection from time series data. KDD '99, pages 33–42.
11. Pong et al. Efficient time series matching by wavelets. In *ICDE99*, pages 126–133.
12. Qiang et al. A multiresolution symbolic representation of ts. In *ICDE05*, pages 668–679.
13. Malinowsky Simon et al. 1d-sax; a novel symbolic representation for time series. In *IDA'13*.
14. Wang et al. Time series analysis with multiple resolutions. *Inf. Syst.*, 35(1):56–74, 2010.
15. Aho et al. Efficient string matching: an aid to bibliographic search. *com. ACM*, 18(6), 1975.
16. Benabderrahmane et al. Intelligo: a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics*, 11(1):588, 2010.
17. Salton et al. *Introduction to Modern Information Retrieval*. McGraw-Hill, NY, USA, 1983.
18. Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*.
19. Raymond et al. On the marriage of lp-norms and edit distance. In *VLDB 04*.
20. Vlachos et al. Indexing ts with support for multiple distance measures. KDD '03.
21. Ucr database. http://www.cs.ucr.edu/eamonn/timeseriesdata.
22. Kelley et al. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein eng.*, 9(11):1063–1065, 1996.

Active Learning based on Random Forest and Its Application to Terrain Classification

Yingjie Gu, Dawid Zydek, and Zhong Jin

1 Introduction

In the machine learning literature many supervised algorithms have been proposed to perform pattern classification tasks. But in many pattern recognition tasks, labels are often expensive to obtain while a vast amount of unlabeled data are easily available. And redundant samples are often included in the training set, thus slowing down the training process of the classifier without improving classification results. To solve this problem, active learning [1][2] techniques are proposed to select the most valuable samples for manually labeling to train a classifier.

Uncertainty, density, and diversity are three of the most important criteria in active learning. Uncertain samples are usually able to improve the current classifier most. The most popular uncertainty sampling is SVM_{active} [3] [4] that selects the sample nearest to the current decision boundary. In density sampling, samples in dense regions are thought to be representative and informative. The cluster structure of unlabeled data is usually exploited to find samples in dense regions. The main weakness of uncertainty and density sampling is that they are unable to exploit the abundance of unlabeled data. Thus the diversity criterion was proposed to select a set of unlabeled samples that are as more diverse as possible in the feature space, which reduces the redundancy among the samples selected at each iteration.

Recently, some active learning algorithms tried to combine two criteria to find the optimal samples. In [5], Huang et al. tried to query informative and representative examples based on the min-max view of active learning [6]. Some active learning techniques also query a batch of unlabeled samples at each iteration by considering both uncertainty and diversity criteria [7] [8]. Shi et al. [9] proposed a batch mode active learning method for Networked Data with three criteria (i.e., minimum redundancy, maximum uncertainty, and maximum impact).

The processing platform for active learning should be considered as well. Among many others, the distributed processing systems are gaining many attention and are suitable for active learning system that gathers samples from many distributed locations, and processes them as one virtual entity. Such solution was proposed in [10] where the system that optimizes the processing task allocation in Peer-to-Peer based computing architecture was proposed. In [11], the decentralized approach was shown, also supporting the multiple data sources (suitable to obtain samples).

Large numbers of active learning algorithms are based on SVM and regression classifier. But there is little work about active learning using random forest classifier. According to the information we have, DeBarr et al. have made an exploration in random forest active learning [12]. In this paper, we proposed a novel active learning algorithm based on random forest that selects samples with large uncertainty, density, and diversity for manual labeling. For each unlabeled samples, we use the difference between the most votes and second most votes from the random forest classifier to measure its uncertainty. The average distance between the sample and its k-nearest unlabeled neighbors is used to measure the density while the distance between the sample and its nearest labeled neighbor is used to measure the diversity.

The rest of this paper is organized as follows. Section 2 describes the proposed active learning based on random

Y. Gu (✉)
Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Department of Electrical Engineering, Idaho State University, Pocatello 83209-8060, USA
e-mail: csyjgu@gmail.com

D. Zydek
Department of Electrical Engineering, Idaho State University, Pocatello 83209-8060, USA
e-mail: zydedawi@isu.edu

Z. Jin
Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
e-mail: zhongjin@njust.edu.cn

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_41,

forest. The experimental settings and results on several data sets are presented in Section 3. Finally, Section 4 discusses the conclusion of this work.

2 Active Learning based on Random Forest

2.1 Random Forest and Active Learning

Random forest is an ensemble classifier that is composed of many decision trees. It was first proposed to solve the classification problem [13]. For a new testing sample, each tree gives a prediction. So the testing sample receives a vote on that prediction class. The prediction label of the sample is the class with the most votes. In recent years, there have been a lot of applications [14] [15] [16] in computer vision, which employ random forest as classifier. Although it has been widely applied, there is little work apply random forest in active learning. According to what we know, DeBarr et al. have made an exploration in random forest based active learning [12]. They queried the sample whose probability assigned by the random forest model is closet to 0. 5. The probability of an instance is computed as the proportion of decision trees assigning the instance label. Similar to Tong's SVM active learning, it just selects the most uncertain sample for manual labeling.

A general active learning procedure is as follows:

step 1. Randomly select several samples to construct an initial training set $\mathcal{L}$ to train a classifier.

step 2. According some criteria, select a set of samples from unlabeled pool $\mathcal{U}$ for manual labeling.

step 3. Selected samples are added to $\mathcal{L}$ and the classifier is retrained by updated training set.

step 4. Repeat 2 and 3 until a stop criterion is satisfied.

The key problem in active learning is how to select a set of samples or a sample from unlabeled pool $\mathcal{U}$ in Step 2. In this paper, a novel active learning algorithm is proposed to select samples with maximum uncertainty, density, and diversity to improve the classifier.

2.2 The Proposed Approach

Given a dataset by $\mathcal{D} = \{x_1, x_2, ..., x_{n+m}\}$, where $\mathbf{x}_i$ is a sample of d dimension vector. The labeled data is $\mathcal{L} = \{x_1, x_2, ..., x_n\}$ while the unlabeled data is $\mathcal{U} = \{x_{n+1}, x_{n+2}, ..., x_{n+m}\}$, so $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$. The label of sample $\mathbf{x}_i$ is $y_i \in \{1, 2, \ldots, c\}$, $i = 1, \ldots, n$.

In the following, we will introduce how to select samples with maximum uncertainty, density, and diversity.

Uncertainty step A model can be trained with random forest based on labeled data $\mathcal{L}$. Perform classification on unlabeled data $\mathcal{U}$, the vote of each sample assigned to each class can be obtained. We use $\mathcal{V} = \{v_{ij} \mid i = n+1, ..., n+m, j = 1, ..., c\}$ to denote the votes of all unlabeled samples assigned to each class. v_{ij} denotes the vote of the unlabeled data $x_i \in \mathcal{U}$ assigned to class $j \in \{1, 2, \ldots, c\}$.

In prediction, random forest assigns each sample to the class that gets the maximum vote. The maximum vote of sample $\mathbf{x}_i$ is defined as $\overline{v}_i$, so

$$\overline{v}_i = \max_{j=1,...,c} \{v_{ij}\} \tag{1}$$

In traditional active learning based on random forest, among all unlabeled samples, the one with the minimum $\overline{v}_i$ [12] is selected for manual labeling. In their opinion, the smaller $\overline{v}_i$ is, the more uncertain the classification result is.

Here we propose a new method to measure the uncertainty of samples. As Figure 1 shows, the sample in left figure is denoted as $\mathbf{s}_1$ while the sample in right figure is denoted as $\mathbf{s}_2$. It can be seen that $\mathbf{s}_1$ got the maximum vote less than 300 while $\mathbf{s}_2$ got the maximum vote more than 300. Traditional random forest active learning will select $\mathbf{s}_1$ since its maximum vote is smaller than that of $\mathbf{s}_2$. However, the maximum vote of $\mathbf{s}_1$ is much larger than the votes of any other class. On the contrary, the maximum vote of $\mathbf{s}_2$ is just slightly larger than the vote of class 4. Thus we suppose $\mathbf{s}_2$ is more uncertain than $\mathbf{s}_1$ since the label of $\mathbf{s}_2$ is more ambiguous.

In view of the above reason, the difference between the maximum vote and the second maximum vote can be used to measure samples' uncertainty. Smaller difference means more uncertainty of a sample.

If sample $\mathbf{x}_i$ get the maximum vote on class p and second maximum vote on class q, namely

$$p = \arg\max_{j=1,...,c} \{v_{ij}\}$$

$$q = \arg\max_{j=1,...,c, j \neq p} \{v_{ij}\}$$

The difference between the maximum vote and the second maximum vote is

$$unc_i = v_{ip} - v_{iq} \tag{2}$$

unc_i is able to measure uncertainty of sample $\mathbf{x}_i$.

Density step Many active learning algorithms select samples that are most representative to unlabeled data. These approaches aim to exploit the cluster structure of unlabeled data, usually by a clustering method. Instead, we propose a novel idea to select representative samples. As we know, samples in dense regions are usually thought to be representative. In other words, a representative point is usually near to

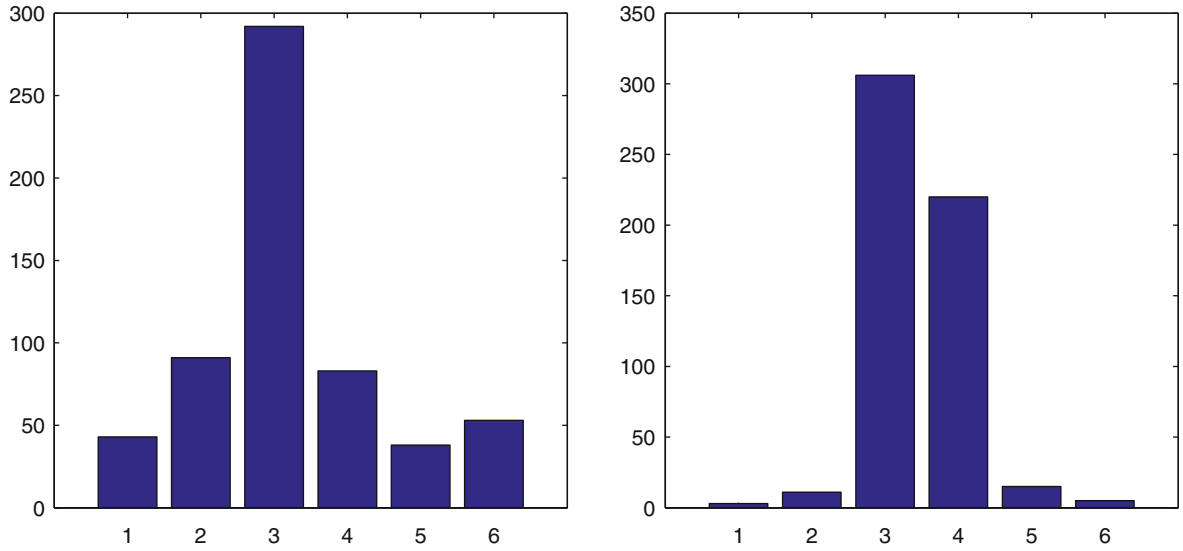


Fig. 1. Votes of two samples: X-axis indicate the class label while Y-axis indicate the vote of each class.

its neighbors. On the contrary, a point far from its neighbors is usually not representative or an outlier. Thus we use the average distance from a sample to its k-nearest neighbors to measure the representativeness of the sample.

For any $x_i \in \mathcal{U}$, define its k nearest neighbors from $\mathcal{U}$ as $x_{i_j}, j = 1, ..., k, x_{i_j} \in \mathcal{U}$. The average distance from $\mathbf{x}_i$ and its k nearest neighbors can be computed:

$$den_i = \frac{1}{k}\sum_{j=1}^{k}\left\|x_i - x_{i_j}\right\|^2 \tag{3}$$

We use den_i to measure the density of sample $\mathbf{x}_i$.

Diversity step Some active learning algorithms select samples that are similar with labeled samples. So it will not improve the classifier obviously. In our proposed approach, the distance between the sample and its nearest labeled neighbor is used to measure the similarity between the sample and labeled samples. If the sample is far from its nearest labeled neighbor, it is dissimilar with other labeled samples. On the contrary, if it is close to the nearest labeled sample, there is at least one sample that is similar with it in the labeled set. Therefore, we select the sample that is far from its nearest labeled neighbor for manual labeling.

For any $x_i \in \mathcal{U}$, compute the distance between $\mathbf{x}_i$ and its nearest labeled neighbor:

$$div_i = \min_{j=1,2,...,n}\left\|x_i - x_j\right\|^2 \tag{4}$$

$$x_i \in \mathcal{U}, x_j \in \mathcal{L}$$

Large div_i indicates that sample $\mathbf{x}_i$ has little similarity with labeled samples.

Selection Function In this section, we want to select an uncertain, representative, and diverse sample that is measured by unc_i, den_i, and div_i. To combine these three criteria together, unc_i, den_i, and div_i $(i = n+1, ..., n+m)$ are normalized to [0, 1], respectively. For any vector $\mathbf{p} = [p_1, \ldots, p_n]$, the normalization operation is as follow:

$$p_i^{'} = \frac{p_i - min}{max - min} \tag{5}$$

where $max = \max_{j=1,...,n}\{p_j\}$ and $min = \max_{j=1,...,n}\{p_j\}$.

The query criteria of proposed active learning can be described as:

$$s = \underset{i=n+1,...,n+m}{argmin}\ \{unc_i + den_i - div_i\} \tag{6}$$

As above analysis, proposed algorithm would select the sample $\mathbf{x}_s$ that has large uncertainty, density, and diversity.

The proposed active learning approach is summarized in Table 1.

3 Experiments

To demonstrate the effectiveness of our proposed algorithm, we compare it with other three active learning methods:

- **Random Sampling** method, which randomly select samples from the unlabeled data.
- **Random Forest active learning** (RFAL), proposed by DeBarr[12] which select samples with min-max votes.
- **Support Vector Machine active learning** (SVMAL) [4], which selects the point closest to the current decision boundary.

Table 1 The proposed active learning approach

Input:
$\mathcal{D}_{train}$: A data set for training
$\mathcal{D}_{test}$: A data set for testing
Initialize:
$\mathcal{L}$: random select 10 samples from $\mathcal{D}_{train}$
$\mathcal{U}$: $\mathcal{U} = \mathcal{D}_{train} - \mathcal{L}$
Repeat:
for $i = n + 1$ to $n + m$ do
Calculate unc_i, den_i, div_i
Calculate $score_i = unc_i + den_i - div_i$
end for
$s = \underset{i=n+1,\ldots,n+m}{argmin} \; score_i$
$\mathcal{L} \leftarrow \mathcal{L} \cup \mathbf{x}_s$
$\mathcal{U} \leftarrow \mathcal{U} - \mathbf{x}_s$
Until the number of selected or the required accuracy is reached

Fig. 2. Patch examples of Outex: bush, grass, tree, sky, road, and building

– **Proposed algorithm**, which query samples with maximum uncertainty, density, and diversity based on random forest. Random forest tool is available here[17].

The analysis of outdoor terrain images for navigating a mobile robot is very challenging. In experiments, above four active learning algorithms are performed on two terrain image data sets.

3.1 Outex Data Sets

Outex data [18] contains two data set: Outex0 and Outex1. Both of them include 20 outdoor scene images and the images' size is 2272 × 1704. The labeled area of each image is cut into patches of size 64 × 64. The patches contain 6 terrain classes defined as bush, grass, tree, sky, road, and building with considerable changes of illumination. Two sample patches of each class are shown in Figure 2. Each terrain patch is represented by a 64 × 64 dimensional vector in image space. It is difficult to classify these terrains directly in image space. We extract color histogram feature [19] and texture feature using rotation-invariant operators $LBP^{riu2}_{8,1+16,3}$ [20] [21]. Both of these features were proved to be effective in performing outdoor scene classification tasks.

For each class, 50 patches are randomly selected to construct a training set while 50 patches are randomly selected for testing. Then 10 patches in training set are randomly selected to construct an initial labeled set $\mathcal{L}$ and the rest in training set construct unlabeled set $\mathcal{U}$. At each iteration, we

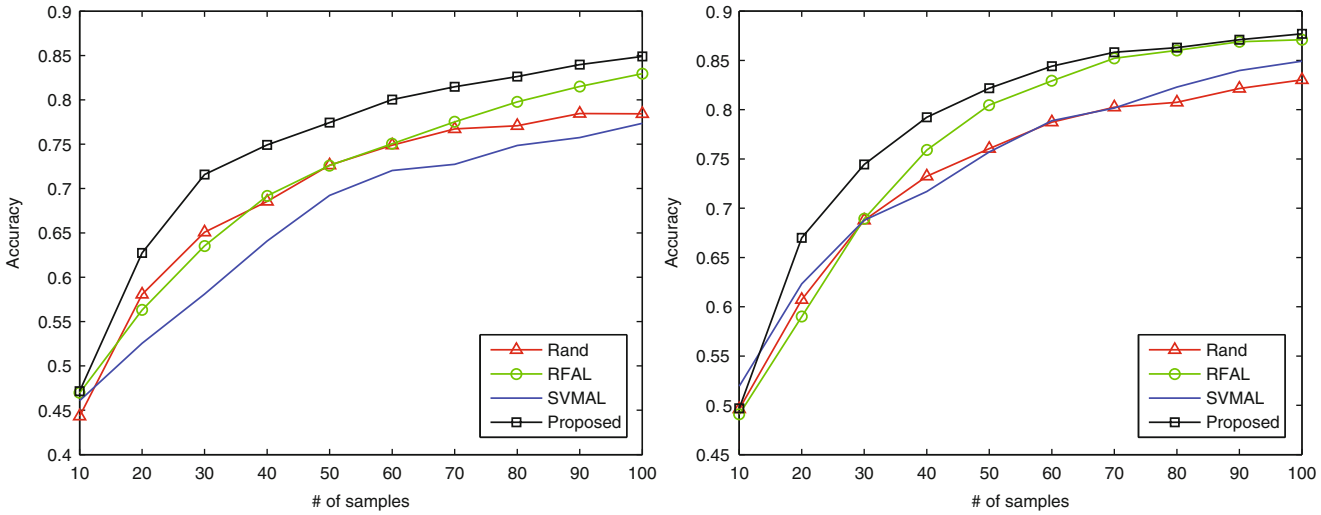


Fig. 3. Classification accuracy on Outex-0, Outex-1

query 10 samples from $\mathcal{U}$ for manual labeling and add them to $\mathcal{L}$. The max number of iteration is fixed at 10. The experiment is repeated for 20 times and the average classification accuracy is shown in Figure 3.

From Figure 3, we can see that our proposed algorithm outperforms other methods. SVM active learning performs the worst nearly all the case since it is not developed for multi-class active learning. Traditional random forest active learning is worse then proposed method because it just selects uncertain sample while ignore the sample's density and diversity.

3.2 Hand-Labeled DARPA LAGR Datasets

Hand-Labeled DARPA LAGR Data sets [19] contain 3 scenes with different lighting condition. Each data set consists of 100-frame, hand-labeled image sequence. Each image was manually labeled, with each pixel being placed into one of three classes: OBSTACLE, GROUNDPLANE, or UNKNOWN. Feature extraction method is fixed as color histogram [22]. To create a color histogram, color intensities in each of the three color channels(R, G, and B) in the neighborhood of the reference pixel are binned. The number of bins is fixed at 5 and the window size is fixed at 16×16. Using three color channels and 5 bins per channel results in a feature image with feature depth of 15 values(3 channels $\times$ 5 bins per channel).

Active learning methods are performed on one of the data sets, DS2A. 5 points of each class in each frame are randomly selected to construct a training set. So the training set consists of 1000 samples. In the same way, a testing set can

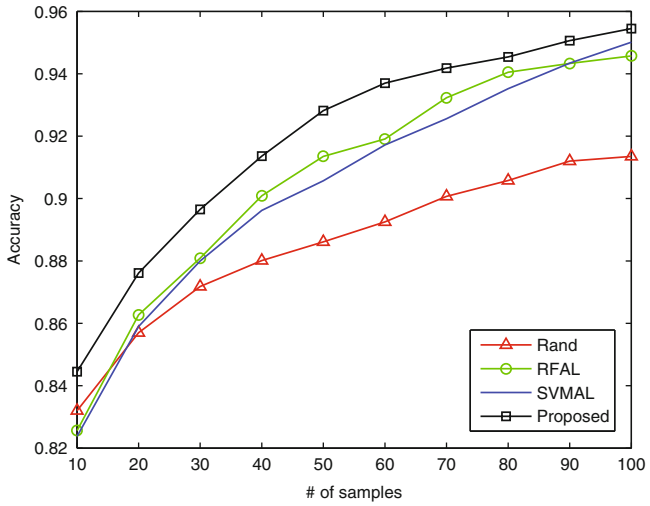


Fig. 4. Classification accuracy on DS2A

be constructed. Firstly, 10 points from training set were randomly selected for manual labeling to construct an initial labeled set $\mathcal{L}$ and the rest in the training set construct an unlabeled set $\mathcal{U}$. At each iteration, we query 5 samples from $\mathcal{U}$ for manual labeling and add them to labeled set $\mathcal{L}$. The max number of iteration is fixed at 10. The experiment is repeated for 20 times and the average classification accuracy is shown in Figure 4.

As can be seen from Figure 4, our proposed active learning algorithm performs the best. SVM active learning and random forest active learning perform worse since they just query the most uncertain sample but ignore the samples' density and diversity. Random sampling performs worst because it selects samples without any criteria.

4 Conclusion

In this paper, we propose a novel active learning technique for solving multiclass classification problem with random forest classifier. The proposed technique combines samples' uncertainty, density, and diversity information and selects the most valuable one. The results of experiment indicate the proposed method outperforms other methods.

There are several advantages of the proposed algorithm:

- The proposed active learning algorithm can also initialize the labeled set $\mathcal{L}$. The selection function is:

$$s = \underset{i=n+1,\ldots,n+m}{argmin} \{unc_i + den_i - div_i\} \tag{7}$$

  if we set $unc_i = 0$ and $div_i = 0$, we can decide which sample should be firstly labeled. Then the labeled set can be constructed according to selection function.
- It can be expanded to other classifiers through altering the first criterion that measures uncertainty of samples.
- There are no other parameters except two parameters in random forest classifier.
- It is independent with samples' label so it can be used for multi-class active learning.

Moreover, there are several interesting directions for extending present work. In this paper, we choose euclidean distance to measure the similarity of two samples, how about mahalanobis distance or consine distance? And how to measure the uncertainty, density, and diversity of samples more effectively. Last but not the least, how to combine different criteria together is an extremely difficult and significant problem.

Acknowledgment This work is partially supported by National Natural Science Foundation of China under Grant Nos. 61373063, 61233011, 61125305, 61375007, 61220301, and by National Basic Research Program of China under Grant No. 2014CB349303.

References

1. D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.
2. B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, 2010.
3. S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proceedings of the ninth ACM international conference on Multimedia*. ACM, 2001, pp. 107–118.
4. S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.
5. S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples." in *NIPS*, vol. 23, 2010, pp. 892–900.
6. S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Semi-supervised svm batch mode active learning for image retrieval," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–7.
7. D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 47, no. 7, pp. 2218–2232, 2009.
8. S. Patra and L. Bruzzone, "A cluster-assumption based batch mode active learning technique," *Pattern Recognition Letters*, vol. 33, no. 9, pp. 1042–1048, 2012.
9. L. Shi, Y. Zhao, and J. Tang, "Batch mode active learning for networked data," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 2, p. 33, 2012.
10. G. Chmaj, K. Walkowiak, M. Tarnawski, and M. Kucharzak, "Heuristic algorithms for optimization of task allocation and result distribution in peer-to-peer computing systems," *International Journal of Applied Mathematics and Computer Science*, vol. 22, no. 3, pp. 733–748, 2012.
11. G. Chmaj and S. Latifi, "Decentralization of a multi data source distributed processing system using a distributed hash table." *International Journal of Communications, Network & System Sciences*, vol. 6, no. 10, 2013.
12. D. DeBarr and H. Wechsler, "Spam detection using clustering, random forests, and active learning," in *Sixth Conference on Email and Anti-Spam. Mountain View, California*, 2009.
13. L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
14. J. Gall and V. Lempitsky, "Class-specific hough forests for object detection," in *Decision Forests for Computer Vision and Medical Image Analysis*. Springer, 2013, pp. 143–157.
15. A. Yao, J. Gall, and L. Van Gool, "A hough transform-based voting framework for action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2061–2068.
16. C. Marsala and M. Detyniecki, "High scale video mining with forests of fuzzy decision trees," in *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*. ACM, 2008, pp. 413–418.
17. Random forest packages. [Online]. Available: http://cran.r-project.org/web/packages/
18. University of oulu texture database. [Online]. Available: http://www.outex.oulu.fi/temp/
19. Hand-labeled darpa lagr datasets. [Online]. Available: http://www.mikeprocopio.com/labeledlagrdata.html
20. M. Pietikäinen, T. Nurmela, T. Mäenpää, and M. Turtinen, "View-based recognition of real-world textures," *Pattern Recognition*, vol. 37, no. 2, pp. 313–323, 2004.
21. T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.
22. M. J. Procopio, J. Mulligan, and G. Grudic, "Learning terrain segmentation with classifier ensembles for autonomous robot navigation in unstructured environments," *Journal of Field Robotics*, vol. 26, no. 2, pp. 145–175, 2009.

Classification of Multichannel EEG Signal by Linear Discriminant Analysis

Mohammad Rubaiyat Hasan, Muhammad Ibn Ibrahimy, S. M. A. Motakabber, and Shahjahan Shahid

1 Introduction

By using the classification algorithm for EEG signal it becomes easy to find out performance of Brain computer interface(BCI). BCI causes direct operation between brain and computer. Studies showed that a person with severe neuromuscular disabilities can learn to use a BCI system by modulating the various features in EEG signal [1]. The efficiency of a BCI depends on 3 operations. They are: signal recording; feature extraction from the recorded signal and classification of the extracted information [2].

The output of the feature extraction unit highly impacts on the performance of the feature classification unit. The probability of correctness identification can be increased if the feature extraction unit transforms the EEG signal in such a way that the signal to noise ratio (SNR) can be maximized as much as possible [3]. This paper presents Linear Discriminate Analysis (LDA), a signal classification algorithm for a MI BCI. For classify EEG signal, it has been used the signal recorded from the motor cortex area while a subject performs the imagination of a motor movement.

The rest of the paper is organized as follows. In Section 2, the materials and methods applied here are mentioned. Performance measurements are discussed in Section 3 and results have been showed in section 4. Finally, section 5 concluded the paper.

M.R. Hasan (✉) • M.I. Ibrahimy • S.M.A. Motakabber
Dept. of Electrical and Computer Engineering, International Islamic University, Gombak, Malaysia
e-mail: mdrubaiyat@yahoo.com; Ibrahimy@iium.edu.my; amotakabber@iium.edu.my

S. Shahid
Dept. of Computing Science, University of Glasgow, Scotland, UK
e-mail: shahjahan.shahid@gmail.com

2 Materials and Methods

In first part the feature extraction has been done by Power spectral density (PSD) for 22 different channel data. Two different frequency bands are calculated for those 22 EEG channels. Then the LDA classifier is used to get more accurate results. The proposed technique has been used to devise to an MI related BCI which has been evaluated with the data provided by the Graz BCI lab as part of the BCI competition IV data-2a. To validate accuracy measurement of accuracy and Cohen's kappa are used in this paper. The accuracy results of the classification are shown in Sections 4.

2.1 Data Selection

In research work, the data set consists of EEG data from 9 subjects. The BCI paradigm consisted of 2 different motor imagery tasks. They are: the imagination of movement of the left hand (class 1) and right hand (class 2). For each subject the two sessions were recorded on different times. Each session is comprised of 6 runs separated by short breaks. One run consists of 24 trials (12 for each of the two possible classes), yielding a total of 144 trials per session. For each session, at the beginning a recording of approximately 5 minutes was performed to estimate the EOG influence. The recording was divided into 3 blocks: two minutes with eyes open (looking at a fixation cross on the screen), one minute with eyes closed and one minute with eye movements.

Fig. 1 shows the timing scheme of paradigm. In this paradigm 0-6 seconds time for one session. Around 2 seconds break makes a total of 8 seconds time for each session. First 3 seconds for for fixation and maintaining the cue then 3-6 seconds the potential time for recording the MI based EEG signal.

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_42,

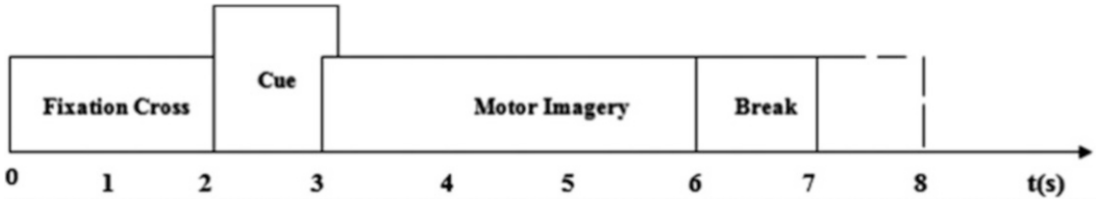


Fig. 1 Timing scheme of the paradigm [4].

2.2 Feature Extraction

PSD shows the strength of the variations as a function of frequency. In other words, it shows at which frequencies variations are strong and at which frequencies variations are weak. The unit of PSD is energy per frequency (width) and you can obtain energy within a specific frequency range by integrating PSD within that frequency range. Computation of PSD is done directly by the method called FFT or computing autocorrelation function and then transforming it.

2.3 Classification

Classification is done by LDA. The aim of is to use hyperplanes to separate the data representing the different classes. LDA assumes normal distribution of the data, with equal covariance matrix for both classes. The separating hyperplane is obtained by seeking the projection that maximize the distance between the two classes means and minimize the interclass variance. To solve an N-class problem ($N > 2$) several hyperplanes are used in LDA. It can be defined by the equation, which is maximized over all linear projections, w:

$$J(w) = |\,m_1 - m_2|^2/(S_1^2 + S_2^2) \tag{1}$$

Here, m represents the mean, S represents a variance, and the subscripts denote the two classes [5]. Limitation of LDA is that, for nonlinear classification it does not provide good performance always. But the high-dimensional and noisy nature of EEG often limits the advantage of nonlinear classification methods [6].

3 Performance Measures

Performance of a BCI system is measured by percentage of accuracy[7].The performance have been measured to see evaluate the classification results of LDA for those data. This is done by percentage of left and right accuracy measurement following the detection of accuracy and Cohen's kappa

3.1 Accuracy

The accuracies have been computed for each instant of data by the classifier's output (estimated label = the sign of d(m), where right means positive and left means negative) is compared with actual left or right of MI to prepare a confusion matrix (CM). Using the CM, the left and right accuracies for each instant of the paradigm are computed by the following formulae:

$$\text{Left accuracy} = \frac{(\text{finally obtained negative in CM} \times 100)}{(\text{total number of input as left})} \tag{2}$$

$$\text{Rightaccuracy} = \frac{(\text{finally obtained positive in CM} \times 100)}{(\text{total number of input as right})} \tag{3}$$

Within the paradigm a total of 50 equi-spaced points were considered from the 3-8 (5 seconds). Considering all trials and their actual and estimated labels (left and right), 50 CM have been made; where positive means both actual and estimated labels are right and they match; negative means both actual and estimated label are left and they match. The left and right accuracies are then computed using equation (1) for each time point of the paradigm. The mean of left and right hand MI accuracy was called here as the overall accuracy.

3.2 Cohen's Kappa

Cohen's kappa is a statistical measurement. It provides an index of interrater reliability. It is an improvement over using the percent of accuracy, as the procedure of computing accuracy does not involve the false positive or false negative effects. The computation of kappa at each instance starts from the CM prepared by comparing the appearance of two raters: the actual events and the estimated events (observed at classifier's output). From the definition, Cohen's kappa can be written as,

$$k = \frac{(P_0 - \mathrm{P_c})}{(1 - P_c)} \quad (4)$$

where Po means the relative observed agreement between raters, and Pc for the hypothetical probability of chance agreement. The maximum possible value of Cohen's kappa is limited to 1 and then the raters are in complete agreement. If there is no agreement among the raters, k = 0 [8].

4 Results and Discussion

The study presented the aim of better performance after using LDA. Table 1 shows the performances of PSD based BCI after applying LDA. These results are obtained upon applying the training and evaluation signals (from 9 subjects) to the respective BCI. Here, column in the table provides the maximum value of a performance by measuring (accuracy or kappa). The maximum measurement has been picked out from its average distribution obtained after averaging across all trials of a session of EEG signal. From results of training stages in Table 1, the average of max accuracy from 9 subjects is above 85%; where, there is around 80% of chances to separate left and right motor imagery signal if we use the ERD/ERS phenomenon in the motor cortex EEG signal. It is only a predictive value of accuracy as the classifier was applied to the same signal on which it was trained. The training stage's results for each subject indicate that the left and right imagination. Hence, the LDA is an acceptable discrimination technique to identify the PSD based features into left and right motor imagery.

Table 1 represents over all accuracy and max kappa for training and evaluation data of 9 different subjects. Comparing with BCI – IVit is found better accuracy applying LDA here. In BCI – IV for 2a data the maximum average kappa was 57%. Here, we have got 61% maximum average kappa using training and valuation data. Individually, for training data we have obtained 73% average kappa and 49% for evaluation data.

Fig. 2 depicts the percentage of overall accuracies for training and evaluation phases. Most of the cases training accuracies are higher than the evaluation accuraceis. Average training accuracy 78% where average evaluation accuracy around 72 in percentage.

Table 1 Results by classification of EEG by LDA based technique

	Training stage		Evaluation stage		Both training and evaluation stages
Subject	Overall accuracy Max (in %)	Max. kappa	Overall accuracy	Max. kappa	Average max kappa
A01	56	0.76	56	0.17	0.47
A02	50	0.56	52	0.40	0.48
A03	99	0.97	93	0.86	0.92
A04	82	0.97	61	0.22	0.60
A05	81	0.64	78	0.57	0.61
A06	69	0.61	69	0.38	0.50
A07	88	0.38	84	0.68	0.53
A08	96	0.75	84	0.68	0.72
A09	81	0.92	72	0.43	0.68
Average	78	0.73	72	0.49	0.61

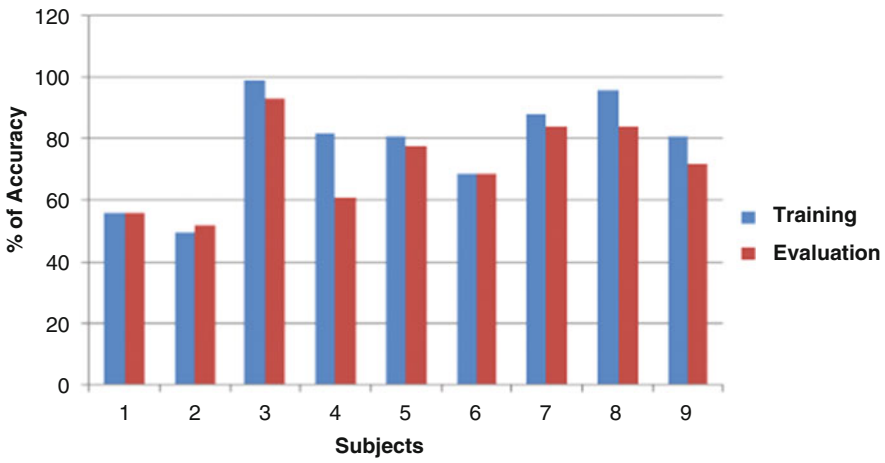


Fig. 2 Comparison of accuracy between training and evaluation data. Blue bar represents training accuracy and red bar represents evaluation accuracy.

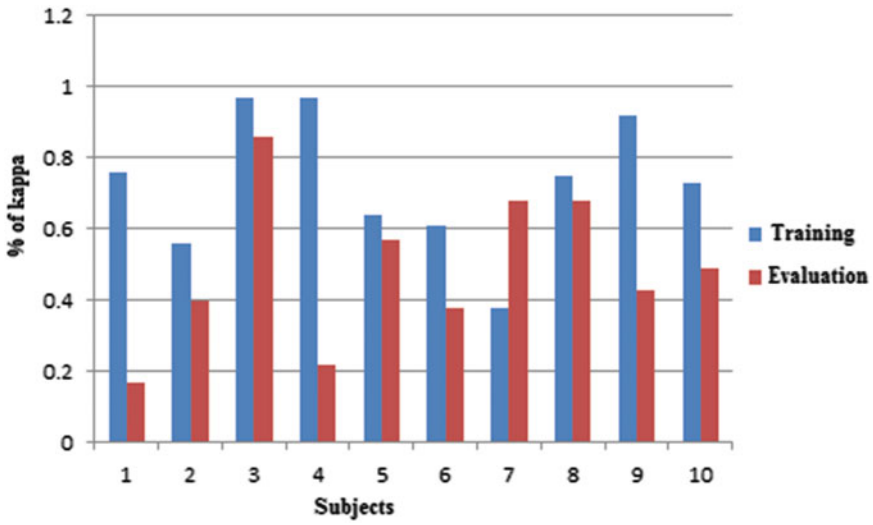


Fig. 3 Comparison of kappa between training and evaluation data. Blue bar represents training kappa and red bar represents evaluation kappa.

Fig. 3 represents the percentage of kappa for training and evaluation phases. Except for subject 7 all other cases the training kappa are higher than the evaluation kappa. Average training kappa73% where average evaluation kappa around 49 in percentage. Overall average kappa for traininig and evaluation is around 61%.

5 Conclusion

Proposed technique shows remarkably much higher and consistent MI task detection accuracy and Cohen's kappa in most of the cases. This paper shows LDA classifier with PSD based feature extraction technique with its application to an MI based BCI. In BCI competition IV average maximim kappa accuracy was 57%, we have obtained 61% by using LDA in this paper. The advantages of using this classification technique are: it uses a simple computation from a sliding windowed EEG signal; it provides performance measures are observed in training and evaluation sessions, and the accuracy or kappa distribution over the time course of paradigm is very similar; In nutshell it is observed that the LDA classification exposes a propitious technique for detecting different brain states.

Acknowledgment This research has been supported by the Ministry of Higher Education of Malaysia through the Exploratory Research Grant Scheme ERGS12-026-0026.

References

1. Ahsan, M.R., Ibrahimy, M.I., Khalifa, O.O.: EMG signal classification for human computer interaction: A review. Eur. J. Sci. Res., 3, 480–501 (2009)
2. Shahid, S., Prasad, G.: Bispectrum based feature extraction technique for devising a practical Brain-Computer Interface. J. Neural Eng., 8(2), Article Number: 025014 (2011)
3. Wolpaw, J.R., Birbaumer, N., McFarland, D.J, Pfurtscheller, G., Vaughan, T.M.: Brain-computer interfaces for communication and control. Clin Neurophysiol, 113(6):767–791, (2002)
4. Brunner, C., Leeb, R., Mü,ller-Putz, G., Schló,gl, A., Pfurtscheller, G.: BCI Competition 2008, Graz Data Set a, Laboratory of Brain-Computer Interfaces, Inst. for Knowledge Discovery, Graz Univ. of Technology (2008)
5. Rathinave, S., Arumugam, S.: Full Shoe Print Recognition based on Pass Band DCT and Partial Shoe Print Identification using Overlapped Block Method for Degraded Images. International Journal of Computer Applications (0975 – 8887), vol. 26 – No.8 (2011)
6. Garrett, D., Peterson, D.A., Anderson, C.W., Thaut, M.H.: Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. IEEE Transactionson Neural System and Rehabilitation Engineering, vol. 11, pp. 141-144 (2003)
7. Hasan, M.R., Ibrahimy, M.I., Motakabber, S.M.A.: Performance Analysis of Different Techniques for Brain Computer Interfacing. International Conference on Computing, Electrical and Electronic Engineering (ICCEEE), pp. 730 – 734 (2013)
8. Sim, J., Wright, C.C.: The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Physical Therapy, vol. 85(3), pp. 257–268 (2005)

Virtual Enterprise Process Monitoring: An Approach towards Predictive Industrial Maintenance

Filipe Ferreira, Ahm Shamsuzzoha, Americo Azevedo, and Petri Helo

1 Introduction

Global business dynamics and volatile market environment pushes companies towards collaboration that can ensure business competitiveness with enhanced sustainability. Manufacturing companies, especially small and medium enterprises (SMEs), where there are often resources scarcity are benefited through such business collaboration. There are various forms of business collaboration exists within the manufacturing domain such as business community, industrial cluster, collaborative network organization (CNO) (Camarinha-Matos et al., 2008), virtual organization (VO) (Carneiro et al., 2010), virtual enterprise (VE) Molina et al., 2007), etc. In such business environment, there needs to continuous monitoring and management of the business processes and associated equipment. The equipment as integrated with the collaborative SMEs processes are need to be real time monitoring in order to enhance its predictive maintenance.

With only a few SMEs having capacity to implement innovative manufacturing technologies, the system must be highly adaptable to any equipment with low cost/investment and without the need of doing hard integrations and setups. Using the approach described in this paper, industrial managers have access to the list of machines installed by their equipment suppliers, as well as their status, location, manuals, technical assistance plans, procedures and drawings. The vendor is then notified in real-time and knows exactly which is the equipment that needs service, having access to the machine status and software, all the machine history, manuals, drawings and service records instantaneously. It's a huge advantage in terms of time-to-market.

Moreover, if the equipment is connected to Internet and the problem is about software, vendor can try to solve it remotely and have real-time feedback from the equipment. If the problem can be solved by doing a couple of procedures, depending on each machine, it can be done remotely by client's technicians and monitored by the vendor too. For customers far from vendor's plant, it's common to subcontract 3rd party firms (partners) to assist their clients. One of the advantages of this new approach is that the vendor becomes able to remotely monitor the process by using equipment's control software integration and the mobile applications where the partner updates the service status. The essential issue tackled in this paper and underlying research study refers to the predicative maintenance and the respective maintenance operations management.

2 Literature review

2.1 Predictive Maintenance

In industrial domain two major maintenance management approaches are available namely; failure-driven and time-based maintenance (Moubray, 1997). There are also other maintenance systems such as conditioned-based maintenance (CBM), statistical-based maintenance (SBM), etc., are used to reduce the uncertainty of maintenance according to the needs indicated by any industrial equipment condition (Yam et al., 2001).

Predictive Maintenance (PM) is used as a maintenance methodology to monitor and detect incipient problems and to prevent catastrophic failure. PM solution opens up innovative new possibilities for companies. It does not depend on industry statistics but relies on real signals demonstrated by a single and specific piece of equipment. Any data or signal from specific sensors monitoring machine condition is

F. Ferreira (✉) • A. Azevedo (✉)
Manufacturing Systems Engineering Unit, INESC TEC (formerly INESC Porto), Porto, Portugal
e-mail: filipe.ferreira@inescport.pt; ala@fe.up.pt

A. Shamsuzzoha (✉) • P. Helo (✉)
Department of Production, University of Vaasa, 700 Wolffintie 34, 65200 Vaasa, Finland
e-mail: ahsh@uva.fi; petri.helo@uva.fi

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_43,

automatically reviewed to pick up any patterns that indicate a possible fault. In addition to early fault detection PM also can be used to avoid unplanned downtimes and both staff and resources can be employed more effectively (Zhou et al., 2007).

One area that many times is overlooked is how to, in an efficient way, transfer the predictive maintenance data to a computerized maintenance management system (CMMS) system so that the equipment condition data is sent to the right equipment object in the CMMS system in order to trigger maintenance planning, execution and reporting. Unless this is achieved, the predictive maintenance solution is of limited value, at least if the predictive maintenance solution is implemented on a medium to large size plant with tens of thousands pieces of equipment.

2.2 OPC (OLE for Process Control)

Object Linking and Embedding (OLE) is a communication standard based on OLE technology from Microsoft that enables interoperability between industry automation and control systems (PLCs, sensors, motor controllers, etc.), field systems (e.g. Manufacturing Execution Systems) and business applications (e.g. ERP).

Object Linking and Embedding (OLE) for Process Control (OPC), is a standard specification developed by a work group of industrial automation in 1996. The standard defines objects, methods and properties to meet the needs of interoperability for process automation applications from different manufacturers in real time. These requirements include a standard technique for accessing field data from plant floor devices, the efficient transfer of data from one device to a business process application and the ability of a client to use multiple servers simultaneously (integration of data from different manufacturer's hardware devices).

Traditionally, every time data access from a device is needed, a custom interface has to be created. The goal of OPC is to define a common interface that is written once and then reused by all business packages, SCADA, HMI, or custom software. There is nothing in the specification to restrict the OPC server to provide access to a device control process. Once an OPC server is written for a specific device, it can be reused by any application that is able to act as an OPC client.

3 Research Methodology

This research is carried out in five steps. First of all, we reviewed the existing literature in areas such as Predictive Maintenance, OLE for Process Control and virtual enterprise for collaborative business. Secondly, we specified a hardware integration tool to apply to the field equipment, which is then integrated with a Virtual Enterprise Management Platform. Thirdly, a requirement elicitation process is carried out, which includes semi-structured interviews to two different business enterprises, namely; a machinery manufacturing SME located in the north of Portugal (Engineer-to-Order business model) and an electronic and automation SME located in the United Kingdom (Engineer-to-Order business model).

The results of these three steps conclude the requirements elicitation process through collecting the expected requirements list, which answers our first research question and functional specification, answering second research question. Fourthly, a set of discussions were carried out with the platform development team to understand the hardware integration and technical requirements. This fourth step answers to the third and last research questions. Finally, a Virtual Enterprise Management Platform is introduced on the shop floor integration hardware in order to get data from the field systems and controllers.

4 Virtual Enterprise Management Platform

The goal of a Virtual Enterprise Management Platform (VEMP) is to simplify the establishment, management, adaptation and monitoring of dynamic manufacturing processes in Virtual Factories. This includes the finding of partners, the design, forecasting and simulation of Smart Processes, and their execution and real-time monitoring.

To establish processes between different companies, data about the partners wishing to collaborate in a virtual factory is needed. Therefore, each Virtual Factory member needs to be able to add data about his company, products, services and processes. To achieve this in a user-friendly way, VEMP has to provide an editor in the scope a Data Provisioning and Discovery component to enter, view, update or delete this data. For reasons of availability, accessibility, access-control and the possibility to have redundant backups if needed, this data should be stored in the cloud. The Cloud Storage component should support several types of data storage like semi-structured data storage (for example for XML or JSON data) used internally by the VEMP, as well as semantic data necessary for semantic company descriptions and also data storage for binary files. Binary files may be used for storing documents such as specifications or even multimedia files.

To design the VEMP, the platform has to provide a Process Designer. To improve and facilitate the usability of all user interfaces should accessible via a single application interface with a single look and feel and a quick learning curve. All the user interfaces should therefore be embedded in the Dashboard, including the process designer.

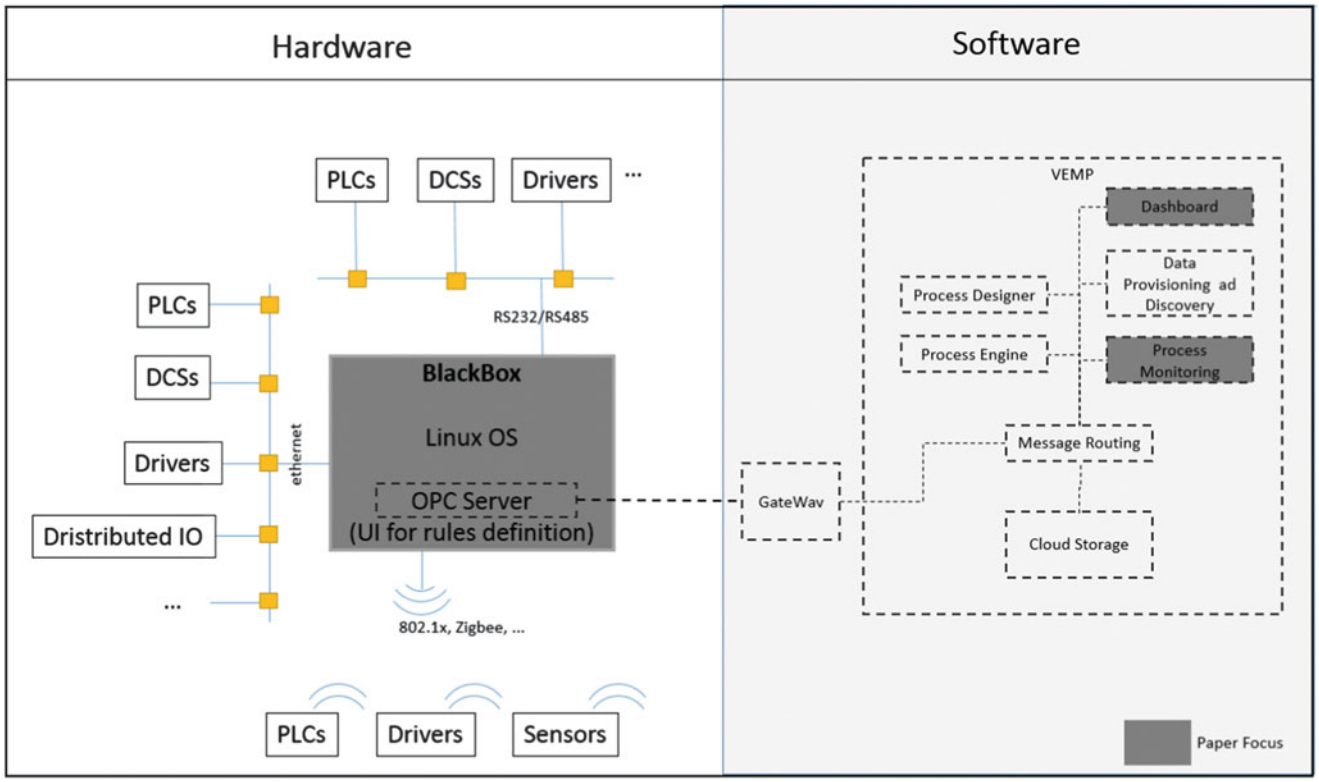


Fig. 1 Black box integration with the VE Management Platform (VEMP)

The Process Workflow Execution component, executes process models and the Real-time Process Monitoring shows the actual status of the process execution and can additionally query machines interfaces for their current state, collecting information for preventive and predictive analysis.

The Machine sensor interfaces as displayed in Figure 1 will be integrated via the Gateway, with acts as an OPC Client for the Black Box OPC server, the gateway will then communicate with the platform via the Message Routing. External systems like legacy systems etc. should also be able to communicate with the VEMP's Message Routing component making use of the Gateway component, which effectively fulfils the role of a bridge, connecting the external system with the platform. The Gateway as well as the Message Routing may invoke Transformation Services that can be used to translate between external (legacy) technology which will use different messaging protocols (including OPC), interfaces and message formats. The Gateways therefore will be the only components that might need to be expanded or recreated when a new Member wants to connect uncovered legacy systems to the Message Routing. The Message Transformation may be used as a base by the Gateways to transform a variety of data formats, hence allowing a wide support of systems.

4.1 Integration black box at shop floor

An integration black box is designed and developed to interface with the virtual enterprise's equipment to be monitored and managed. As displayed in Figure 1, the black box Runs an OPC Server on top of Linux Operating System, enabling the integration with the overall virtual enterprise platform via a Gateway implemented Services. The black box itself works as a smart object with sensors collects monitored raw data to be used for predictive maintenance from programmable logic controllers (PLCs), Industrial PCs, DCSs, Sensors from different vendors using different communication protocols and physical layers.

The black also counts with several interfaces such as Ethernet, 802.1x, RS482, RS232, USB as well as direct digital and analog IO which is helpful for old equipment's, where there's no communication at all and all is based on relays and discrete controllers (temperature, pressure, flow, timers, counters, encoders, etc. . .). Moreover, it counts with an intuitive User interface allowing the definition of alarm trigger rules. Different monitoring events are exposed through the gateway to the workflow engine with retrieves and update the information to the process monitoring.

4.2 Gateways

A gateway will comprise of standard components and custom components with functionality developed or created for connecting to a specific external system type and/or instance. A gateways mission is to communicate with a specific system, meaning that a significant part of a gateway implementation is tailored for specific (e.g. SAP ERP) technology or communication/interface protocol.

Gateways are only about connecting 3rd party systems, having an agnostic view about what is the content exchanged and what the format is. Checking and processing of the information gathered through the gateways will be performed at the destination components (as per example the Process Execution component).

A gateway is based on the following subcomponents:

- Communication: This sub-component is the bridge to the Message Routing component and will send and receive the messages on behalf of the Gateway. It will use a common interface to communicate with the Custom Connector component of the Gateway.
- Custom Connector: Set of customised adaptations that have to be performed in the 3rd Party System side to assure that the data exchanges are performed successfully.
- Black Box Integration: component that allow the black box (sensors, etc.) to be integrated into the platform. It can communicate directly to the Communication subcomponent.
- Legacy/ERP systems: Companies own systems (can also be SaaS systems), that manage company data that should be exposed to Virtual Enterprise Management Platform to assure the functionality envisaged.

4.3 Workflow process engine

The Workflow Process Execution component will be at the heart of the platform, as it will orchestrate all interaction in a virtual factory. Its purpose is to execute Processes, modelled in the Process Designer. This component will deal with Processes, Process Instances and the communication with gateways and logging. From a high level perspective Processes and Instances have the following minimum requirements:

- Processes: Each process has to have a set of attributes, independent of the language in order to capture the process model: Endpoints to enumerate a list of gateways/partners used in a particular process; Data elements (variables) to act as intermediate buffer for capturing results from calls to gateways and providing input for calls to gateways. Description to hold process mode.
- Instances: Each process has to be instantiated, each instance has to be stopped and started. The Name of an instance is intended to hold a description of the circumstances under which the process has been instantiated. The purpose of position is to show a set of activities that are currently executed by the instance. There can be multiple positions because there may be parallel execution of activities. While State will provide a fixed model consisting of a set of conditions and their transition, Status will hold a machine readable semantic/domain specific description of what is currently going on in the process: e.g. “Process is finished, but the results mixed”, “Process has an error which is related to machine X”, “The results of the process can be found in data element X”.

4.4 Process designer

The module’s main object is the Smart Process Model and it provides the required functionality to manage its lifecycle and perform tasks related to the different phases. A new Process Model is created in the Process Designer either as an empty model or based on a ready-to-use template from templates repository. As the broker starts to design (edit) the process, the model enters its ‘in design’ phase. This is the core phase for the Process Designer and the designer can perform several types of tasks in it:

- Manage process metadata with the goal to enable automation and make the process discoverable
- Design process models, e.g. add/configure/remove process activities or other process model elements, using the notation and semantics supported by the Process model design tool
- Use the simulation module to trigger simulation of the process and verify its qualities before executing it and potentially return to redesign for better results
- Use the optimization module to trigger optimization of the process for optimal business goal results and potentially return to redesign for better results
- Save for further work or share the designed process model
- Load, if the process model has been saved earlier or has been shared by another user

4.5 Process monitoring

The process monitoring data will show the actual status information, log and performance data relating the virtual factory processes. The details sub-components and their interdependencies with each other are presented in Figure 2.

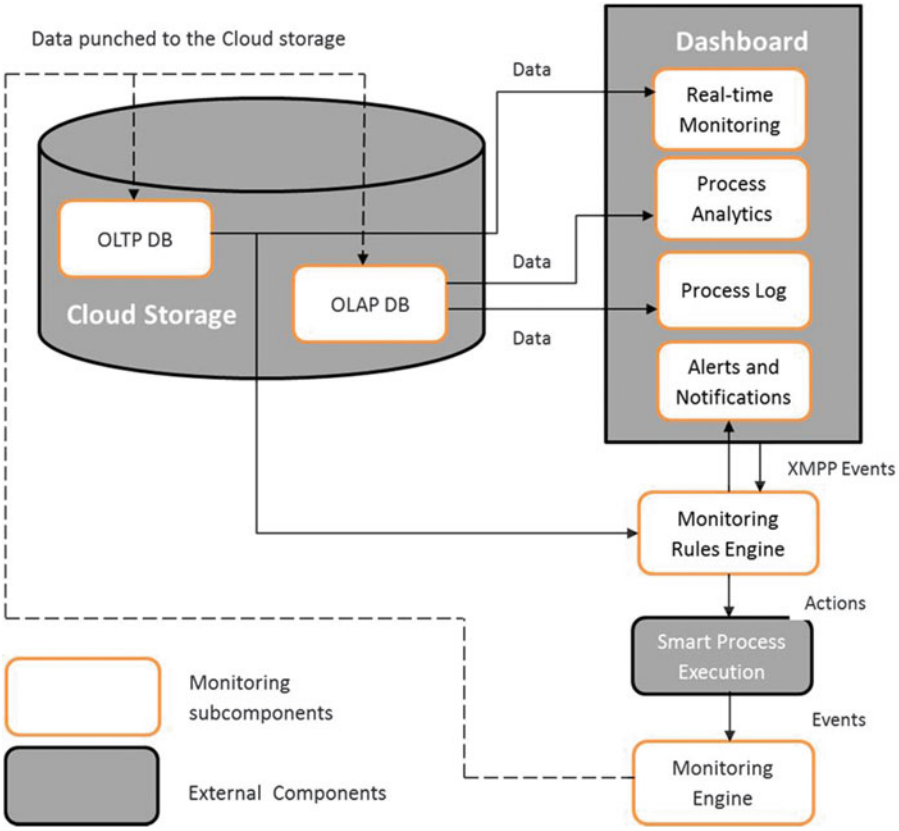


Fig. 2 Process Monitoring sub-components

From Figure 2 it is noticed that process monitoring component consists of the following sub-components:

- Real-time Monitoring
- Process Analytics
- Process Log
- Monitoring Rules Engine
- Monitoring Engine

Details of each of the sub-components can be explained as follows:

Real-time Monitoring: This sub-component shows the actual status of process instances. It uses the same user interface as Process Designer in order to have the same look and feel, allowing identification and tracking of the process instances.

Process Analytics: This sub-component provides an independent service that just queries finished process instances and show them to the user. It collects the key performance indicators associated with the manufacturing processes.

Process Log: It includes a search engine and shows historical data of process instances. It allows users to search for finished process instances and display its data in a graphical interface.

Monitoring Rules Engine: This sub-component mainly responsible to allow the definition of rules and corresponding actions. It throws alerts to the Dashboard, as well as performs action upon the Smart Process Engine. The rules are evaluated based on throwing events and notifications.

Monitoring Engine: It allow the visualization of key performance indicators related to the manufacturing processes and aggregates and analyses data. It provides a graphical display with the objective to track KPIs.

It is observed from Figure 2 that the Monitoring Engine receives data via XMPP events from the Smart Process Execution and stored the relevant event information in the Cloud Storage component, so that all data will be available for predictive maintenance analysis. The events data are stored within the cloud within two separate databases such as OLTP (OnLine Transaction Processing) database and OLAP (OnLine Analytical Processing) data base according to events types after following the XMPP protocol. The Real-time Monitoring sub-component displays a live view of the event data as stored within the OLTP database using the process editor interface. This data visualization helps virtual factory brokers to improve the performance of the manufacturing processes.

The Process Analytics sub-component displays the key performance indicators related to the manufacturing process over the Dashboard user interface, which were stored in the OLAP data base within the Cloud Storage. The Process Log also receives finished process instances from the OLAP database and visualizes in a graphical interface over the Dashboard as displayed in Figure 2. The Monitoring Rules Engine evaluates rules based on the events stored within the OLAP database in the Cloud Storage and triggers events and

notifications. Finally, the Alerts and Notifications sub-component provides the definition of the rules based on process execution delays which are evaluated by the Monitoring Rules Engine and throws alerts to the Dashboard.

5 ADVENTURE Dashboard: tool to VE process monitoring

In order to visualize the data as are collected from various sources (e.g. smart objects, sensors), an interactive user interface layer is designed and developed within the scope of this research. This user interface layer termed as 'Dashboard' displays the different processes monitored data in various formats (e.g. tables, graphs, texts). This data visualization offers real-time information update of the VE processes and resources which directly influence to overall decision making processes within the VE partners. A snap shot of the dashboard interface as developed within the scope of ADVENTURE project [ADVENTURE, 2011] is displayed in Figure 3.

From Figure 3, it is seen that the dashboard interface contains several porlets or widgets, such as, process instances list, my smart objects, resources, etc. These widgets are responsible for visualizing individual data or information according to the widget type. For instance, 'Resource' widgets as displayed in Figure 4 visualizes the corresponding resource status such as CO_2 footprint, energy consumption and steps finished per day. Often this

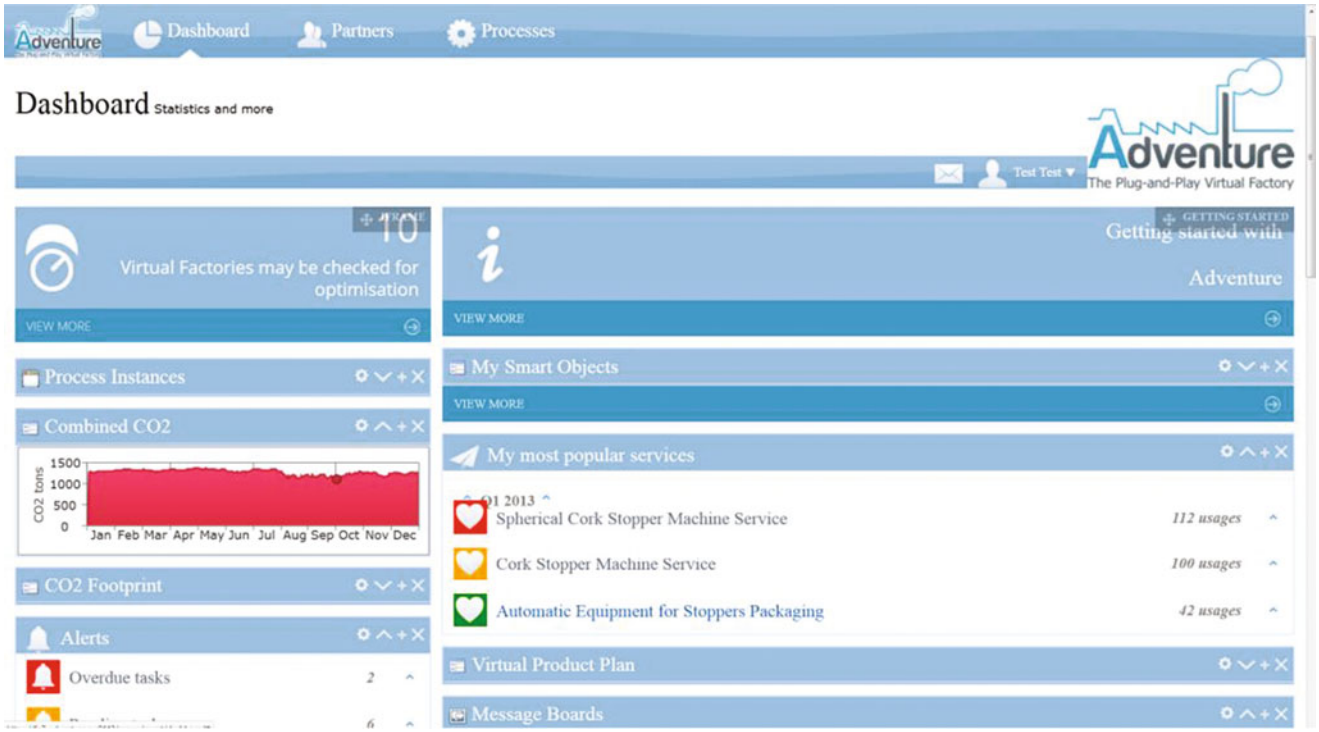


Fig. 3 A snap shot of the ADVENTURE Dashboard homepage

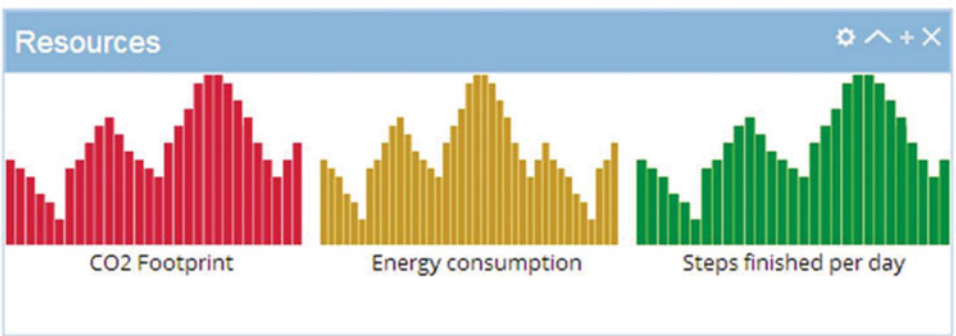


Fig. 4 Snap shot of ADVENTURE dashboard's Resources widget

widget also displays other relevant resource information such as resource ready, breakdown, downtime, shortage, etc.

6 Conclusions

The main focus of this research is to highlight a complete loosely coupled virtual enterprise management tool applied to the Predictive Virtual Enterprise Maintenance Processes. This tool composed of with nine modules: (i) The integration Black Box with OPC server, (ii) Gateway with OPC Client, (iii) Process Execution Engine, (iv) Message Routing, (v) Process Designer, (vi) Process Monitoring,(vii) Data Provisioning and Discovery, (viii) Cloud Storage and (ix) Dashboard. This study mainly highlights two components such as Process Monitoring and Dashboard that are directly interfaced with VE business process monitoring and management. All other components are the supporting ones and are responsible to execute the virtual enterprise management platform successfully. The integration 'black box' as highlighted in this research collects resources or equipment data and visualizes over the dashboard through gateway services. The collected data from an individual equipment or resource by the integration black box acts as the source of predictive maintenance of the specific equipment.

User interactive component Dashboard as presented in this study enables the overall data visualization that can be used as critical decision making process. The information visibility also ensures the current condition of the equipment and/processes which to be used as the necessary corrective actions in case of abnormality or failure.

Acknowledgements The authors would like to acknowledge the co-funding of the European Commission in NMP priority of the Seventh RTD Framework Programme (2007-13) for the ADVENTURE project (Adaptive Virtual Enterprise Manufacturing Environment), Ref. 285220. The authors also acknowledge the valuable collaboration provided by the project team during the research work.

References

ADVENTURE (2011), Adaptive Virtual Enterprise Manufacturing Environment, European RTD project, Grant agreement no: 285220, Duration 01.9.2011-31.08.2014.

Camarinha-Matos, L.M., Afsarmanesh, H,, Galeano, N. & Molina, A. (2008) "Collaborative Networked Organizations – Concepts and practice in manufacturing enterprise", Computers & Industrial Engineering.

Carneiro, L., et al.: An innovative framework supporting SME networks for complex product manufacturing. Collaborative networks for a sustainable world, Volume 336, p. 204-211 (2010).

Molina, A., Velandia, M. & Galeano, N. (2007) "Virtual Enterprise Brokerage: A Structure-driven Strategy to Achieve Build to Order Supply Chains", International Journal of Production Research, Vol. 45, No. 17, pp. 3853- 3880.

Moubray, J. (1997), Reliability-Centered Maintenance, 2nd edn, Industrial Press, New York.

Yam, R.C.M, Tse, P.W., Li, L. and Tu, P. (2002), "Intelligent predictive decision support system for condition-based maintenance", International Journal of Advance Manufacturing Technology, Vol. 17, pp. 383-391.

Zhou, X., Xi, L. & Lee, J. (2007) "Reliability-centered predictive maintenance scheduling for a continuously monitored system subject to degradation", *Reliability Engineering & System Safety*, Vol. 92, No. 4, pp. 530-534.

Module-based release management for technical changes

Günther Schuh, Sasa Aleksic, and Stefan Rudolf

1 Related research

The release planning methodology, which will be presented in the following, covers issues of the design of modular product architectures, technical change management and release management. This chapter purposes at introducing basic knowledge of the issues at the one hand and presents selected relevant previous approaches on the other.

1.1 Modular product architecture design

Modular product architectures represent one of the favorite tools of modern complexity management. ULRICH & EPPINGER define a product architecture as the assignment of the functional elements of a product to the physical building blocks of the same [1]. In a modular product architecture ULRICH & EPPINGER see every function of the product being exactly matched with one component and all components having defined interfaces between each other [1]. The term component relates as well to single parts as to whole assembly groups, which on a deeper level consist of single parts or other assembly groups again [2]. Looking at the design of modular product architectures, literature shows a myriad of approaches and concepts. The Modular Function Deployment (MFD) by ERICSSON and ERIXON postures a viable approach with the aim of facilitating a systematic design [3]. The approach focusses the development of a modular concept by the application of twelve module drivers which are the basis for mapping of functional requirements to certain modules. By doing so, the module drivers, e.g. „Planned Product Changes“, “Styling“ and “Upgrading” aim at a change oriented product structure [3]. Another notable approach is the Design for Variety by MARTIN [4], with a focus on both the creation of robust platforms for modular products and the reduction of interdependencies between system elements. In order to achieve this aim MARTIN defines the indices Generational Variety Index (GVI), a description of the effort for the design of a certain component, and the Coupling Index (CI), specifying the strength of the connection between the components of a product [4].

1.2 Technical change management

BELENER focused his technical change management approach on the deeper connection of technical change management and the explained modular product structures [5]. The main aspect of the approach is the analysis and management of change risks and subsequently the development of procedures for the analysis and evaluation of technical changes. This aspect is divided into four subcategories, the design of easily-modifiable modules, the classification of changes, release scheduling and technical change management [5]. The analytical network process (ANP) approach by HAKYEON ET AL. follows a similar train of thought, resulting in a tool for the comparative evaluation of multiple technical changes in a modular product [6]. The ANP serves the purposes of both selection and prioritization of technical change alternatives, but also offers the ability to take indirect impacts among elements into account. In order to represent change relationships among parts and modules, a design dependency network of the product is modelled in this approach. Consequently the relative change impact (RCI) values of the modules or parts are defined [6].

1.3 Release management

The term “release” is defined in information technology as the collection of one or more new or changed configuration

G. Schuh (✉) • S. Aleksic • S. Rudolf
RWTH Aachen University, Laboratory for Machine Tools and Production Engineering WZL, Steinbachstraße 19, 52074 Aachen, Germany
e-mail: G.Schuh@wzl.rwth-aachen.de

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_44,

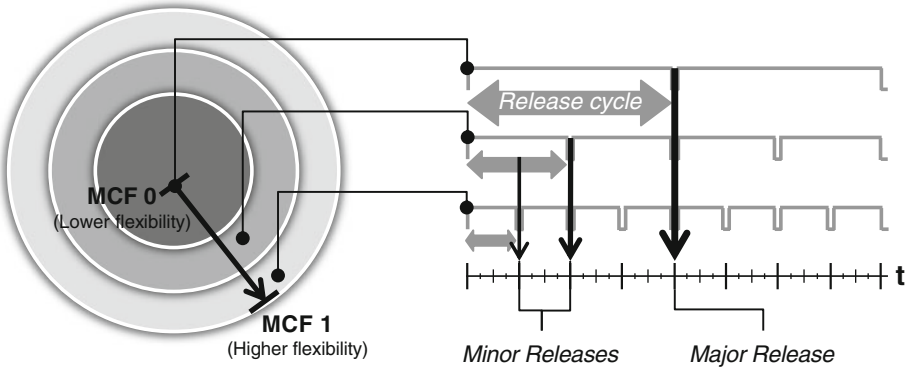


Fig. 1 Onion peel model and resulting release-cycles

items deployed into the live environment as a result of one or more changes [7]. In information technology there exist standardized proceedings for the management of releases, as well as profound research into the timing and prioritization of changes and releases, offering well developed tools for those matters [8–10]. Nevertheless the different entities of software make the adaption to physical product structures difficult. Although opportunity and potential for the adaption of IT proceedings might exist in a later stage, they are not yet in focus of research, since the basic requirements differ and IT research admittedly is already concerned with more mature problems. Release management in the background of this research paper includes the planning, collection, synchronization and time wise separation and implementation of technical changes of modules within a modular product architecture. Set time-frames of reoccurring releases are referred to as release-cycles.

1.4 Need for research

While ERICSSON and ERIXON'S approach for modular product architecture design does not address future planned and unplanned technical changes [3], MARTIN provides a bottom-up approach for the handling of planned technical changes [4]. Nevertheless the determination of the flexibility of modules is not focused in detail and there is no planning for the further lifecycle of the architecture. The technical change management approaches intensely examine the handling of actual changes to the product and do not set the focus on implementation strategies and implementation into a product architecture [5, 6].

2 Flexibility of modules and release-cycles

The flexibility of a module in the product architecture needs to be quantified to support future planning of changes. Five basic measurement dimensions are chosen to be evaluated. Consumer demands implement market driven requests as a dimension in the evaluation. Costs as a dimension are separated into production costs, one-time complexity costs and running complexity costs. The last dimension is represented by the module interaction. The module interaction is examined from the perspective of active interaction, which describes active influences from one module on others and from passive interaction view, which describes the level of impact by others [11]. Obviously the above mentioned dimensions do not only influence the flexibility of a module, but influence each other, too.

$$MCF_g = \frac{1}{\sum_{(i=1)}^{5} g_i \times x_i} \times \sum_{i=1}^{5} g_i \times D_{MCF_i} \qquad (1)$$

To reduce complexity and provide transparency for the process, a module change flexibility classification number (MCF) is derived from these five dimensions. D_{MCFi} represents the classification number of each dimension with a maximum value of nine. The factor g_i is used to weigh a dimension and makes the process adjustable for different companies and targets. The sum of their product is taken and normalized by the sum of the product of the weighting factor and x_i, which is consistently given the maximum value of nine. Therefore MCF ranges between zero and one. Finally the MCF is put into the onion peel model visualization in which the MCF increases from the inside to the outside (see Fig. 2). Modules at the inner shell are rather not flexible, while modules at the outer shell are of high flexibility. Latter produce a small amount of work when they are changed, while modules with low flexibility influence other modules strongly and therefore generate a lot of effort when being object of change.

With this results the release-cycles for modules of the product architecture can be systematically planned. Technical changes and innovations shall only be implemented in predefined release-cycles. In that way, larger numbers of technical changes and innovations can be unified and synergy effects are used. The release-cycles themselves are structured into different release frequencies which are multiples of each other. High flexibility modules are chosen

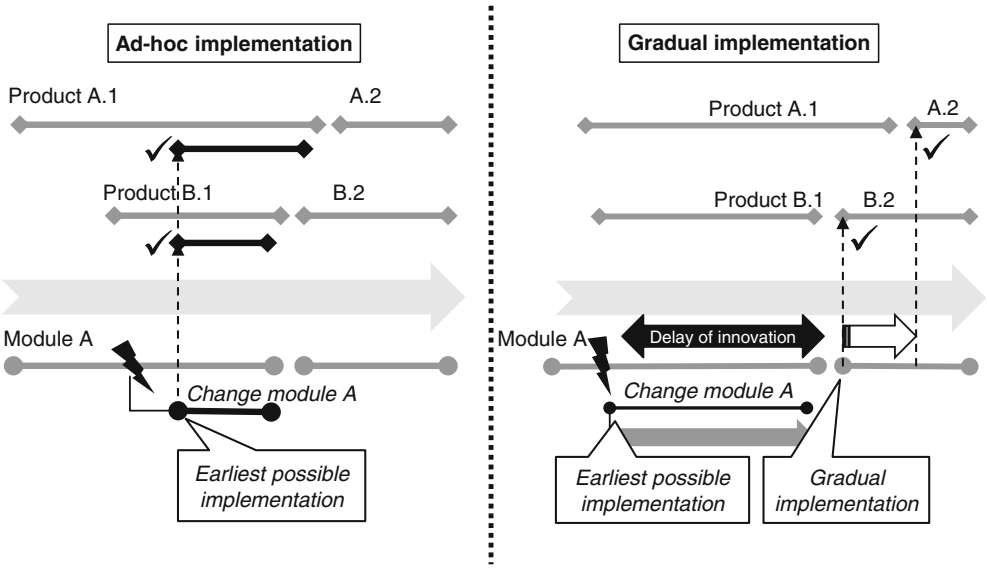


Fig. 2 Strategies of introduction

to be changed more frequently, because of their lower change effort. Modules with lower flexibility are assigned to higher frequencies. Thus, there are "major releases" where changes are bundled and implemented in each release-cycle at the same point of time and as a consequence "minor releases" in which changes are applied in a smaller number of release-cycles.

At this point each module is structured into its individual innovation frequency. To reduce complexity, the interdependencies of the modules are optimized. This is achieved by examining and rating the correlations and interactions of the modules. The direction (outgoing, ingoing) is taken into account, at that point the correlation is classified into a physical, a force, a signal and a material interaction. After this the particular sensitivity of the interaction is valued. In that way, a sensitivity of nine represents high influence on the module when a correlated module experiences a small change. The results are put into the Module-Correlation-Matrix (MCM). The MCM is then used for optimization of the module interdependencies by determining significant fields of the matrix. For example a field with at least one value of nine is considered significant. Fields where higher frequency modules influence low frequency modules are of special interest for the examination. These fields need to be optimized by either lowering the frequency of the influencing module or reducing the sensitivity by redesigning the interface or using "overdesign". The changing of the component specification should be considered as well. If the field is within the same frequency cluster, the fusion of the modules can be taken into account. For a significant field, where a lower frequency module influences a higher frequency module, optimization is only recommended if the effort is small. As a result, the modules of the product architecture are organized into optimized, predefined release-cycles for the implementation of future innovations and adjacent technical changes.

3 Release Management

The approach of implementing release management into lifecycle management of modular product architectures aims at realizing sustainable success with the modular approach. At this point sustainability is understood as to minimalize effort, reduce consequences of changes, as well as maximizing economies of scale. Two dimensions build the framework for the identification of implementation strategies - urgency and consolidation potential. The urgency of the handled change is determined in a first step. This is followed by the analysis of the consolidation potential for the bundling with other changes. The next step is the deduction of the nature of change for the module, since there is the possibility to change the existing module or to create a new module variant. The final and key step is the identification of the implementation strategy in form of an effort-utility portfolio.

3.1 Analysis of the urgency of technical change

The assessment of the urgency of change aims at determining whether a technical change needs to be implemented into the product portfolio straight away or possibly at a later point of time. All changes which do not require immediate implementation, are potentially able to be consolidated with other

changes and thus need to be analyzed for their optimum time of synchronization with the product portfolio. The urgency criteria is analyzed first, since it is a K.O. criteria which makes further analysis obsolete in case the reasons of change do not allow any delay.

To enable efficient release-building three dimensions of urgency are defined - *„immediate implementation*“, *„contemporary implementation„* and *„soonish implementation*“. An immediate implementation requires the change to be addressed straight away and is not subject to further examination with regard to a later synchronization with the product portfolio. “Contemporary implementation” and “soonish implementation” both allow the change to be at least postponed until the next defined point of release.

3.2 Analysis of the consolidation potential for bundling

In possible scenarios usually multiple changes are in the planning stage at the same time. Considering that some changes will always affect the same module or even component, a consolidation of these changes appears logical. This way multiple changes to one component over a short time span and thus additional effort can be minimized. In an ideal surrounding upcoming changes are always to be carried out consolidated and thus in one go.

To identify a potential for consolidation with other changes two steps appear necessary. On the one hand a collection of all changes in relation to one module has to be created. This seems relevant in order to gather both the causes of change, the time needed for the change and the affected components and sub-modules. On the other hand all restrictions in terms of resources like e.g. time, personnel, capital and machine capacity have to be examined. This enables the identification of combinations which are limited by those resources and thus not feasible.

3.3 Nature of change of the module

By its definition and nature modules can be developed, produced and flexibly combined or replaced with other modules independently [12]. When planning a change within a modular product the inevitable decision has to be made, how the change is to be conducted inside the architecture. This decision implicates two possible ways of acting. On the one hand the existing module could be modified while on the other hand a completely new module variant could be developed. The second option offers the opportunity to only make the relevant adjustments to the module or to develop a new variant of the module. Nevertheless both ways should lead to a state in which all technical requirements are met. Disregarding which option is being chosen the existing number of variants is increased for at least a short time until the final synchronization with the complete relevant product portfolio. The length of this time with less commonalities and thus less economies of scale depends on the chosen implementation strategy which will be determined in the next step.

3.4 Identification of the implementation strategy

This last step of the procedure includes the identification of the optimal implementation strategy of a technical change within a module. To achieve this all relevant implementation strategies will be analyzed in an effort-utility portfolio.

In this research paper two basic implementation strategies are distinguished. On the one hand the *ad-hoc implementation* of change and on the other hand the *gradual implementation* of change. Based on the analysis of the consolidation potential these two strategies can be split up further into consolidated changes and single changes, creating four possible strategies out of the two basic strategies.

The *ad-hoc implementation* strategy is characterized by the fact that changes to many elements of the product portfolio are conducted disregarding the timing of the follow up products at one point of time. Due to this fast reaction high costs of change are accepted to keep the delay of innovation as short as possible. In addition to the high costs of change another disadvantage has to be mentioned, namely the high strain on resources and capacities. Nevertheless, an advantage not yet mentioned is the possibility to maintain the product market cycles and not having to prepone or postpone market introductions of future products. One of the most important points is the realization of high economies of scale, which remain on the highest possible level during the conduction of change due to the direct introduction of the changed module into all products.

The *gradual implementation* strategy is characterized by the consecutive introduction of the change into the product portfolio. Thus different advantages are realized, like the compliance with the product market cycles, as well as the low costs for the change since not all products have to be changed at the same time. As a disadvantage it has to be noted that certain products are exposed to an increased delay of innovation, as seen in Figure 3. Furthermore the economies of scale for the module also only reach the optimum, when the change has been implemented into all products after a period of time.

It is examined at this point, by which means the single change is to be conducted and what the result means for the whole product architecture. The explicit target is to minimize effort and maximize utility where possible.

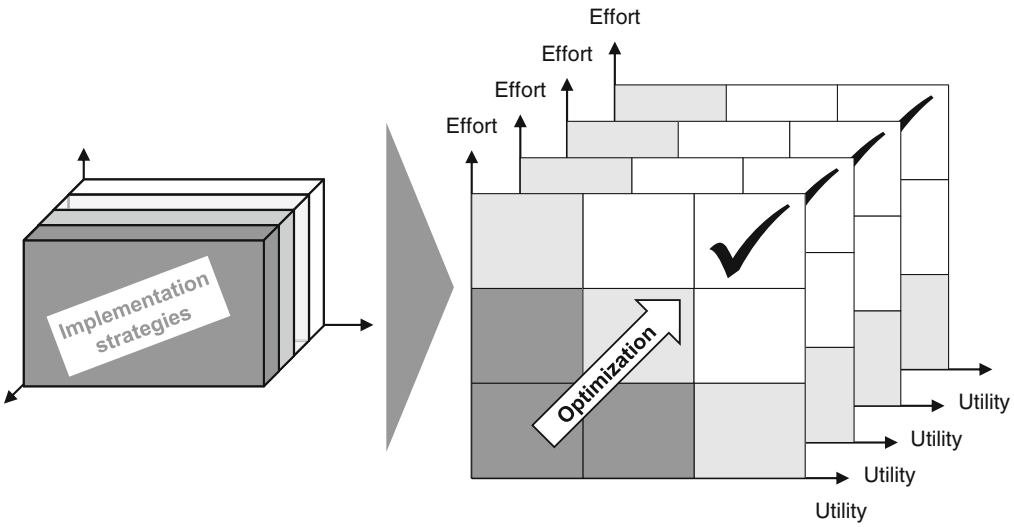


Fig. 3 Effort-utility portfolio

To determine the effort dimension both the production costs and the complexity costs have to be analyzed. In order to calculate the production costs of a module by terms of unit costs accounting, manufacturing costs and material costs have to be identified [13]. The complexity costs are determined by a resource oriented process costs calculation [14]. This is done by identifying the complexity drivers along the process line. The overall cost of the change are then calculated by comparing the initial state of the module with the changed state.

With regard to the utility both internal and external utility have to be taken into account. For internal utility both cost effective and cost ineffective utility changes are relevant, while for the external utility different drivers can be identified. One important factor here is the fulfillment of market requirements. To measure the external utility the Kano model is employed, which differentiates between the fulfillments of basic needs, performance needs as well as delighters [15]. Finally the aggregated utility has to be calculated and weighted.

Figure 3 shows the utilization of the effort-utility portfolio. The two initial basic implementation strategies, ad-hoc or *gradual implementation*, are split up further into four implantation strategies. The determining factor is whether the implementation includes consolidation or not. The four strategies are analyzed in the effort-utility portfolio and the analysis ends with the selection of the strategy with the most favorable result.

4 Conclusion

The analysis of the related research showed the initial necessity of creating an approach for release management in modular product architectures. While the onion peel model shows the flexibility of a module, the release-cycles can be determined by the examination of the flexibility. In series production the approach enables the analysis whether the requirement for the change is urgent and has to be synchronized straight away with the product portfolio or whether the urgency is low and the synchronization can take place at a later point of time. If the latter is the case, the approach gives the opportunity to bundle technical changes and release them at a defined release-point. This research is detailed further in the future and has to be validated in a practical case study.

References

1. Ulrich, K.T., Eppinger, S.D.: Product design and development. 2. Vol., Irwin/McGraw-Hill, Boston (2000)
2. Göpfert, J.: Modulare Produktentwicklung: zur gemeinsamen Gestaltung von Technik und Organisation. Dt. Univ.-Verl. [u.a.], Wiesbaden (1998)
3. Ericsson, A., Erixon, G.: Controlling design variants. Society of Manufacturing Engineers, Dearborn (1999)
4. Martin, M.V.: Design for variety: A methodology for development of product platform architectures. Mechanical Engineering. Stanford University, Stanford (1999)
5. Belener, P.: Technisches Änderungsmanagement modularer Produkte und Prozesse. Dissertation, Bochum (2008)
6. Hakyeon, L., Hyeonju, S., Nakhwan, S., Yoo S., S., Yongtae, P.: An analytic network process approach to measuring design change impacts in modular products. In: Journal of Engineering Design 21 (1), p. 75–91 (2010)
7. IEEE Standard, Adoption of ISO/IEC 20000-2:2012, Information technology, Service management, Part 2: Guidance on the application of service management systems. IEEE Std 20000-2-2013. In: IEEE Std 20000-2-2013, p. 1–105 (2013)
8. Van Bon, J., Wilkinson, J.: Foundations of IT Service Management: basierend auf ITIL V3. 3. Vol..Van Haren Publishing, Zaltbommel (2008)

9. Carlshamre, P.: Release Planning in Market-Driven Software Product Development: Provoking an Understanding. In: Requirements Eng 7 (3), S. 139-151 (2002)
10. Wright, H.K.: Release Engineering Processes: Their Faults and Failures. Dissertation. University of Texas at Austin, Austin (2012)
11. Schuh, G., Arnoscht, J., Aleksic, S.: Systematische Gestaltung von Kommunalitäten in Produkten und Prozessen. In: ZWF Produktentwicklung, 107, No. 5, p. 322–326 (2012)
12. Rudolf, S.: Produktionsgerechte Baukastengestaltung. 1. Vol. Apprimus-Verl. (Produktionssystematik, 2013,30), Aachen (2013)
13. Kemmetmüller, W., Bogensberger, S.: Handbuch der Kostenrechnung. Das Grundlagenwerk zu Kostenrechnung und Kostenmanagement. 8., aktualisierte und erw. Aufl. Servive-Fachverl., Wien (2004)
14. Schuh, G.: Lean Innovation. Der deutsche Weg; VDI-Buch. Springer, Berlin (2013)
15. Sauerwein, E.: Das Kano-Modell der Kundenzufriedenheit. Reliabilität und Validität einer Methode zur Klassifizierung von Produkteigenschaften. Dt. Univ.-Verl. (Gabler Edition Wissenschaft), Wiesbaden (2000)

Trajectory Optimization by Particle Swarm Optimization in Motion Planning

Jeong-Jung Kim and Ju-Jang Lee

1 Introduction

Motion planning finds a sequence of actions that move from an initial configuration to a goal configuration that satisfy constraints. The problem can be solved optimally when a dimension of a workspace and number of action is low. However, when a dimension of a workspace is high and additional constraints such as obstacles, velocity, and force are added to a problem, there is no optimal algorithm to solve it. Such problems only can be solved approximately because a calculation time of the problems is exponentially increased as the number of the state space is increased [1].

The primary goal in motion planning is a collision-avoidance. Previously, a simple methods such as a potential field method [6, 7] or an elastic band method [8] produces a trajectory having no collision. However, those methods are limited in low-dimension problems and cannot deal with complex constraints. Optimization-based motion planning methods have been recently proposed for dealing with complex constraints. In these approaches, a trajectory is divided into N steps and the trajectory modification amount is encoded into parameters. The parameters are optimized with covariant Hamiltonian optimization in [9, 10] and derivative-free stochastic optimization, path integral [4, 5] in [11]. The stochastic optimization-based method shows fast convergence rate, high success rate and possibility of dealing with various constraints. However, the method is based on a step-based optimization that only consider a cost for each step and has a high chance to produce a trajectory that is stuck in local minima.

In this paper, trajectory optimization by particle swarm optimization (PSO) is suggested. A parameterization of amounts of the trajectory modification and a cost function for constraints are designed for PSO. PSO is a population-based stochastic global optimization method inspired by social behavior such as bird flocking or fish schooling [2]. The main advantages of the PSO are simple to understand, easy to implement and fast in convergence compared to other stochastic global optimization method. Its advantage is utilized in the proposed trajectory optimization methods. A normalized-step cost (NSC) concept is also suggested and it is used for an initialization of particles in PSO.

The organization of the paper is as follows. In section 2, the trajectory optimization by PSO is provided. In section 3, a simulation setup is addressed and results of the proposed method are analyzed. Finally concluding remarks and discussion are provided in section 4.

2 Trajectory optimization by particle swarm optimization

2.1 Encoding of trajectory modification vector for PSO

Particle swarm optimization (PSO) is a population-based stochastic global optimization method inspired by social behavior such as bird flocking or fish schooling [2]. The main advantages of the PSO are simple to understand, easy to implement and fast in convergence compared to other stochastic global optimization methods. Trajectory modification amounts are encoded in particles and they are optimized with PSO. When PSO is used, encoding of candidate solution to the particles and cost function should be designed. In the case of the time duration from a start configuration to a goal configuration is T and sampling time is dt, number of steps for the trajectories is $TN = T/dt$. The amounts of trajectories modification $\mathbf{P}[t]_{mod}$ for the TN steps is encoded into the particles. $\mathbf{P}[t]_{mod}$ can be represented in Cartesian coordinate or rotation coordinate. When an initial trajectory $\mathbf{P}[n]_{init}$ is given, final trajectory $\mathbf{P}[t]$ is defined as

J.-J. Kim (✉) • J.-J. Lee
Department of Electrical Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea
e-mail: rightcore@kaist.ac.kr; jjlee@ee.kaist.ac.kr

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_45,

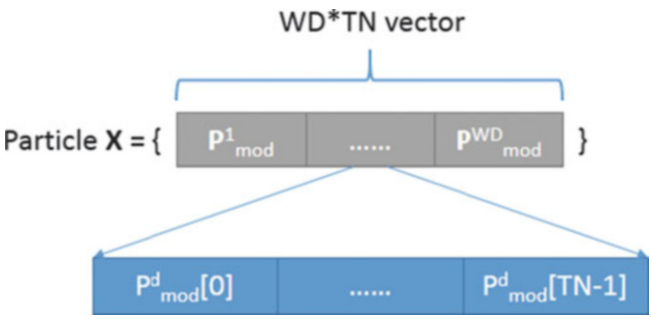


Fig. 1 $D = WD \times TN$ dimensional modification vector. A modification vector in each workspace is consists of TN dimensional vector. The D dimensional vector is encoded into a particle of PSO.

$$\mathbf{P}[n] = \mathbf{P}_{init}[n] + \mathbf{P}[n]_{mod}, \quad \text{for } n = 0, \cdots, TN-1. \tag{1}$$

There are $D = WD \times TN$ dimension of possible modification points if dimension for a workspace is WD. The WD modification vectors that each has TN dimension are stacked into a particle and it has D dimensions. The structure for a particle is shown in Fig. 1.

2.2 Cost function design

The particles are optimized by minimizing the following cost function

$$C(\mathbf{P}[n]) = \sum_{i=1}^{NC} W_i Const_i(\mathbf{P}[n]) \tag{2}$$

where $Const_i$ is a function that represents each constraint, W_i is a weighting factor that balances the importance for each constraint and NC is a number of constraints.

First constraint is for obstacle avoidance that has a form of

$$Const_{obsavo}(\mathbf{P}[n]) = \sum_{o=1}^{O} \delta(R[o], \|\mathbf{Obs}[o] - \mathbf{P}[\mathbf{n}]\|), \quad \text{for } n = 1, \cdots, TN-2 \tag{3}$$

where $\mathbf{Obs}[o]$ is a center location of obstacle, $R[o]$ is radius of an obstacle o, O is a number of obstacles, and a function $\delta(a_1, a_2)$ is an activation function that has a one when a_2 is smaller than a_1 and zero in another case. In this paper, obstacles are simplified to spheres for a convenience.

Second constraint is for a minimum distance between steps that has a form of

$$Const_{mindis}(\mathbf{P}[n]) = |\mathbf{P}[n+1] - \mathbf{P}[n]|, \quad \text{for } n = 0, \cdots, TN-2 \tag{4}$$

where function $\| p_1 - p_2 \|$ is a Euclidean distance measure between two points p_1 and p_2. A trajectory that has a short distance produces a low cost value.

Third constraint is for minimum acceleration that has a form of

$$Const_{minacc}(\mathbf{P}[n]) = \|acc(\mathbf{P}[\mathbf{n}])\|, \quad \text{for } n = 1, \cdots, TN-2 \tag{5}$$

where function $| acc(p) |$ is a function that calculates a magnitude of acceleration for a step and can be calculated with numerical differentiation. It is calculated for each dimension of workspace and an normalization is conducted for the cost term. Usually minimum acceleration is important for saving energy and avoiding a sudden change of motion.

Fourth constraint is for a minimum modification of a trajectory that has a form of

$$Const_{minmod}(\mathbf{P}[n]) = \sum_{d=1}^{WD} \left| P^d_{mod}[n] \right|, \quad \text{for } n = 0, \cdots, TN-1. \tag{6}$$

The constraint is applied when some parts of an initial trajectory should be maintained.

The total cost function to be optimized is

$$\begin{aligned} C(\mathbf{P}) = \sum_{n=0}^{TN-1} \{ & W_{obsavo} Const_{obsavo}(\mathbf{P}[n]) \\ & + W_{mindis} Const_{mindis}(\mathbf{P}[n]) \\ & + W_{minacc} Const_{minacc}(\mathbf{P}[n]) \\ & + W_{minmod} Const_{minmod}(\mathbf{P}[n]). \end{aligned} \tag{7}$$

The final cost value is summed up for each cost value between 0 and $TN - 1$.

2.3 Normalized step cost-based particle initialization

The initialization of particles in PSO is an important element because a quality of optimization is determined by that. In PSO community the particles are uniformly initialized in a search space. However, when it is applied to the optimization that is dealt in the paper, the method produces poor result because search space is huge. In the previous section, the cost for PSO was calculated for whole trajectories. It is changed to calculate for each step and used for initializing particles. A cost value is calculated by multiplication between degree of constraint violation and weighting factor. In this paper a normalized-step cost (NSC) is calculated by

dividing a step cost with weight factor that has a maximum value as in

$$NSC(\mathbf{P}[n]) = \sum_{i=1}^{NC} W_i Const_i(\mathbf{P}[n])/W_{max}. \quad (8)$$

The procedure for obtaining the vector is shown in Algorithm 1.

The normalized-step cost is used as a standard deviation of normal distribution and the mean of the distribution is set to zero. So each particle is initialized with Normal distribution, $N(0, NSC[n]^2)$. In the motion planning problem, the weight for obstacle avoidance has the largest value. So the standard deviation is similar to the distance between the trajectory and obstacle bound. The steps where their cost value are high are explored than other steps to find parameters that reduce the cost value. The procedure for the initialization of particles with NSC is shown in Algorithm 2.

2.4 Overall optimization process of trajectory optimization

The optimization procedure of the trajectory optimization is summarized in Algorithm 3 and it is identical to the procedure of PSO except the initialization step and the encoding method. First, the positions of particles are initialized in the search space with the Algorithm 2. And then cost value of each particle is evaluated with Eq. 7 and positions of particle are updated until stopping condition is not true. Particles store their best experience during optimization process and velocity and position of each particle are updated by Eq. 10 and Eq. 9 respectively.

Algorithm 1 Procedure for calculating normalized-step cost vector

Require: $\mathbf{P_{init}}$ and $\mathbf{P_{mod}}$
$\mathbf{P}[n] = \mathbf{P}_{init}[n] + \mathbf{P}_{mod}[n], n = 0, ..., TN - 1$
$\mathbf{Cost}[n] \leftarrow 0, n = 0, ..., TN - 1$
$\mathbf{NSC}[n] \leftarrow 0, n = 0, ..., TN - 1$
$W_{max} \leftarrow max\{W_i\}, i = mindis, obsavo, minacc, minmod$
for $n = 0, \ldots, TN - 1$ **do**
  if $n == TN - 1$ **then**
    $\mathbf{Cost}[n] += W_{minmod} Const_{minmod}(\mathbf{P}[n])$
  else if $n == 0$ **then**
    $\mathbf{Cost}[n] += W_{minmod} Const_{minmod}(\mathbf{P}[n])$
    $\mathbf{Cost}[n] += W_{mindis} Const_{mindis}(\mathbf{P}[n])$
  else
    $\mathbf{Cost}[n] += W_{minmod} Const_{minmod}(\mathbf{P}[n])$
    $\mathbf{Cost}[n] += W_{mindis} Const_{mindis}(\mathbf{P}[n])$
    $\mathbf{Cost}[n] += W_{obsavo} Const_{obsavo}(\mathbf{P}[n])$
    $\mathbf{Cost}[n] += W_{minacc} Const_{minacc}(\mathbf{P}[n])$
  end if
end for
for $n = 0, \ldots, TN - 1$ **do**
  $\mathbf{NSC}[n] = \mathbf{Cost}[n]/W_{max}$
end for
return NSC

Algorithm 2 Initialization of particles with normalized-step cost

Require: NSC
$X[idx] \leftarrow 0, idx = 0, ..., TN \times WD$
for $d = 0, \ldots, TN - 1$ **do**
  for $wd = 0, \ldots, WD - 1$ **do**
    idx = d × TN + wd
    X[idx] = $NormalDistribution(\mu = 0, \sigma^2 = \mathbf{NSC}[d]^2)$
  end for
end for
return X

Algorithm 3 Overall procedure for trajectory optimization

```
Initialize N particles that have D-dimensional vector with Algorithm 2
while stopping condition is not true do
    for each particle n = 1, ..., N do
        Update the velocity using Eq. 10
        Update the position using Eq. 9
        Fitness evaluation usng Eq. 7
    end for
end while
```

$$\mathbf{X}(i+1) = \mathbf{X}(i) + \mathbf{V}(i+1), \tag{9}$$

$$\mathbf{V}(i+1) = |randn|(\mathbf{p}_{best} - \mathbf{X}(i)) + |randn|(\mathbf{g}_{best} - \mathbf{X}(i)). \tag{10}$$

where $\mathbf{V}(i+1)$ is a velocity and $\mathbf{X}(i+1)$ is a position of a particle at $i+1$ iteration, and $|\ randn\ |$ is positive random numbers generated according to the absolute value of the Normal distribution, i.e., abs[N(0, 1)]. And $\mathbf{p}_{best}$ is a previous best position for each particle and $\mathbf{g}_{best}$ is a best position of the whole particle obtained so far. The velocity update Eq. 10 is different from an original PSO velocity update equation and an equation suggested by R. A. Krohling [3] is used in the paper. Advantage of the equation is that it does not require PSO cognition learning factor and social learning factor and converges faster than canonical PSO. As increase an iteration, the particles cooperate and finally reach optimized parameters.

3 Simulations

In this section, the proposed method is verified by simulations. The robot for the simulations is shown in Fig. 2 and the parameters are $L1 = 1.\ 7$ cm, $L2 = 1.\ 7$ cm, $L3 = 6.\ 7$ cm, and $L4 = 10$ cm. We generated a minimum jerk trajectory [12] in joint space that starts from $(0, -1.7, -18.4)cm$ to $(12.\ 0, -4.\ 0, 4.\ 0)cm$ and it was used as an initial trajectory. Obstacles were located at $(15.0, -1.0, -5.0)cm$ and $(10.0, -2.0, -14.0)cm$ and their radius are $4cm$ and $2cm$ respectively. The trajectory collides with the obstacles with this setting. The settings for an initial trajectory and obstacles are shown in Fig. 3.

To adapt a trajectory to constraints, the trajectory was modified with STOMP [11] and trajectory optimization by PSO (TOP). Number of sampling in STOMP was set as 100 and number of particles in PSO was set 100 and maximum iteration was set as 100. The simulations were conducted by 20 times for each simulation and maximum, minimum, and average value were included because the STOMP and TOP are stochastic optimization method. The cost function for the simulations was set as Eq. 7 and the weight parameters were set as $W_{mindis} = W_{minacc} = W_{minmod} = 10$ and $W_{obsavo} = 1000$ because the obstacle avoidance is important factor in motion planning. In this setting, each algorithm tries to minimize the cost function.

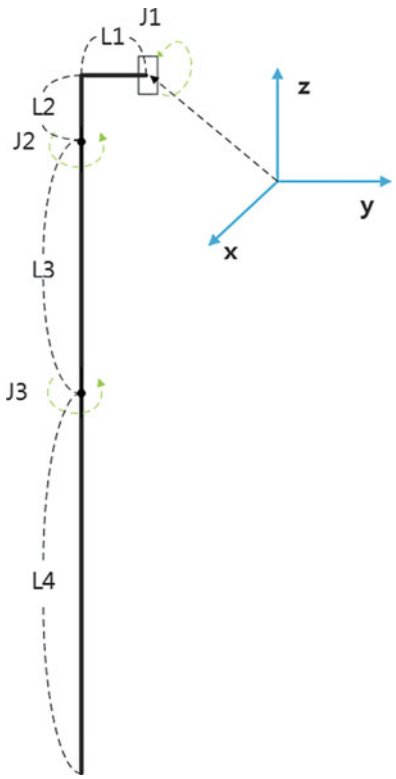


Fig. 2 Kinematic model of arm. The robot has 3 joints and 4 links. First joint rotates along y axis and second and third rotate along x axis.

The trajectories in each axis for STOMP and TOP is shown in Fig. 4 and Fig. 5. The steps that collide with obstacles were modified and the trajectory generated with TOP is smoother than the trajectory generated with STOMP. The cost value for each generation is shown in Fig. 6. As shown in Fig. 6, STOMP produces a modification vector that shows higher cost than TOP. As generation goes, the cost value by STOMP was stuck into a value and cost value by TOP was decreased. Those result can be interpreted that it was came from the differences between a local optimization

and a global optimization. STOMP calculates the cost value for each step and updates the step base on the cost value and TOP calculates the cost value for whole steps and updates steps base on the cost value of whole trajectory. Usually local optimization has more chance to fall into local minima. The phenomenon was also appeared in the result.

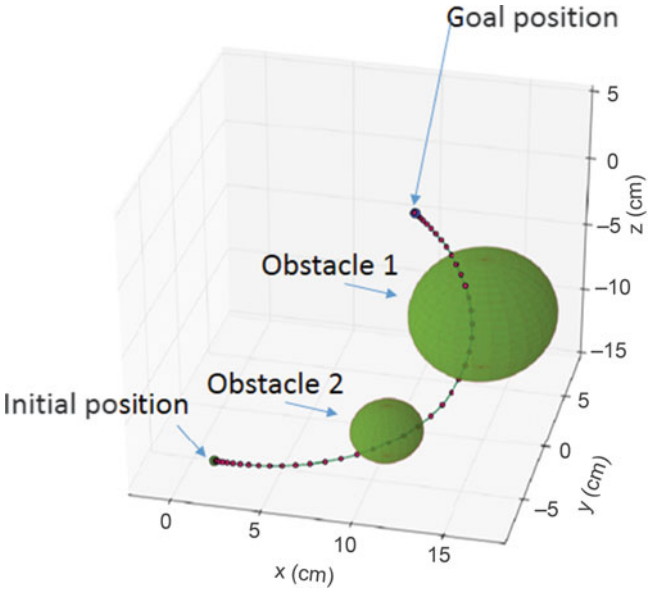


Fig. 3 Settings for initial trajectory and obstacles. STOMP and TOP modify initial trajectory for reducing violation of constraints

4 Conclusion

In this paper, trajectory optimization by particle swarm optimization was proposed. The PSO was used as trajectory optimizer and cost function design for additional constraints, normalized step cost concept, particle initialization method for motion planning were provided. The simulations for generating a trajectory that reaches a goal point in three-dimensional with three link robot were conducted. The proposed particle initialization method produced a particle that modifies parts that has a high cost and try to unchange parts that has a low cost. We expect that the proposed can be applied to trajectory optimization problems that should consider various constraints with less chance to fall into a local minimum.

Acknowledgment This research was supported by the MOTIE (The Ministry of Trade, Industry and Energy), Korea, under the Technology Innovation Program supervised by KEIT (Korea Evaluation Institute of Industrial Technology),10045252, Development of robot task intelligence technology.

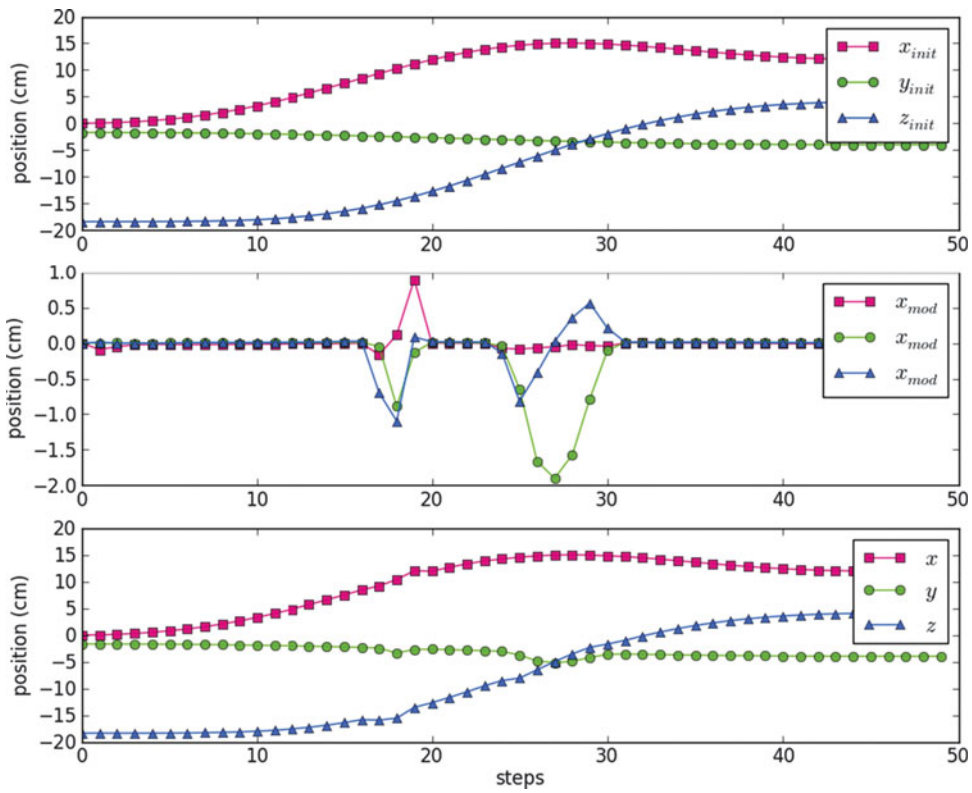


Fig. 4 Optimized trajectory with STOMP.

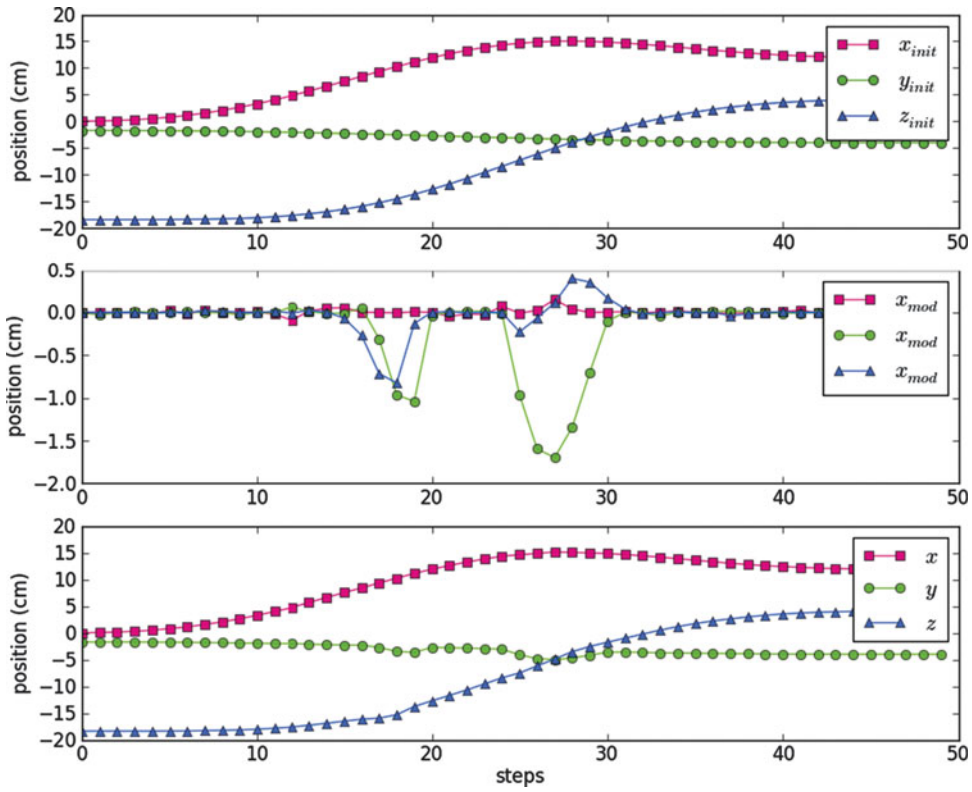


Fig. 5 Optimized trajectory with TOP.

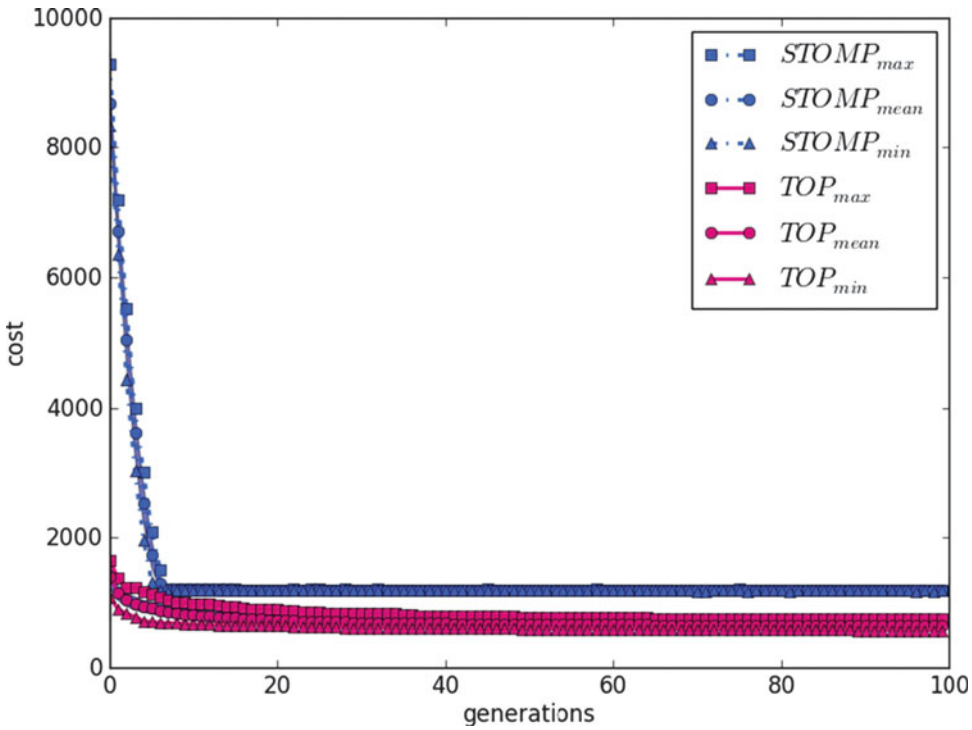


Fig. 6 Cost value for each generation.

References

1. S. M. LaValle: Planning algorithms, Cambridge University Press, 2006.
2. J. Kennedy, R. Eberhart: Particle swarm optimization, Proc. of the IEEE Int. Conf. on Neural Networks, vol. 4, pp. 1942–1948 (1995)
3. R. Krohling: Gaussian swarm: a novel particle swarm optimization algorithm, IEEE Conf. on Cyber. and Int. Sys., vol. 1, pp. 372–376 (2004)
4. E. A. Theodorou, J. Buchli and S. Schaal "A Generalized Path Integral Control Approach to Reinforcement Learning, Journal of Machine Learning Research, vol. 11, pp. 3137–3181 (2010)
5. E. A. Theodorou, J. Buchli and S. Schaal "Reinforcement Learning of Motor Skills in High Dimensions: A Path Integral Approach, Proc. of the IEEE Conf. on Robotics and Automation, pp. 2397–2403 (2010)
6. O. Khatib: Real-time obstacle avoidance for manipulators and mobile robots, Int. Journal of Robo. Res., vol. 5, no. 1, pp. 90–98 (1986)
7. C.W. Warren: Global path planning using artificial potential fields, Proc. of the IEEE Conf. on Robotics and Automation, pp. 316–321 (1989)
8. S. Quinlan and O. Khatib. "Elastic bands: Connecting path planning and control, Proc. of the IEEE Conf. on Robotics and Automation, pp. 802–807 (1993)
9. N. Ratliff, M. Zucker, J.A. Bagnell, and S. Srinivasa: CHOMP: Gradient optimization techniques for efficient motion planning, Proc. of the IEEE Conf. on Robotics and Automation, pp. 489–494 (2009)
10. M. Zucker, N. Ratliff, A.D. Dragan, M. Pivtoraiko, M. Klingensmith, C.M. Dellin, J.A. Bagnell, and S.S. Srinivasa: CHOMP: Covariant hamiltonian optimization for motion planning, Int. Journal of Robo. Res., vol. 32, no. 9–10, pp. 1164–1193 (2013)
11. M. Kalakrishnan, S. Chitta, E. Theodorou, P. Pastor, and S. Schaal: STOMP: Stochastic trajectory optimization for motion planning, Proc. of the IEEE Conf. on Robotics and Automation, pp. 4569–4574 (2011)
12. T. Flash, N. Hogan: The coordination of arm movements: an experimentally confirmed mathematical model, The Journal of Neuroscience, vol. 5, no. 7, pp. 1688-1703 (1985)

Cost model for an integrated load carrier design process in the lithium-ion battery production

Achim Kampker, Christoph Deutskens, Heiner Hans Heimes, Mathias Ordung, and Andreas Haunreiter

1 Introduction

Due to the finite nature of fossil energy, car manufacturers focus more and more on developing electric fueled vehicles. As traction batteries are a key component in an electric drive train, a growing demand of Lithium-Ion battery cells can be observed.[1] Nevertheless, the traction battery has a share of 60 % in production costs and 40 % in total costs, respectively.[2]

To satisfy the growing demand for battery cells, an automated production is required, although several issues have to be resolved to accomplish this. Open points in terms of quality and safety among others exist, e.g. investigated by [3] and [4]. Another key issue is the specialization of suppliers. An internal study by the Laboratory of Machine Tools and Production Engineering identifies that among more than 200 German and international suppliers of production equipment and automation technology for batteries only a few possess the competences for more than one process step. Subsequently, battery manufacturing companies need to align equipment from many different suppliers by adjusting interfaces between process steps.[5] One of the enablers to increase sales is reducing the battery price and this means for manufacturers to be under pressure with regard to costs. A potential to raise the scale of efficiency and minimize costs offers production logistics. As certain production processes are time consuming, a significant number of load carriers is needed to store and transport the goods in process. Hence, load carriers are a remarkable cost factor [6] and in general, 70 % of the costs of a product are determined in the development phase [7,8]. Consequently, the definition of a standardized process of developing the logistics process and load carriers is desirable to identify possible proficiency to reduce costs at an early stage.

This paper introduces a methodology to approach a standardized load carrier development process by extending the integrated product and process development. In the course of this, a model is needed to evaluate costs in an early stage. A framework for such a cost model, which has to contain all relevant cost factors and attain a practical feasibility, is established in this article.

2 Background

As market demands have shifted to more individual products, manufacturers have to satisfy customers by a greater product variety. In consequence, production tends to a lower volume for single product types.[9] Together with short product life cycles, a broad variety leads to the necessity of more ramp-ups in short periods. Due to these challenges, a reevaluation of product and process development has taken place and new approaches have been developed. A common approach is the aggregation to the integrated product and process development, called IPPD in the following.[10] Whereas the order of product and process engineering was chronological formerly, their deployment is overlapping time wise in the IPPD.

Figure 1 visualizes the general framework of the IPPD. It is shown that the majority of the two overlapping processes happens simultaneously, but they do not have to begin and end at the same time. Prior to the IPPD, the product definition is determined by combining technology planning and market demands. This approach ensures that developers consider all necessary elements in the product life cycle, including conception, user requirements, quality, costs and disposal.[13] This system of concurrent engineering, which is a synonym to the IPPD, accomplishes a shorter time span between starting the product development and a possible start of production (SOP). One of its concerns is to make all required information accessible to carry out development tasks.[9] The results are improved quality, improved time to market and reduced redesign times.[14] The logistics as a

A. Kampker (✉) • C. Deutskens • H.H. Heimes
M. Ordung • A. Haunreiter
RWTH Aachen, Laboratory for machine tools and production engineering, Steinbachstraße 19, 52074 Aachen, Germany

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_46,

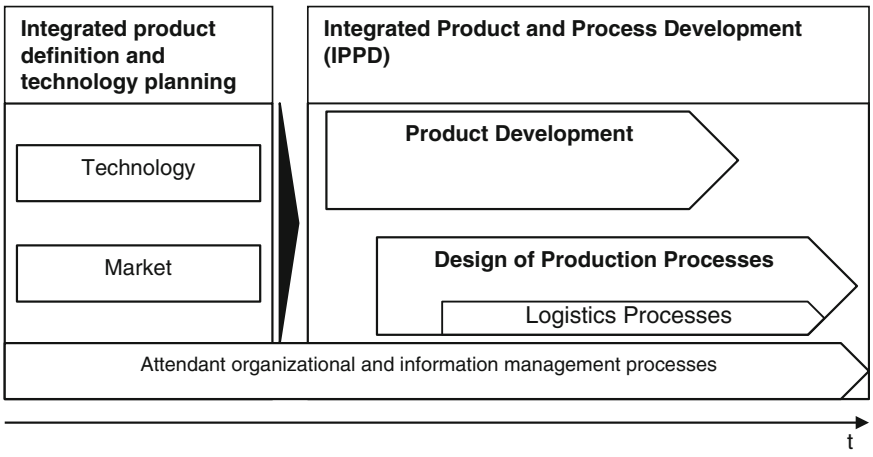


Fig. 1 Framework of Integrated product and process development [11,12]

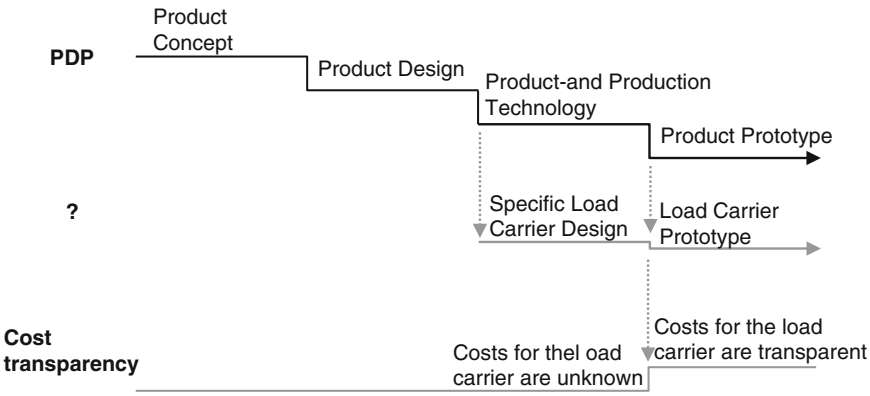


Fig. 2 Typical phases of the load carrier development aligned with the battery development

system often is understood as a subsystem of the production process. This is visualized in figure 1, where the production process design also contains determining logistics processes.

3 Purpose

In chapter 2 it was pointed out that the IPPD serves to provide a faster ramp-up. But not only is the time perspective of interest in this context. As Kampker debunks simultaneous engineering makes potentials in terms of cost reduction and adhering to target costs accessible. Furthermore, cost innovations, which lead to massive cost savings, are a possible consequence of the IPPD.[15]

Load carriers consume a remarkable amount of invest and have a profound influence on the production efficiency. A significant cost factor is the low level of standardization in production equipment of different process steps, which causes the production of numerous different load carrier types for one identical product design.[7] Figure 2 shows the typical phases of the battery production development and the late integration of the logistics and load carrier design. Whereas the product development process (PDP) starts with the product concept, which is followed by product design, the development of a load carrier starts parallel to the development of product- and production technology.

Not only does the load carrier development start late, but also a standardized development process for the load carrier does not exist. The lack of feedback from a simultaneous load carrier development to the battery development has a negative impact on costs. Furthermore, there is no cost transparency, which helps to identify and exploit potentials in cost reduction.

Overall, various heterogeneous components and processes have to be coordinated to achieve a complete functionality as a system.[15] There is the task of outlining the logistics system, its implications on the transporting elements and the understanding of cost drivers. Therefore, the risk is high that with a late start of load carrier development problems may occur, which affect the planned volume increase to series production negatively. In case of such a delay, the objective of matching target costs is at risk. Considering this fact, this stresses a necessity for such a standard.

4 Methodology

This chapter defines a methodic approach to standardize and align load carrier development with the actual product development process. In passage 4.1, a general framework with regard to the IPPD (cf. chapter 2) is emphasized. This is followed by the outline of a cost model to increase cost transparency (see chapter 4.2).

4.1 Introducing a load carrier development process

Within the IPPD presented in chapter 2, logistics processes are part of the production process development. In chapter 3, the problematization showed that these processes concerning logistics and especially the development of load carriers starts rather late with risks to affect the SOP negatively.

To minimize those risks, the development of load carriers (LDP) needs to be integrated into the IPPD as a third category starting simultaneously with product development. This is shown in figure 3 below. The establishment of an LDP provides key advantage. Figure 2 showed that load carrier design only used little amount of time compared with the PDP.

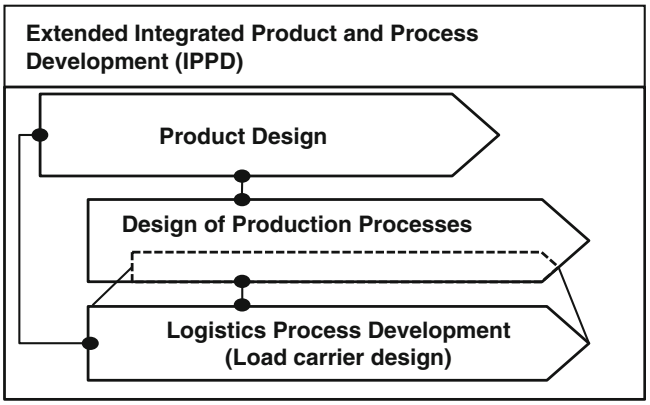


Fig. 3 Extending the IPPD by a logistics perspective

With the new alignment the conception of load carriers begins earlier. This means that with the applicable instruments not only risks can be minimized, but also the most cost-effective solution can be exposed and pursued. Cost effectiveness relates to two categories here. By modularizing load carriers, invest is kept low. Secondly, planning the load carrier system at such an early stage makes it possible to influence production efficiency with a specific load carrier design. For example, a load carrier that allows the product to rest in it during a process step, stints an additional handling step.

According to Christopher, costs are a measure for logistics' efficiency, which profit and related to this the return on investment.[16] To accomplish the increase of cost effectiveness, it is necessary to have a reliable cost model, which allows calculating costs, caused by the load carrier system. Such a model secures a high level of cost transparency. The framework of such a cost model is detailed in chapter 5.

In terms of aligning the PDP and LDP, figure 4 exposes a detailed view on the different stages and their synchronization. The first phase of the PDP covers the entire product concept development. It starts with an initial strategic phase that defines a scope for the battery system. During this phase, the logistics strategy can also be determined. Knowledge about in- and outsourcing of process steps, the structure of the production network and the location of customers allow answering the first questions about the requirements for external transport. This is the basis for the first step in the load carrier development process (LDP), which starts with an outlining of the logistics strategy itself. This determines initial decisions about the usage of load carriers not only in the production itself, but also for transfers between productions, i.e. as elements of larger transportation devices. In this phase, the cost transparency is rather low, compared to the estimates for product costs, which can usually be

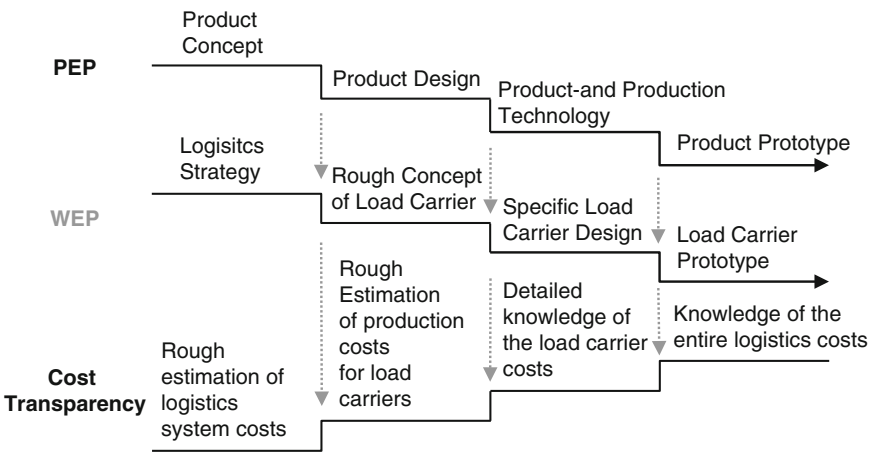


Fig. 4 Product and load carrier development processes aligned

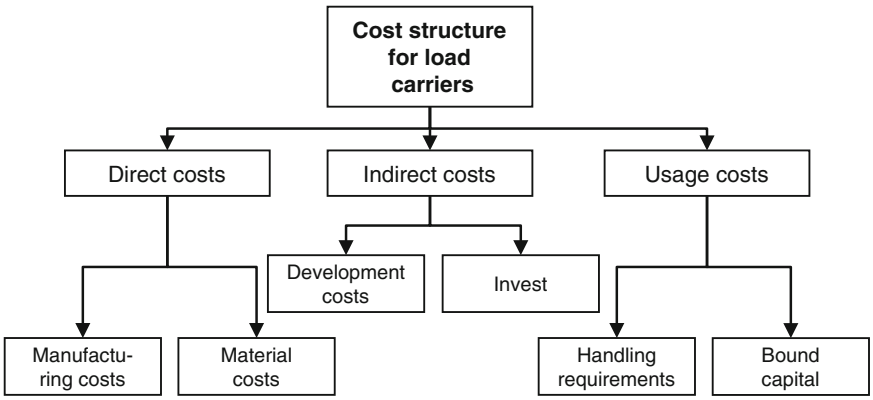


Fig. 5 Cost structure for the load carrier cost model

derived from earlier versions of the product. Since battery generations often evolve instead of taking revolutionary steps, cost assumptions can be quite precise even in the initial phase. The cost model in the LDP already allows rough estimations about the costs for the entire logistics system, assuming costs from earlier generations, which are being set in context to the expected changes in cost-relevant design features.

The second phase in the product development process allows estimations about the actual design of the battery itself as it advances more into detail. This usually includes decisions such as the basic geometrical properties, the number of cells per module and pack, respectively, as well as an outline of total production capacities per annum basing upon strategic decisions during the first phase. The decision, whether the cell has a solid outer structure or a pouch cell, determinates which process steps can be performed right in the load carrier and where handling is needed otherwise.

Different phases within the LDP are supported additionally by the use of different requirement categories, which will need to be defined. These categories range from geometrical to maintenance issues and depend on the degree of maturity of product development. Ergo, there must be an alignment to consider the right requirements at the right time. The result of a successful is a load carrier system, which is as standardized as possible, but as individualized as necessary, serves two objectives. On the one hand, it minimizes invest for the load carrier and, on the other hand, a load carrier system can maximize production efficiency. This equals a maximum in cost reduction.

4.2 Focusing on a load carrier cost model within the LDP

According to figure 4, cost transparency rises with the advance of PDP and LDP synchronically, starting with a rough estimation and finishing with a complete cost transparency. Reducing the production costs as a major target in the product development requires the knowledge about the type and significance of different cost influences in the early stages of the IPPD.[7] In this context, the concept of process cost management is introduced as the basis for further research related to costs. Principles about cost models in production and logistics can be found in [8,17,18]. Here, the focus lies on the specific cost structure for load carriers in battery production.

In this case the load carrier cost model uses a clear cost structure (see figure 5), which differentiates between direct and indirect costs at first. Direct costs can be divided into manufacturing and material costs. Indirect costs result from costs for developing the load carrier and the necessary investments for the setup of a production line. Additionally to direct and indirect costs, the cost model also includes the implications on the use of the load carrier. This allows the setup of a business case to show, if there are raises in costs due to the lack of certain design features.

The cost model bases on the early awareness of the requirements that are being addressed to the load carrier during the various stages of the product development process. These requirements can be clustered into ten categories, which cover 83 requirements in total.

Figure 6 shows the ten requirement clusters. The cost model itself not only considers cost estimation, but also indicates the level of data maturity. The model uses a standardized set of typical load carrier requirements and their dimensions, but this set can be changed by manual data input as well. The result is a cost overview that shows the cost elements in respect to their maturity levels. The level of cost transparency in percent is an indicator, how plausible a result appears, and can be derived from an essential formula:

$$Cost\ transparency\ [\%] = \frac{\sum_{x=1}^{n}(Cost\ impact_x\ [€] * \ data\ maturity\ level_x)}{Number\ of\ maturity\ levels * 100} \quad (1)$$

In the formula, a data maturity level is a number on a scale. This scale can range from 1 to 3, for example, with 1 for low

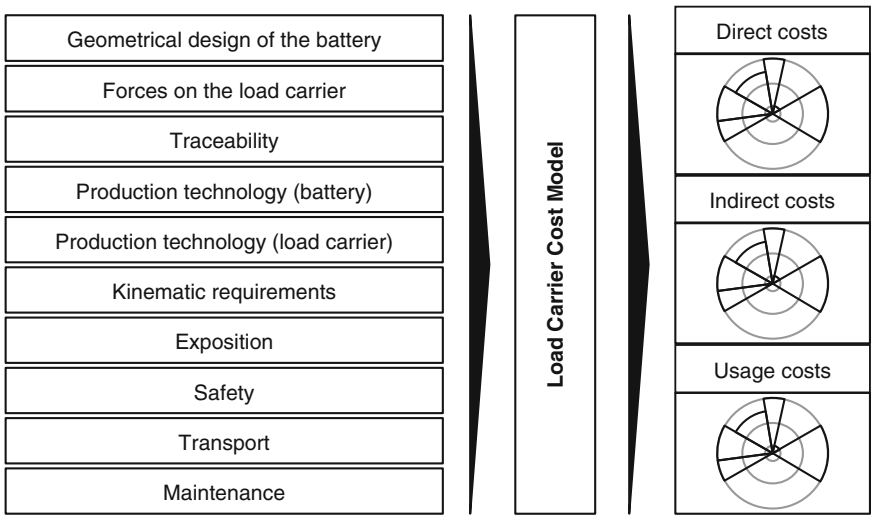


Fig. 6 Input categories for the load carrier cost model

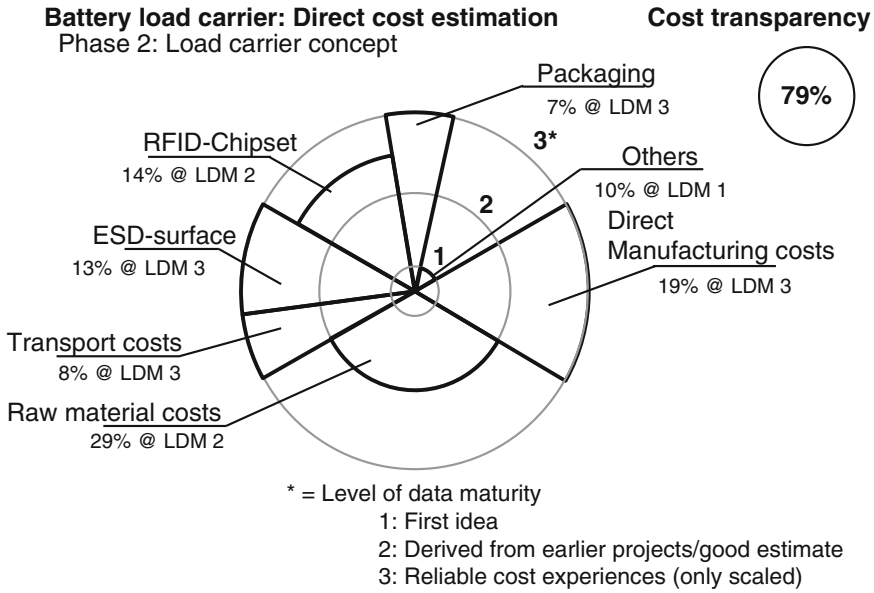


Fig. 7 Example of the cost model output for phase 2 in the LDP

data maturity. The division by the number of maturity levels on the scale and by 100 provides a relative indicator to track the development of cost transparency.

The result of the calculation is a chart, that allows the identification of cost drivers, the relations between the costs and an initial attention on the fields with low data maturity, allowing intervention if necessary. Not only can these resulting plots be obtained for direct, but also for indirect costs and the costs, that derive from the use of the load carrier in the logistics system.

Figure 7 shows an example output of the load carrier cost model, using early estimates from the product and load carrier concept phase. Thus the cost data maturity remains rather low in terms of raw material costs and unknown factors – 'others' with a 10 % placeholder that has been proven a valid assumption from recent projects. With these plots, the developer is able, to maintain transparency and prioritize the development efforts to the largest impacting cost factors.

5 Validation

An example for the benefit of an LDP with a high level of cost transparency is the position of electrodes of pouch cells, which is important during the formation process. The formation process is the first loading of the battery cell.[19] For this, the cell rests in its load carrier and spring-supported contact electrodes approach from above to establish an electric connection to the batteries' electrodes. In order to achieve the connection, the contact electrodes apply pressure to get through the outer oxide layer. In specifications for

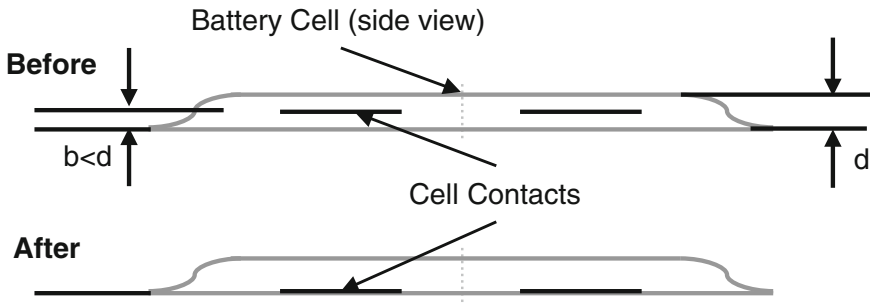


Fig. 8 Adjusted positions of the cell contacts due to load carrier influences

battery development, there do not exist any requirements concerning the position of electrodes. In consequence, a position between lower and upper edge is possible, which is illustrated in figure 8 in the first picture. Nevertheless, a support for the electrodes is necessary to endure the pressure in the formation process.

In the known LDP (cf. chapter 2), the solution for this problem would be a load carrier, which has got a heel to support the electrodes. A change in the load carrier design, to provide an elevated heel element as support, causes increased material costs and lets commonality decrease. Commonality of load carriers allows the usage of a load carrier model for multiple battery sizes differing in size of cell contacts.

In the extended IPPD (cf. chapter 4.1), a cost optimum is reached by positioning cell electrodes at the lower edge of the cell. This is shown in the bottom illustration of figure 8. This still matches the battery's specifications and does not cause any additional costs in production. For the load carrier it means less specific design, which allows a higher modularization and more scale effects. With the cost model outlined in chapter 4.2 it is possible to evaluate the two scenarios from figure 8 and to pursue the least cost intensive scenario.

6 Summary

As cost reduction is a major objective of the extended IPPD, a cost model was developed in this paper, which serves a high level of cost transparency. Moreover, a catalog of requirements is the basis of calculating costs for possible scenarios to put a load carrier system into practice. The cost model uses those requirements and integrates their impact on specific cost situations into the calculation. An overview about the cost situation of the current stage of development can be provided with the given information. Subsequently, it was shown, how the alignment of the LPD with the PDP affects the potential of cost reduction positively. An early consideration of the effect of certain product attributes on a load carrier design makes these potentials accessible. The next steps for further research include detailing the general framework of the extended IPPD and defining different stages of maturity. Furthermore, the requirements need a value, which is approved by experts and serves as a standard value for the cost model. Once this has been accomplished, the cost model needs to receive a higher level of usability. Hence, it must be converted into a tool, which can be modified in a simple way by the user.

Acknowledgement The approach has been designed by the Chair of Production Engineering for E-Mobility Components (PEM) of RWTH Aachen University within the publicly funded research projects LAKOBAT (funded by the ZIEL2-programme) and E-PRODUCTION (funded by the Federal Ministry of Education and Research (BMBF)).

Reference

1. B. Mayer, Lithium-ion race pics up, Autom. News Europe, Vol. 13, June 2008.
2. BMU, Konzept eines Programms zur Markteinführung von Elektrofahrzeugen, Berlin, September 2009, p.6.
3. M. Westermeier, G. Reinhart, T. Zeilinger, 2013, Method for quality parameter identification and classification in battery cell production, Electric Drives Prod. Conf. (EDPC), Nuremberg.
4. R. Ranzinger, 2013, Design of safe assembly processes for live working in traction battery series production, Electric Drives Prod. Conf. (EDPC), Nuremberg.
5. A. Kampker, H. Heimes, C. Sesterheim, M. Schmidt, 2013, Conception of Technology Chains in Battery Production, In: V Prabhu, M Taisch, D Kiritsis (Editors), Advances in Production Management Systems. Sustainable Production and Service Supply Chains, IFIP Advances in Information and Communication Technology, Vol. 414, 2013, pp.199-209.
6. J. Oser, 2007, Integrated Unit Load and Transport System Design in Manufacturing, In: IET Int. Conf. on Agile Manufacturing, ICAM 2007, Durham, pp. 119-123.
7. G. Pahl, W. Beitz., 2007, Engineering Design, 3. Edition, Springer London Ltd.
8. H. Warnecke, H.-J. Bullinger, R. Hichert, A. Voegele, 1996, Kostenrechnung für Ingenieure, 5. Auflage, Hanser-Verlag, München, pp.232-238.
9. F. Hull, P. Collins, J. Liker, 1996, Composlite Forms of Organization as a Strategy for Concurrent Engineering Effectiveness, IEEE Transactions on engineering management, Vol. 43, No. 2
10. J. Usher, U. Roy, H. Parsaei (Editors), 1998, Integrated Product and Process Development: Methods, Tools, and Technologies, John Wiley & Sons, p. IX

11. G. Schuh, W. Stölzle, F. Straube (Editors), 2008, Anlaufmanagement in der Automobilindustrie erfolgreich umsetzen, Springer, pp.1-5
12. T. Waggoner, 1995, Concurrent engineering strategies in electrical component manufacturing, Electrical Electronics Insulation Conference, 1995, and Electrical Manufacturing & Coil Winding Conference. Proceedings.
13. R. Winner, J. Pennell, H. Bertrand, M. Slusarczuk, 1988, The role of concurrent engineering in weapons system acquisition, Institute for Defense Analyses, Virginia.
14. A. Yassine et al., 1999, A Decision Analytic Framework for Evaluating Concurrent Engineering, IEEE Trans. On Eng. Management, Vol. 46, No. 2.
15. P. Thomes et al., Grundlagen, In: A Kampker, D Vallée, A Schnettler (Editors), 2013, Elektromobilität, Springer-Verlag, pp.5-58
16. M. Christopher, 2012, Logistics and Supply Chain Management, fourth edition, Prentice Hall, p.85
17. D. Ellström, J. Rehme, M. Björklund, H. Aronsson, 2012, Logistics Cost Management Models and Their Usability for Purchasing, Journal of Modern Accounting and Auditing, Vol. 8, No. 7.
18. Z. Bokor, 2012, Integrating Logistics Cost Calculation into Production Costing, Acta Polytechnica Hungarica, Vol. 9, No. 3.
19. A. Kampker et al., 2012, Process Alternatives in the Battery Production, Electrical Systems for Aircraft, Railway and Ship Propulsion (ESARS), 16-18 Oct. 2012.

Sensorless Force Estimation for a Two-Link Manipulator Based Upon Linear Dynamics

Douglas R. Isenberg

1 Introduction

The need for robotic manipulators that are capable of not only positional control but also force control has been recognized for decades [1]. Typically, forces are measured via a multi-axis force/torque sensor mounted on the end-effector of a manipulator. However, there are a variety of other techniques that have been developed to measure such forces [2]. In addition to the techniques presented in [2], model-based force-sensorless techniques resulting in a disturbance observer formulation have also been developed [3]-[6]. In [4], a nonlinear observer is developed to estimate friction forces. A Kalman filter implementation of a disturbance observer is applied to a linear motor in [5]. A nonlinear disturbance observer is designed in [6] for estimating end-effector forces on a manipulator utilized for a friction stir welding process.

In this paper, it is recognized that a two-link planer manipulator's equations of motion can be decomposed into a linear and nonlinear part. The linear part of the dynamics, as evident from the results of nonlinear system identification of the manipulator, are dominant. These linear dynamics, with the addition of the identified static friction terms, are utilized to design a discrete-time unknown input observer to estimate joint torques that arise from forces acting on the end-effector of the manipulator. The resulting observer makes use of only the two measured joint angles and the two motor armature duty-cycles.

The paper is organized in the following manner. First, the equations of motion for the composite two-link planer manipulator/actuator are described. Then, a force observer based upon the linear part of the equations of motion is developed. Numerical simulations of the proposed observer are then presented. This is followed by a presentation of experimental results of the observer applied to the planer two-link manipulator depicted in Fig. 1.

2 Manipulator Dynamics

The dynamics of a planer two-link manipulator can be expressed as

$$H(\gamma)\ddot{\gamma} + D(\gamma,\dot{\gamma}) - \tau_e = \tau_M \tag{1}$$

where $\gamma = [\theta_1, \theta_2]^T$ are the joint angles of the manipulator, $\tau_M \in \mathbb{R}^2$ are the torques applied by the motors, and $\tau_e \in \mathbb{R}^2$ are the joint torques due to force $F_e \in \mathbb{R}^2$ acting on the end-effector. The system mass matrix is

$$H(\gamma) = \begin{bmatrix} m_2 d_{y_1}^2 + 2 \cdot {}^2_2\Gamma_y d_{y_1} cos(\theta_2) + {}^2_2 J_{zz} + {}^1_1 J_{zz} & {}^2_2 J_{zz} + {}^2_2\Gamma_y d_{y_1} cos(\theta_2) \\ {}^2_2 J_{zz} + {}^2_2\Gamma_y d_y cos(\theta_2) & {}^2_2 J_{zz} \end{bmatrix} \tag{2}$$

D.R. Isenberg (✉)
Department of Aerospace and Mechanical Engineering, Embry-Riddle Aeronautical University, Prescott, AZ 86301, USA
e-mail: isenberd@erau.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_47,

and the vector of coupling torques is

$$D(\gamma,\dot{\gamma}) = \begin{bmatrix} -2\cdot{}_2^2\Gamma_y d_{y_1}\,\dot{\theta}_1\,\dot{\theta}_2 sin(\theta_2) - {}_2^2\Gamma_y d_{y_1}\dot{\theta}_2^{\,2} sin(\theta_2) + \alpha_1\,\dot{\theta}_1 + \alpha_2 sgn(\dot{\theta}_1) \\ {}_2^2\Gamma_y d_{y_1}\dot{\theta}_1^{\,2} sin(\theta_2) + \alpha_3\,\dot{\theta}_2 + \alpha_4 sgn(\dot{\theta}_2) \end{bmatrix}. \tag{3}$$

Here, $d_{y_1} = 0.178 \quad (m)$ is the is the distance between joint 1 and joint 2 along the y-axis of frame 1, ${}_1^1J_{zz}$ is the moment of inertia of link 1 about the z-axis of frame 1 measured with respect to frame 1, m_2 is the mass of link 2, ${}_2^2\Gamma_y$ is the first mass moment of link 2 along the y-axis of frame 2 measured with respect to frame 2, ${}_2^2J_{zz}$ is the moment of inertia of link 2 about the z-axis of frame 2 measured with respect to frame 2, and α_{1-4} are coefficients of viscous and static friction.

The relationship between τ_e and ${}^IF_e$, the force applied to the end-effector as measured in the inertial frame, can be found through a virtual work balance, $(\delta\gamma_e)^T\tau_e = (\delta r_e)^{T\,I}F_e$, where r_e is the forward kinematic expression for the location of the end-effector,

$$r_e = \begin{bmatrix} -d_{y_1} sin(\theta_1) - d_{y_2} sin(\theta_1+\theta_2) \\ d_{y_1} cos(\theta_1) + d_{y_2} cos(\theta_1+\theta_2) \end{bmatrix}. \tag{4}$$

Here, $d_{y_2} = 0.229 \quad (m)$ is the distance along the y-axis of frame 2 to the end-effector. Thus, the virtual displacement, δr_e, is

$$\delta r_e = \begin{bmatrix} -d_{y_1} cos(\theta_1) - d_{y_2} cos(\theta_1+\theta_2) & -d_{y_2} cos(\theta_1+\theta_2) \\ d_{y_1} sin(\theta_1) + d_{y_2} sin(\theta_1+\theta_2) & d_{y_2} sin(\theta_1+\theta_2) \end{bmatrix} \delta\gamma. \tag{5}$$

τ_e is therefore

$$\begin{aligned} \tau_e &= \begin{bmatrix} -d_{y_1} cos(\theta_1) - d_{y_2} cos(\theta_1+\theta_2) & d_{y_1} sin(\theta_1) + d_{y_2} sin(\theta_1+\theta_2) \\ -d_{y_2} cos(\theta_1+\theta_2) & d_{y_2} sin(\theta_1+\theta_2) \end{bmatrix} {}^IF_e \\ &= \Omega(\gamma)^I F_e. \end{aligned} \tag{6}$$

The force on the end-effector is typically measured with force-sensors fixed to the end-effector, so ${}^EF_e$ is the actual measured quantity. Thus, $\tau_e = \Omega(\gamma)^I T_E {}^EF_e$, so

$$\tau_e = \begin{bmatrix} -d_{y_2} - d_{y_1} cos(\theta_2) & d_{y_1} sin(\theta_2) \\ -d_{y_2} & 0 \end{bmatrix} {}^EF_e, \tag{7}$$

where ${}^IT_E$ is the coordinate transformation from frame E to frame I.

Fig. 1 *Two-link planer manipulator.*

The links of the manipulator are actuated by brushed DC motors with an output to input gear ratio, N. Thus,

$$Ki_a = N\tau_M + J_a N^{-1}\ddot{\gamma} + \beta N^{-1}\dot{\gamma} + Ssgn(\dot{\gamma}) \tag{8a}$$

$$V_a = R_a i_a + KN^{-1}\dot{\gamma}. \tag{8b}$$

Here, $K \in \mathbb{R}^{2\times2}$ is a diagonal matrix of the motors' back e.m.f. / torque constants, $i_a \in \mathbb{R}^2$ are the motors' armature currents, $J_a \in \mathbb{R}^{2\times2}$ is a diagonal matrix of the armatures' moment of inertia, $N \in \mathbb{R}^{2\times2}$ is a diagonal matrix of the gear ratios, $\beta \in \mathbb{R}^{2\times2}$ is a diagonal matrix of the motors' viscous friction coefficients, $S \in \mathbb{R}^{2\times2}$ is a diagonal matrix of the motors' static friction coefficients, $V_a \in \mathbb{R}^2$ are the armatures' voltage pulse-width modulation duty cycles, and $R_a \in \mathbb{R}^{2\times2}$

is a diagonal matrix of the armatures' resistances. It is assumed that the armatures' inductances are negligible.

From (1), (8a), and (8b), the composite manipulator/motor equations of motion are

$$V_a = H_c(\gamma)\ddot{\gamma} + D_c(\gamma,\dot{\gamma}) - \Omega_c(\gamma)^E F_e \tag{9}$$

where

$$H_c(\gamma) = \begin{bmatrix} b_1 + 2b_2 cos(\theta_2) & b_3 + b_2 cos(\theta_2) \\ b_3 + b_2 cos(\theta_2) & b_4 \end{bmatrix} \tag{10a}$$

$$D_c(\gamma,\dot{\gamma}) = \begin{bmatrix} -2b_2\dot{\theta}_1\dot{\theta}_2 sin(\theta_2) - b_2\dot{\theta_2}^2 sin(\theta_2) + b_5\dot{\theta}_1 + b_6 sgn(\dot{\theta}_1) \\ b_2\dot{\theta_1}^2 sin(\theta_2) + b_7\dot{\theta}_2 + b_8 sgn(\dot{\theta}_2) \end{bmatrix} \tag{10b}$$

$$\Omega_c(\gamma) = \begin{bmatrix} -b_9 d_{y_2} - b_9 d_{y_1} cos(\theta_2) & b_9 d_{y_1} sin(\theta_2) \\ -b_{10} d_{y_2} & 0 \end{bmatrix}. \tag{10c}$$

The constants, b_{1-10} in (10), which are functions of the manipulator and motor parameters, are obtained experimentally via system identification for the planer two-link manipulator depicted in Fig. 1 as

$$b = [6.8768, 1.3476, 0.4412, 1.9434, 27.9565, 3.9869, 15.8666, 1.6459, 489, 368]^T. \tag{11}$$

3 Force Observer

The most straightforward method of estimating ${}^E F_e$ is to measure V_a and γ, differentiate γ twice, substitute the values into (9), and solve for ${}^E F_e$. However, this method has a few significant drawbacks. Differentiating the joint angles, typically with a finite-difference, amplifies noise, and although the first derivative may be acceptable, the noise in the second derivative will generally be too large. To overcome this problem, low-pass filtering of the finite-differences can be utilized [7]. Due to the low bandwidth of the system, however, the necessary low-pass filters will introduce a significant time-delay between the application of the force and the estimation of the force. Hence, this method is not entirely well suited for applications requiring even moderate bandwidth. An alternative approach to this open-loop estimation method is to utilize a force observer.

The composite manipulator / motor dynamics, (9), can be decomposed into the sum of the linear dynamics and the nonlinear dynamics. The linear part of the dynamics plus the static friction terms is:

$$V_a = \begin{bmatrix} b_1 & b_3 \\ b_3 & b_4 \end{bmatrix}\ddot{\gamma} + \begin{bmatrix} b_5 & 0 \\ 0 & b_8 \end{bmatrix}\dot{\gamma} - \begin{bmatrix} b_9 & 0 \\ 0 & b_{10} \end{bmatrix}\tau_e + \begin{bmatrix} b_6 & 0 \\ 0 & b_8 \end{bmatrix} sgn(\dot{\gamma}). \tag{12}$$

Assuming that $\dot{\tau}_e \approx 0$, a state-space description for the system is thus,

$$\begin{bmatrix} \dot{\gamma} \\ \ddot{\gamma} \\ \dot{\tau}_e \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{-b_4 b_5}{a} & \frac{b_3 b_7}{a} & \frac{b_4 b_9}{a} & \frac{-b_3 b_{10}}{a} \\ 0 & 0 & \frac{b_3 b_5}{a} & \frac{-b1 b_7}{a} & \frac{-b_3 b_9}{a} & \frac{b_1 b_{10}}{a} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \gamma \\ \dot{\gamma} \\ \tau_e \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{b_4}{a} & \frac{-b_3}{a} & \frac{-b_4 b_6}{a} & \frac{b_3 b_8}{a} \\ \frac{-b_3}{a} & \frac{b_1}{a} & \frac{b_3 b_6}{a} & \frac{-b_1 b_8}{a} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} V_a \\ sgn(\dot{\gamma}) \end{bmatrix} \tag{13}$$
$$= Ax + BV_a,$$

where $a = b_1 b_4 - b_3 b_3$ is the determinate of the linear part of the system mass matrix. Additionally, since incremental encoders are utlized to measure the angular displacement of the links, the measurement equation for the system is described as

$$y = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \gamma \\ \dot{\gamma} \\ \tau_e \end{bmatrix} = Cx. \tag{14}$$

It is simple to verify that the system states, x, are observable from the output vector, y. Furthermore, since the system is linear and time-invariant, a closed form expression for the sampled-data system can be found. Utilizing the parameter vector, (11), and a sampling time of $T_s = 0.02(s)$, the sampled-data system dynamics are described as

$$\begin{aligned}\dot{x}[k+1] &= \begin{bmatrix} 1 & 0 & 0.0192 & 9.793\cdot 10^{-5} & 0.01403 & -0.002271 \\ 0 & 1 & 1.725\cdot 10^{-4} & 0.01843 & -0.003018 & 0.03638 \\ 0 & 0 & 0.9209 & 0.009393 & 1.384 & -0.2179 \\ 0 & 0 & 0.01655 & 0.8474 & -0.2895 & 3.54 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} x[k] \\ &+ \begin{bmatrix} 2.87\cdot 10^{-5} & -6.172\cdot 10^{-6} & -1.144\cdot 10^{-4} & 1.016\cdot 10^{-5} \\ -6.172\cdot 10^{-6} & 9.886\cdot 10^{-5} & 2.461\cdot 10^{-5} & -1.627\cdot 10^{-4} \\ 0.00283 & -5.92\cdot 10^{-4} & -0.01128 & 9.744\cdot 10^{-4} \\ -5.92\cdot 10^{-4} & 0.009619 & 0.00236 & -0.01583 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} V_a[k] \\ sgn(\theta_1) \\ sgn(\theta_2) \end{bmatrix} \\ &= A_d x[k] + B_d u[k], \end{aligned} \tag{15}$$

where the input is subjected to a zero-order hold.

An unknown input linear observer [8] based upon this linear part of the manipulator's dynamics then takes on the form,

$$\begin{aligned}\hat{x}[k+1] = A_d\hat{x}[k] + B_d[V_a[k], sgn(\hat{x}_3[k]), \hat{s}gn(\hat{x}_4[k])]^T \\ + L(y[k] - C\hat{x}[k]).\end{aligned} \tag{16}$$

Here, estimates of the joint velocity are utilized for computing the static friction compensation terms.

There are a variety of methods that can be applied to select the gain matrix, L, in (16). For this work, L is selected based upon a linear quadratic regulator approach [9]. Through experimental tuning of weights, the observer gain is selected as

$$L = \begin{bmatrix} 0.602 & -0.01207 \\ -0.01165 & 0.7329 \\ 6.434 & -0.322 \\ -0.2954 & 9.905 \\ 0.7221 & -0.001771 \\ 0.01086 & 0.6747 \end{bmatrix} \tag{17}$$

The observer in (16) will provide estimates of τ_e. In order find the estimates of the forces acting on the end-effector, the inverse of (7) is solved.

$${}^{E}F_e = \begin{bmatrix} 0 & -\dfrac{1}{d_{y_2}} \\ \dfrac{1}{d_{y_1}\sin(\theta_2)} & -\dfrac{d_{y_2} + d_{y_1}\cos(\theta_2)}{d_{y_1} d_{y_2}\sin(\theta_2)} \end{bmatrix} \tau_e. \tag{18}$$

This transformation is only possible when $\theta_2 \neq \pm n\pi \quad for\ n = 0, 1, \cdots$ which correspond to the kinematic singularities for the manipulator.

4 Simulation

The observer is simulated by integrating (9) with a fourth-order Runge-Kutta method using an integration step size of 0.00005 seconds. At every 0.02 seconds within the integration loop, estimates of the joint position, velocity, and torque are computed with (16).

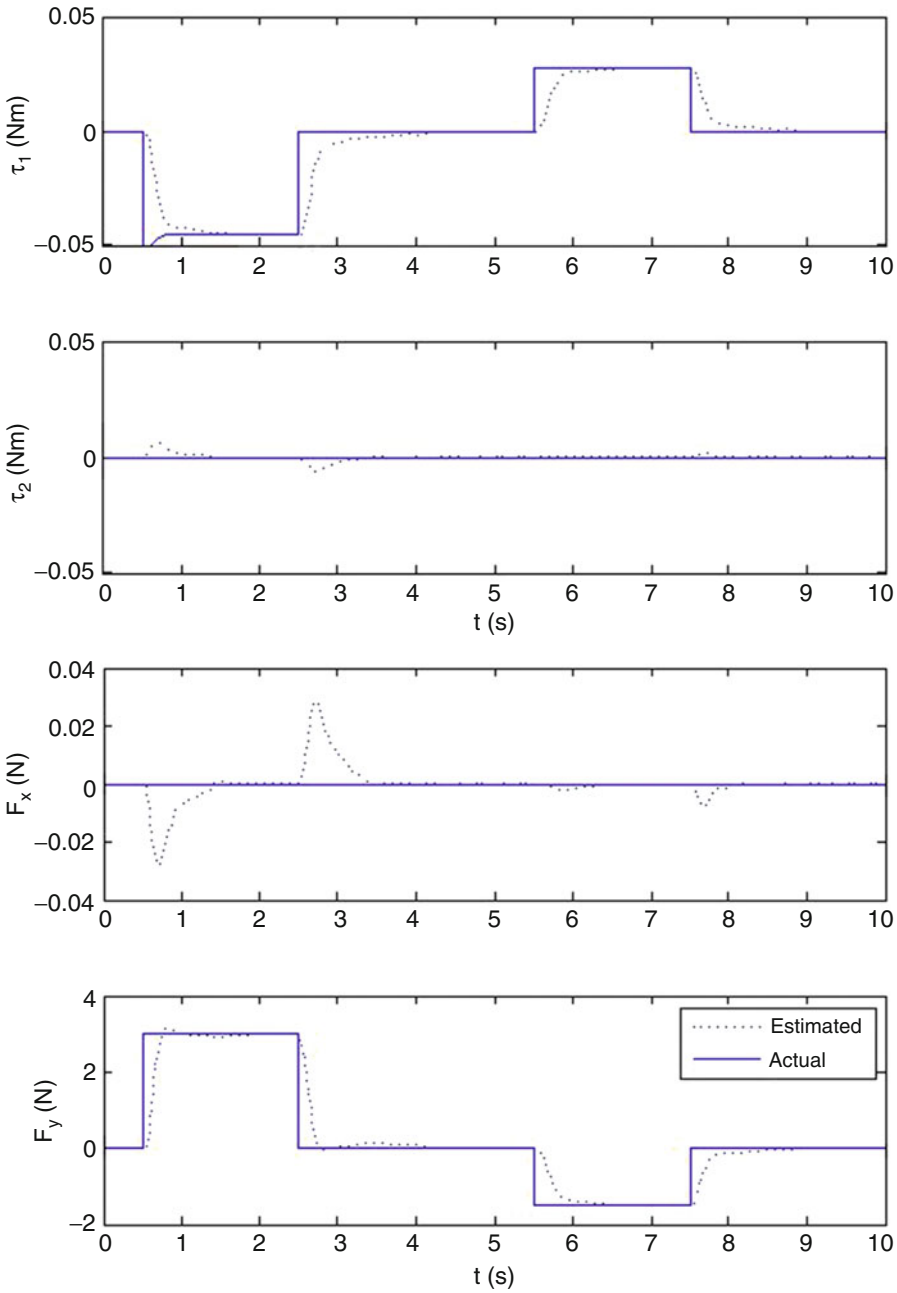


Fig. 2 *Estimation of force parallel to the y-axis of end-effector frame.*

Fig. 2 depicts the observer response to a force comprised of two pulses along the y-axis of the end-effector frame. For this simulation, $\theta_1(t_0) = 0\ (rad)$ and $\theta_2(t_0) = -0.1 \quad (rad)$. The first two plots depict the joint torques which are estimated directly. It is evident that there is some delay in the torque estimation, however the estimates do converges to the correct values. The bottom two plots depict the forces acting on the end-effector which are indirectly computed from the estimated joint torques utilizing (18). The estimated end-effector forces clearly converge to the actual end-effector forces.

Fig. 3 depicts the observer response to a force comprised of two pulses along the x-axis of the end-effector frame. For this simulation, $\theta_1(t_0) = 0\ (rad)$ and $\theta_2(t_0) = 3.\ 1\ (rad)$. The estimated joint torques in response to an x-axis force is not as accurate as the estimated joint torques in response to the y-axis force as evident by the relatively significant error in the estimated F_y. This is caused by the fact that a force along the x-axis of the end-effector's frame will result in substantial motion of θ_2 since F_x is normal to the radial displacement vector of link two and also the static and viscous friction coefficients for joint two are nearly half that of

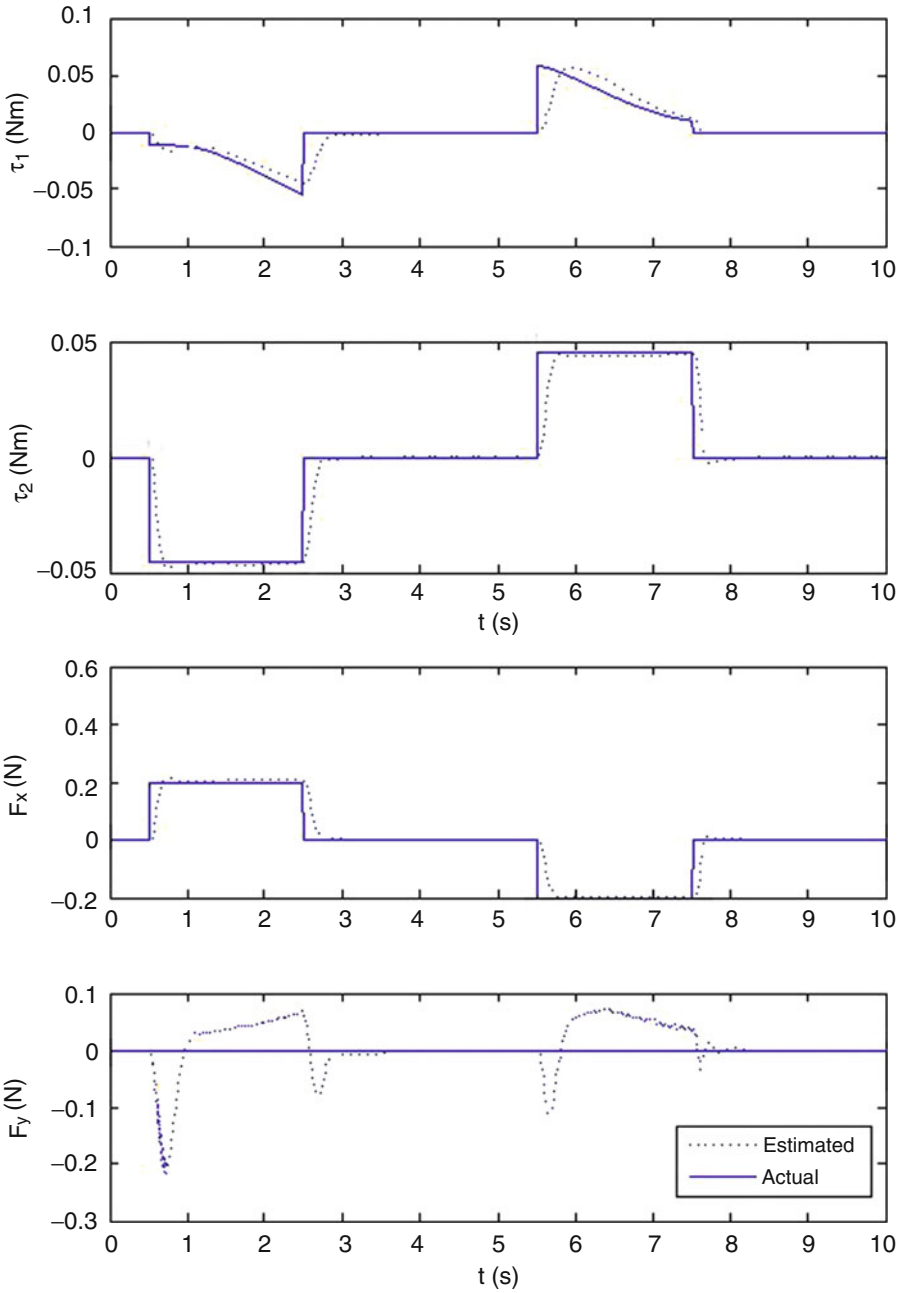


Fig. 3 *Estimation of force parallel to the x-axis of end-effector frame.*

joint one. Due to the periodic nature of the singularities in the joint space, such a constant force will always tend to drive the system towards a singular configuration whereupon the linear relation, (18), is ill-conditioned and ultimately singular.

5 Experimental Results

The observer is experimentally tested utilizing the two-link planer manipulator depicted in Fig. 1. The observer is implemented in Python on a desktop personal computer that communicates with the manipulator's embedded controller via an RS-232 connection. A two-axis force sensor comprised of four orthogonally situated load cells is utilized to measure the force that is provided to a circular disk located at the end-effector. This directly measured force in the end-effector's frame is utilized to evaluate the performance of the observer.

Fig. 4 depicts both the actual and estimated joint torques and end-effector forces. It is evident from the figure that the estimates for both torque and force lag behind the actual forces and torques. This effect was also seen in the simulation results and should be expected since the

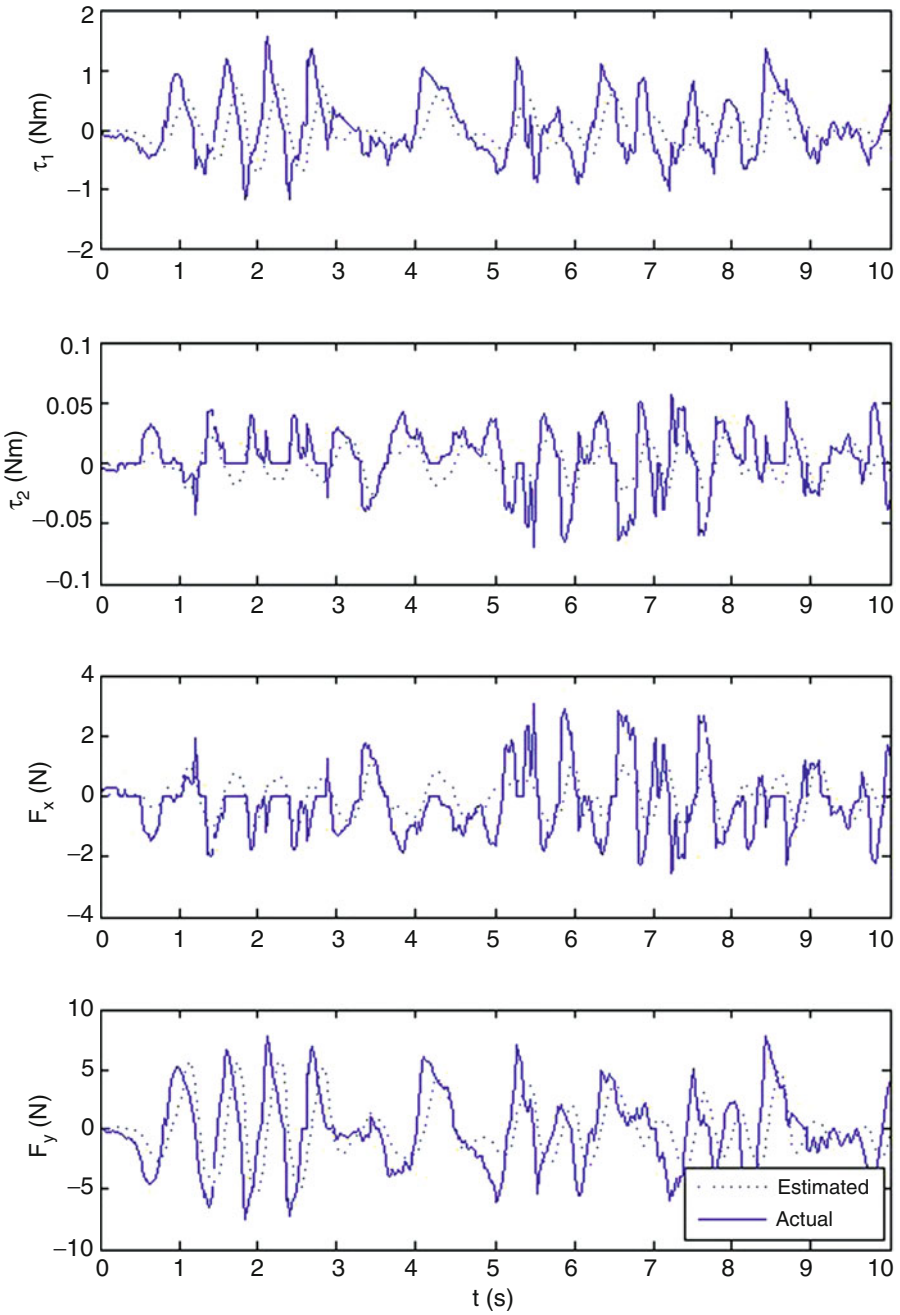


Fig. 4 *Estimation of external force on experimental system.*

observer was based upon a model for which the torques were assumed to have zero velocity and thus slowly changing. Although the observer's estimates trend well with the actual data, the lagging phenomenon warrants further investigation. Once solution to this problem might be to increase the speed of the observer by altering the observer gains. This solution, however, has been seen to introduce chattering into the estimates during the development of this work. Another approach to be investigated involves characterizing the external forces, which will be application dependent, so that the external torque model, $\dot{\tau}_e \approx \mathbf{0}$ can be replaced with a nontrivial state equation. This will result in an observer that performs the estimation of the external torques not just from system measurements but also from a model.

6 Conclusion

This paper has presented a discrete-time external force observer for a planer two-link manipulator. The observer is based upon the linear part of the manipulator's

equations of motion and also makes use of the nonlinear static friction terms. Through simulations and experimental results, the observer is seen to provide adequate estimates of low-bandwidth forces acting along the y-axis of the end-effector. The performance of the observer is degraded when estimating forces along the x-axis of the end-effector.

Improvements to the proposed observer should be examined in future work. This includes increasing the discrete-time sampling frequency, investigating different observer gain matrices and observer gain selection techniques, characterizing the nature of the external forces, and incorporating additional nonlinear compensation into the observer.

References

1. Daniel E. Whitney, "Force Feedback Control of Manipulator Fine Motion", *Journal of Dynamic Systems, Measurement, and Control*. vol. 99, no. 2, pp. 91-97, 1977
2. Daniel Helmick, Avi Okon, Matt DiCicco, "A Comparison of Force Sensing Techniques for Planetary Manipulation", *Proceedings of the IEEE Aerospace Conference*. 2005
3. Toshiyuki Murakami, Fangming Yu, Kouhei Ohnishi, "Torque Sensorless Control in Multidegree-of-Freedom Manipulator", *IEEE Transactions on Industrial Electronics*. vol. 40, no. 2, pp.259-265, 1993
4. Wen-Hua Chen, Donald J. Ballance, Peter J. Gawthrop, John O'Reilly, "A Nonlinear Disturbance Observer for Robotic Manipulators", *IEEE Transactions on Industrial Electronics*. vol. 47, no. 4, pp. 932-938, 2000
5. Chowarit Mitsantisuk, Kiyoshi Ohishi, Shiro Urushihara, Seiichiro Katsura, "Kalman Filter-Based Disturbance Observer and its Applications to Sensorless Force Control", *Advanced Robotics*. vol. 25, pp.335-353, 2011
6. Jinna Qin, Francois Lenard, Gabriel Abba, "Experimental External Force Estimation Using Non-Linear Observer for 6 Axes Flexible-Joint Industrial Manipulators", *Proceedings of 9th Asian Control Conference* pp. 1-6, 2013
7. Frank L. Lewis, Chaouki T. Abdallah, Darren M. Dawson, *Control of Robot Manipulators*. Macmillan, New York, 1993.
8. Aaron Radke, Zhiqiang Gao, "A Survey of State and Disturbance Observers for Practitioners" *Proceedings of the American Controls Conference*. pp. 5183-5188, 2006
9. Brian D.O. Anderson, John B. Moore, *Optimal Control Linear Quadratic Methods*. Dover Publications, Mineola, New York, 1990

A Joint-Space Parametric Formulation for the Vibrations of Symmetric Gough-Stewart Platforms

Behrouz Afzali-Far and Per Lidström

1 Introduction

Symmetric Gough-Stewart Platforms (SGSPs), which are also known as symmetric hexapods, have an increasing number of modern applications such as e.g. in collimation systems of large optical telescopes, CNCs and surgical robots [1–3]. An SGSP, being a parallel robot, has some inherent advantages over a serial robot including higher load-carrying capacity, stiffness and precision. An SGSP provides six degrees of freedom with the aid of its six linear actuators. The symmetry property referred to here is a rotationally symmetric structure of the six struts along with a symmetric platform (with principal moments of inertia $I_{xx} = I_{yy}$, see Figs. 1, 2).

Despite the wide range of research carried out on the SGSP kinematics, its vibrational behavior is not thoroughly investigated in the literature. To set up a complete parametric model for the SGSP vibrations, stiffness and inertia matrices need to be formulated parametrically. These matrices, as well as the eigenvalue problem, can be formulated in terms of the Cartesian-space or joint-space coordinates.

On the one hand, regarding the Cartesian-space approach, some researchers only obtain the stiffness matrix based on the Jacobian matrix focusing on statics without considering the dynamics of the system and solving the corresponding eigenvalue problem [4–7]. Whereas, some other researchers extend the static approach to a dynamic one by carrying out modal analyses with semi-analytical (see [8–11]) and parametric approaches (see [12–15]). Among the parametric approaches, Tian, Jiang, He and Tong [13], without giving a parametric solution to the damped eigenvalue problem, take the damping into account and study the influence of the damping on the decoupling in the Cartesian space. Tong, He and Jiang [14], calculate the eigenvalues of the undamped eigenvalue problem in the Cartesian space. Afzali-Far, Lidström and Nilsson develop a comprehensive parametric model of Gough-Stewart platforms vibrations in the Cartesian space [12, 15].

On the other hand, joint-space analyses are rarely reported in the literature (see [16–19]). Jiang, He and Tong [16] formulate the eigenvalue problem in the joint space and study the dynamic isotropic conditions, but a complete parametric formulation of the inertia matrix and the eigenvectors are not presented in this work. Furthermore, references [17–19], assume a reduced diagonal inertia matrix for the struts in the joint space, without giving a parametric formulation in terms of the system variables.

The joint-space formulation is crucial in order to obtain a better understanding of the system. In particular, the joint-space formulation can be employed to form an equivalent inertia matrix in order to consider the influence of the inertia of the struts. However, in the literature neither a parametric formulation for the inertia matrix nor for the eigenvectors in the joint space has yet been presented. This might be due to the complexity concerning the higher number of non-zero elements in the inertia matrix belonging to the joint-space.

In this paper, we present a complete parametric model of the SGSP vibrations at the neutral configuration. For the first time, the inertia matrix and the eigenvectors are formulated parametrically in the joint space. Firstly, a structured Jacobian-matrix is developed based on the kinematics of the system at the neutral configuration. Secondly, the linearized equations of motion are presented. Then, the inertia matrix is parametrically formulated in the joint space. Next, having solved the eigenvalue problem, the eigenvectors and eigenfrequencies are given parametrically. Finally, a reference SGSP is studied numerically where the results are compared with those obtained from a Cartesian-space based approach.

B. Afzali-Far (✉) • P. Lidström
Department of Mechanical Engineering, Lund University, Lund, Sweden
e-mail: Behrouz.afzali_far@mek.lth.se

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_48,

2 Kinematics

Generally, an SGSP consists of a fixed and a mobile platform which are known as *base* and *platform*, respectively (see Figs. 1, 2). An SGSP has six linear actuators which define the platform's position and orientation. Employment of universal joints for the connection of struts to the base and spherical joints for the connection to the platform is the most common design choice. The configuration of the platform can be defined either in the joint space or in the Cartesian space. The Cartesian-space representation of the configuration is defined by the coordinates

$$q_C = (\, t_x \quad t_y \quad t_z \quad \alpha \quad \beta \quad \gamma \,)^T \tag{1}$$

This representation consists of the position coordinates t_x, t_y, t_z of the platform's center of mass and its orientation by three Bryant angles α, β, γ. On the other hand, the joint-space representation is defined by coordinates equal to the length of struts as follows

$$q_J = l = (\, l_1 \quad \ldots \quad l_6 \,)^T \tag{2}$$

Note that in this paper the following non-dimensional parameters are used to simplify the equations

$$\mu_R = \frac{R_p}{R_b},\ \ \mu_H = \frac{H}{R_b},\ \ \mu_h = \frac{h}{R_b},\ \ \mu_l = \frac{l_0}{R_b} \tag{3}$$

where the parameters are shown in Figs. 1, 2 and l_0 is the length of the struts in the neutral configuration.

The base-frame and the platform-frame are the two right-handed orthonormal frames which are represented by their bases $\underline{\boldsymbol{e}}_b = \left(\boldsymbol{e}_b^x \ \ \boldsymbol{e}_b^y \ \ \boldsymbol{e}_b^z\right)$ and $\underline{\boldsymbol{e}}_p = \left(\boldsymbol{e}_p^x \ \ \boldsymbol{e}_p^y \ \ \boldsymbol{e}_p^z\right)$ as well as their origin points O_b and O_p, respectively (see Fig. 1).

The kinematic equation is, in the vector form, written

$$\boldsymbol{l}_i = \boldsymbol{t} + \boldsymbol{p}_i - \boldsymbol{b}_i,\ i = 1, \ldots, 6 \tag{4}$$

Accordingly, the matrix form of this equation is given by [12]

$$[\boldsymbol{l}_i]_{\underline{\boldsymbol{e}}_b} = [\boldsymbol{t}]_{\underline{\boldsymbol{e}}_b} + [\boldsymbol{R}]_{\underline{\boldsymbol{e}}_b}[\boldsymbol{p}_i]_{\underline{\boldsymbol{e}}_p} - [\boldsymbol{b}_i]_{\underline{\boldsymbol{e}}_b} \tag{5}$$

where the rotation matrix, based on the Bryant angles, is

$$[\boldsymbol{R}]_{\underline{\boldsymbol{e}}_b} = \begin{pmatrix} c\,\beta\,c\gamma & -c\beta s\gamma & s\beta \\ c\,\alpha s\,\gamma + s\,\alpha s\,\beta c\gamma & c\alpha\,c\,\gamma - s\alpha s\,\beta s\,\gamma & -s\,\alpha c\beta \\ s\alpha s\,\gamma - c\,\alpha s\beta c\gamma & s\alpha c\,\gamma + c\,\alpha s\,\beta s\gamma & c\alpha c\beta \end{pmatrix} \tag{6}$$

3 Jacobian Matrix

The Jacobian matrix related to the transformation $q_J = \hat{q}_J(q_C)$ between the Cartesian-space and the joint-space coordinates at the neutral configuration is

$$J = \left.\frac{\partial \hat{q}_J}{\partial q_C}\right|_{q_{C,0}} \in \mathbb{R}^{6\times 6} \tag{7}$$

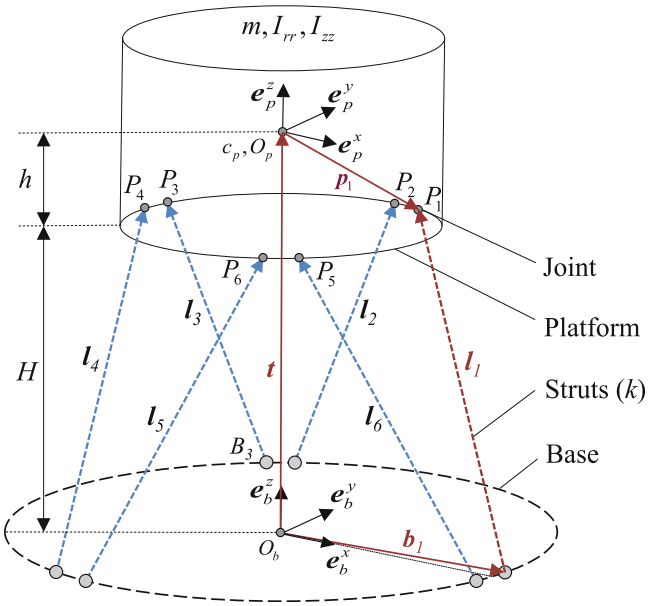


Fig. 1 Parameterized layout of a general SGSP

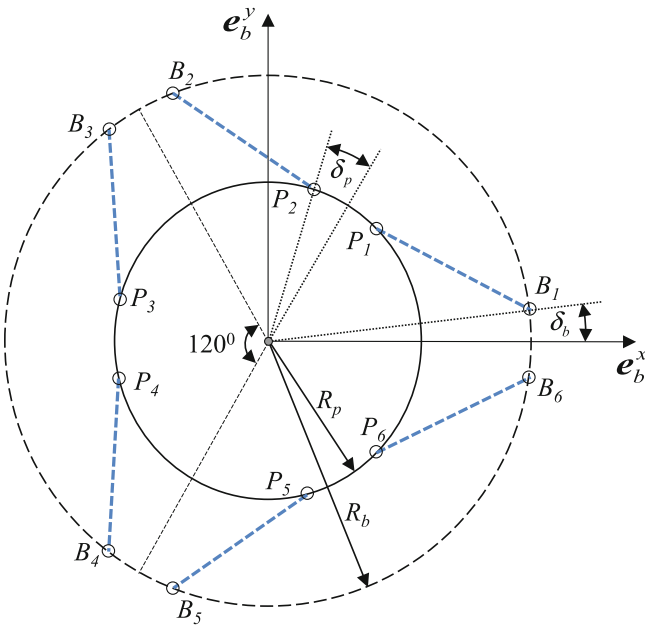


Fig. 2 Top view of a general SGSP

Consequently, the matrix elements j_{ik} of J are

$$j_{ik} = \frac{\partial \hat{l}_i}{\partial q_{C,k}}\bigg|_{q_{C,0}} = \frac{1}{2l_0}\left[\frac{\partial(\boldsymbol{l}_i \cdot \boldsymbol{l}_i)}{\partial q_{C,k}}\right]\bigg|_{q_{C,0}} \tag{8}$$

The structure of the 6×6 Jacobian matrix, containing 14 different elements, is given in [12] as follows

$$J = \frac{1}{2\mu_l}\begin{pmatrix} j_{11} & j_{12} & j_{13} & j_{14} & j_{15} & j_{16} \\ j_{21} & j_{22} & j_{13} & j_{24} & j_{25} & -j_{16} \\ j_{31} & j_{32} & j_{13} & j_{34} & j_{35} & j_{16} \\ j_{31} & -j_{32} & j_{13} & -j_{34} & j_{35} & -j_{16} \\ j_{21} & -j_{22} & j_{13} & -j_{24} & j_{25} & j_{16} \\ j_{11} & -j_{12} & j_{13} & -j_{14} & j_{15} & -j_{16} \end{pmatrix} \tag{9}$$

Here we apply further manipulations in order to obtain a simpler Jacobian matrix. This simplification can be done considering the fact that $JM_C^{-1}J^T$ is a centro-symmetric matrix, where M_C is the inertia matrix in the Cartesian space representation. By applying the centro-symmetric constraint to the $JM_C^{-1}J^T$ matrix and solving the obtained system of equations, the 14 parameters in (9) can be reduced to only 10 parameters. Hence, the simplified Jacobian matrix at the neutral configuration, based on ten parameters defined in (11), is

$$J = \frac{1}{2\mu_l}\begin{pmatrix} j_1 & j_3 & j_5 & j_6 & j_8 & j_{10} \\ j_2 & j_4 & j_5 & j_7 & j_9 & -j_{10} \\ -j_1 - j_2 & j_4 - j_3 & j_5 & j_7 - j_6 & -j_8 - j_9 & j_{10} \\ -j_1 - j_2 & j_3 - j_4 & j_5 & j_6 - j_7 & -j_8 - j_9 & -j_{10} \\ j_2 & -j_4 & j_5 & -j_7 & j_9 & j_{10} \\ j_1 & -j_3 & j_5 & -j_6 & j_8 & -j_{10} \end{pmatrix} \tag{10}$$

where

$$\begin{cases} j_1 = \mu_R \cos\delta_p - 2\cos\delta_b + \sqrt{3}\mu_R \sin\delta_p \\ j_2 = \mu_R \cos\delta_p + \cos\delta_b - \sqrt{3}\mu_R \sin\delta_p - \sqrt{3}\sin\delta_b \\ j_3 = \sqrt{3}\mu_R \cos\delta_p - \mu_R \sin\delta_p - 2\sin\delta_b \\ j_4 = \mu_R \sin\delta_p - \sqrt{3}\cos\delta_b - \sin\delta_b + \sqrt{3}\mu_R \cos\delta_p \\ j_5 = 2\mu_H \\ j_6 = R_b\left[\mu_R(\mu_H + \mu_h)\left(\sqrt{3}\cos\delta_p - \sin\delta_p\right) - 2\mu_h \sin\delta_b\right] \\ j_7 = R_b\left[\mu_R(\mu_H + \mu_h)\left(\sqrt{3}\cos\delta_p + \sin\delta_p\right) - \mu_h\left(\sqrt{3}\cos\delta_b + \sin\delta_b\right)\right] \\ j_8 = R_b\left[2\mu_h \cos\delta_b - \mu_R(\mu_H + \mu_h)\cos\delta_p - \sqrt{3}\mu_R(\mu_H + \mu_h)\sin\delta_p\right] \\ j_9 = R_b\left[\sqrt{3}\mu_R(\mu_H + \mu_h)\sin\delta_p + \sqrt{3}\mu_h \sin\delta_b - \right. \\ \left. \mu_h \cos\delta_b - \mu_R(\mu_H + \mu_h)\cos\delta_p\right] \\ j_{10} = R_b\mu_R\left[\sqrt{3}\cos\left(\delta_b + \delta_p\right) - \sin\left(\delta_b + \delta_p\right)\right] \end{cases} \tag{11}$$

and to make the final expressions more compact the following four parameters are used for the rest of the paper

$$\begin{cases} j_1' = -j_1 - j_2 \\ j_2' = j_4 - j_3 \\ j_3' = j_7 - j_6 \\ j_4' = -j_8 - j_9 \end{cases} \tag{12}$$

Note that the centro-symmetry property of the $JM_C^{-1}J^T$ matrix can be proven analytically. The proof is based on the tri-symmetry property of SGSPs which is further studied in Section 4.

4 Vibrations in the joint space

In this chapter we explain how it is possible to parametrically obtain the following linearized equations of motion in the joint space

$$M_J \ddot{q}_J + K_J q_J = 0_{6\times 1} \tag{13}$$

The inertia and the stiffness matrices M_J and K_J have to be expressed using the design variables of the system. For an SGSP the design space corresponding to the design variables is defined as

$$D_s = \{(d, I_{xx}, I_{rr}, m, k) \in \mathbb{R}^{10} | d \in G_s, m, k > 0, i = 1, ..., 6\} \tag{14}$$

where the geometrical design space is defined as [12]

$$G_s = \left\{\left(R_b, \mu_R, \mu_H, \mu_h, \delta_b, \delta_p\right) \in \mathbb{R}^6 |\, R_b > 0,\ \mu_R > 0,\right.$$
$$\left. 0 < \mu_H < \mu_l,\ \delta_b + \delta_p < \pi/3,\ \delta_b, \delta_p \geq 0\right\} \tag{15}$$

To set up a parametric stiffness matrix the elastic potential energy V is formulated. It can be directly written in terms of the joint-space coordinates as follows

$$V = \frac{1}{2}\sum_{i=1}^{6} k_i \Delta l_i^2 \tag{16}$$

where $\Delta l_i = l_i - l_0$. Hence, the corresponding stiffness matrix is simply obtained as

$$K_J = diag(\,k \quad \ldots \quad k\,) \in \mathbb{R}^{6\times 6} \tag{17}$$

where $k \in \mathbb{R}^+$ is the strut stiffness and the struts are here assumed to have identical stiffness properties.

To obtain the inertia matrix in the joint space and based on the rotational symmetry properties of the SGSP, one can simplify the following general inertia matrix in the joint space

$$M_J = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} & m_{15} & m_{16} \\ & m_{22} & m_{23} & m_{24} & m_{25} & m_{26} \\ & & m_{33} & m_{34} & m_{35} & m_{36} \\ & & & m_{44} & m_{45} & m_{46} \\ & sym & & & m_{55} & m_{56} \\ & & & & & m_{66} \end{pmatrix} \tag{18}$$

Assuming the tri-symmetry property for the struts and the platform, at the neutral configuration, one can simplify (18) step by step. Firstly, the symmetric and identical geometry of the struts implies identical diagonal elements for the inertia matrix

$$m_{11} = m_{22} = m_{33} = m_{44} = m_{55} = m_{66} \rightarrow m_1 \tag{19}$$

Secondly, only using the tri-symmetry property of the struts and the platform, one can write the following relations for the inertia matrix in the joint space

$$\begin{cases} m_{12} = m_{34} = m_{56} \rightarrow m_2 \\ m_{13} = m_{15} = m_{24} = m_{26} = m_{35} = m_{46} \rightarrow m_3 \\ m_{14} = m_{25} = m_{36} \rightarrow m_4 \\ m_{16} = m_{23} = m_{45} \rightarrow m_5 \end{cases} \tag{20}$$

Hence, this lead to the formation of the following simple inertia matrix

$$M_J = \begin{pmatrix} m_1 & m_2 & m_3 & m_4 & m_3 & m_5 \\ & m_1 & m_5 & m_3 & m_4 & \\ & & m_1 & m_2 & & \\ & & . & . & & \\ & . & cen. & sym. & . & \\ . & & & & & . \end{pmatrix} \tag{21}$$

a positive definite centro-symmetric matrix which is structured by only five inertia elements. Interestingly, the obtained centro-symmetric structure in (21) is valid even if one takes the inertia of the struts into account. Thus this method has the potential of considering the inertia of the struts as well.

The formulation of the joint-space inertia matrix starting from the non-linear kinetic energy of the system is quite complicated. Therefore, the joint-space inertia matrix can be formulated directly, based on a linearized approach, by the following equation

$$M_J^{-1} = J M_C^{-1} J^T \tag{22}$$

where M_C is given by [12]

$$M_C = diag(\, m \quad m \quad m \quad I_{rr} \quad I_{rr} \quad I_{zz}\,) \in \mathbb{R}^{6\times 6} \tag{23}$$

and the platform is considered symmetric ($I_{xx} = I_{yy} = I_{rr}$). Having calculated (22), ultimately the same structure as given in (21) is also obtained here.

To obtain a simpler formulation of the eigenvalue problem, it is more efficient to parameterize the inverse inertia matrix M_J^{-1} instead of M_J. The inverse of a positive definite centro-symmetric (*cen.*) matrix is also centro-symmetric (See *Remark 1* below). Therefore, the inverse of the inertia matrix in the joint space can also be written, using five parameters, as follows

$$M_J{}^{-1} = \begin{pmatrix} m_1' & m_2' & m_3' & m_4' & m_3' & m_5' \\ & m_1' & m_5' & m_3' & m_4' & \\ & & m_1' & m_2' & & \\ & & . & . & & \\ & . & cen. & sym. & . & \\ . & & & & & . \end{pmatrix} \tag{24}$$

where, based on (22), the inertia matrix components are calculated as

$$\begin{cases} m_1' = \dfrac{j_1^2 + j_3^2 + j_5^2}{m}, + \dfrac{j_6^2 + j_8^2}{I_{rr}}, + \dfrac{j_{10}^2}{I_{zz}} \\ m_2' = \dfrac{j_1 j_2 + j_3 j_4 + j_5^2}{m} + \dfrac{j_6 j_7 + j_8 j_9}{I_{rr}} - \dfrac{j_{10}^2}{I_{zz}} \\ m_3' = \dfrac{j_1 j_1' + j_3 j_2' + j_5^2}{m}, + \dfrac{j_6 j_3' + j_8 j_4'}{I_{rr}}, + \dfrac{j_{10}^2}{I_{zz}}. \\ m_4' = \dfrac{j_1 j_1' - j_3 j_2' + j_5^2}{m}, + \dfrac{j_8 j_4' - j_6 j_3'}{I_{rr}}, - \dfrac{j_{10}^2}{I_{zz}} \\ m_5' = \dfrac{j_1^2 - j_3^2 + j_5^2}{m}, + \dfrac{j_8^2 - j_6^2}{I_{rr}}, - \dfrac{j_{10}^2}{I_{zz}} \end{cases} \tag{25}$$

At this stage, all the necessary matrices belonging to the joint-space equations of motion in (13) are established parametrically. In other words, those matrices are formulated in terms of the design variables of the system specified in (14) and (15).

Remark 1: Introduce the matrix $L \in \mathbb{R}^{6\times 6}$ defined by

$$L = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{26}$$

then $L \in sym(\mathbb{R}^{6\times 6})$ and $L^2 = 1_{6\times 6}$ A matrix $A \in sym(\mathbb{R}^{6\times 6})$ is centro-symmetric if and only if $LAL = A$. If $A \in sym^{+}(\mathbb{R}^{6\times 6})$ is centro-symmetric then A^{-1} is centro-symmetric, Cantoni & Butler [20].

5 Parametric Modal Analysis

Having parametrically formulated the inertia and stiffness matrices in the joint space, the eigenvalue equation $(M_J^{-1}K_J - \omega^2 I_{6\times 6})X_J = 0$ has to be solved parametrically in order to obtain eigenvectors and eigenfrequencies as functions of the design variables.

The undamped eigenfrequencies ($\omega_i \in \mathbb{R}^+, i = 1, \ldots, 6$), based on the inverse inertia matrix parameters given in (25), are

$$\begin{cases} \omega_{1,2} = \sqrt{k\left(m_1' - m_3' - \Phi\right)} \\ \omega_3 = \sqrt{k\left(m_1' + 2m_3' - m_2' - m_4' - m_5'\right)} \\ \omega_4 = \sqrt{k\left(m_1' + 2m_3' + m_2' + m_4' + m_5'\right)} \\ \omega_{5,6} = \sqrt{k\left(m_1' - m_3' + \Phi\right)} \end{cases} \tag{27}$$

where

$$\Phi = \sqrt{m_2'^2 + m_4'^2 + m_5'^2 - m_2'm_4' - m_2'm_5' - m_4'm_5'} \tag{28}$$

Note that $\Phi \in \mathbb{R}^+$, since

$$m_2'^2 + m_4'^2 + m_5'^2 - m_2'm_4' - m_2'm_5' - m_4'm_5' =$$

$$\frac{\left(m_2' - m_4'\right)^2}{2} + \frac{\left(m_2' - m_5'\right)^2}{2} + \frac{\left(m_4' - m_5'\right)^2}{2} \geq 0$$

The associated normalized modal matrix in the joint space is given by

$$X_J = \begin{pmatrix} \chi_{J,1}/\chi_{J,4} & \chi_{J,2}/\chi_{J,4} & 1 & 1 & \chi_{J,1}/\chi_{J,4} & -\chi_{J,2}/\chi_{J,4} \\ -\chi_{J,2}/\chi_{J,4} & -\chi_{J,3}/\chi_{J,4} & -1 & 1 & \chi_{J,2}/\chi_{J,4} & -\chi_{J,3}/\chi_{J,4} \\ -\chi_{J,3}/\chi_{J,4} & -\chi_{J,2}/\chi_{J,4} & 1 & 1 & -\chi_{J,3}/\chi_{J,4} & \chi_{J,2}/\chi_{J,4} \\ \chi_{J,2}/\chi_{J,4} & \chi_{J,1}/\chi_{J,4} & -1 & 1 & -\chi_{J,2}/\chi_{J,4} & \chi_{J,1}/\chi_{J,4} \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & -1 & 1 & 0 & 1 \end{pmatrix} \tag{29}$$

where the modal matrix elements are parameterized by the inverse inertia matrix elements according to

$$\begin{cases} \chi_{J,1} = m_2' - m_4' \\ \chi_{J,2} = \Phi \\ \chi_{J,3} = m_5' - m_2' \\ \chi_{J,4} = m_4' - m_5' \end{cases} \tag{30}$$

Note that the modal matrix given in (29) can be normalized differently. Regardless of the normalization method, equations (29) and (30) interestingly show that the modal matrix X_J is only a function of the joint-space inertia matrix elements (and not the stiffness matrix elements). This is due to the fact that the joint-space stiffness matrix is a scaled identity matrix and the inertia matrix represents the mass properties as well as the geometrical properties. However, this is not the case for the eigenvectors parametrically obtained by using the Cartesian-space coordinates where the eigenvectors are dependent on the elements of both inertia and stiffness matrices [12]. It should also be noted that, for a given SGSP, the sequence of the six eigenfrequencies presented in (27), may change according to its geometrical and mass properties.

6 Numerical Example

To exemplify the established parametric model for the joint-space vibrations, a reference SGSP is considered. The reference SGSP has the following properties [12]

$$\begin{cases} m = 100\ (\text{kg}),\ I_{rr} = 5\ (\text{kgm}^2),\ I_{zz} = 10\ (\text{kgm}^2), \\ R_b = 0.500\ (\text{m}),\ k = 10^6\ (\text{N/m}), \\ \delta_b = 0.1745\ (\text{rad}),\ \delta_p = 0.2618\ (\text{rad}), \\ \mu_R = 0.8,\ \mu_H = 1.8,\ \mu_h = 0.2 \end{cases} \tag{31}$$

First, according to (11), the ten elements of the Jacobian matrix are calculated and consequently the Jacobian matrix is obtained as

$$J = \begin{pmatrix} -0.22 & 0.21 & 0.95 & 0.29 & -0.25 & 0.12 \\ 0.29 & -0.09 & 0.95 & 0.36 & -0.13 & -0.12 \\ -0.07 & -0.30 & 0.95 & 0.07 & 0.38 & 0.12 \\ -0.07 & 0.30 & 0.95 & -0.07 & 0.38 & -0.12 \\ 0.29 & 0.09 & 0.95 & -0.36 & -0.13 & 0.12 \\ -0.22 & -0.21 & 0.95 & -0.29 & -0.25 & -0.12 \end{pmatrix} \tag{32}$$

Second, using (17), the stiffness matrices in the joint space is given by

$$K_J = diag\begin{pmatrix} 10^6 & \ldots & 10^6 \end{pmatrix} \in \mathbb{R}^{6\times 6} \tag{33}$$

Next, to calculate the inertia matrix in the joint space, one first needs to have the inertia matrix in the Cartesian space using (23)

$$M_C = diag\begin{pmatrix} 10^2 & 10^2 & 10^2 & 5 & 5 & 10 \end{pmatrix} \in \mathbb{R}^{6\times 6} \tag{34}$$

Once the inertia matrix in the Cartesian space is written, using (24) and (25), the inverse inertia matrix and consequently the joint-space inertia matrix is calculated as

$$M_J = \begin{pmatrix} 146.4 & -125.1 & -40.4 & 74.6 & -40.4 & 3.1 \\ -125.1 & 146.4 & 3.1 & -40.4 & 74.6 & -40.4 \\ -40.4 & 3.1 & 146.4 & -125.1 & -40.4 & 74.6 \\ 74.6 & -40.4 & -125.1 & 146.4 & 3.1 & -40.4 \\ -40.4 & 74.6 & -40.4 & 3.1 & 146.4 & -125.1 \\ 3.1 & -40.4 & 74.6 & -40.4 & -125.1 & 146.4 \end{pmatrix} \tag{35}$$

Then, the eigenfrequencies, obtained using (27), are as follows

$$\begin{cases} \omega_{1,2} = 52.555\,\text{rads}^{-1}, \omega_3 = 94.066\,\text{rads}^{-1}, \\ \omega_4 = 233.374\,\text{rads}^{-1}, \omega_{5,6} = 295.361\,\text{rads}^{-1} \end{cases} \tag{36}$$

It is noteworthy that exactly the same results are obtained in [12] using a Cartesian-space approach. Finally, the eigenvectors are calculated according to (29) as

$$X_J = \begin{pmatrix} -2.793 & -2.451 & 1 & 1 & -2.793 & 2.451 \\ 2.451 & 1.793 & -1 & 1 & -2.451 & 1.793 \\ 1.793 & 2.451 & 1 & 1 & 1.793 & -2.451 \\ -2.451 & -2.793 & -1 & 1 & 2.451 & -2.793 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & -1 & 1 & 0 & 1 \end{pmatrix} \tag{37}$$

Equation (37) presents six modes including two lateral-dominant, two tilting-dominant, one axial and one twisting mode. Since the position and orientation of the platform is directly given by the Cartesian-space coordinates, it is of interest to obtain the Cartesian space modal matrix X_C as well. It can be obtained from the joint-space modal matrix X_J as follows

$$X_C = J^{-1} X_J = \begin{pmatrix} -26.25 & -9.37 & 0 & 0 & -0.0019 & -0.0053 \\ 30.90 & 26.25 & 0 & 0 & -0.0016 & -0.0019 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -1.18 & -1 & 0 & 0 & -0.8495 & -1 \\ -1 & 0.12 & 0 & 0 & 1 & 2.8000 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \tag{38}$$

A proof of the non-singularity of the J matrix is given in [12]. Applying linear combinations using the first two columns and accordingly the last two columns of the matrix given in (38) leads to pure eigenvectors in the $x - \beta$ and the $y - \alpha$ directions. Having done the calculations, the Cartesian-space modal matrix is obtained as

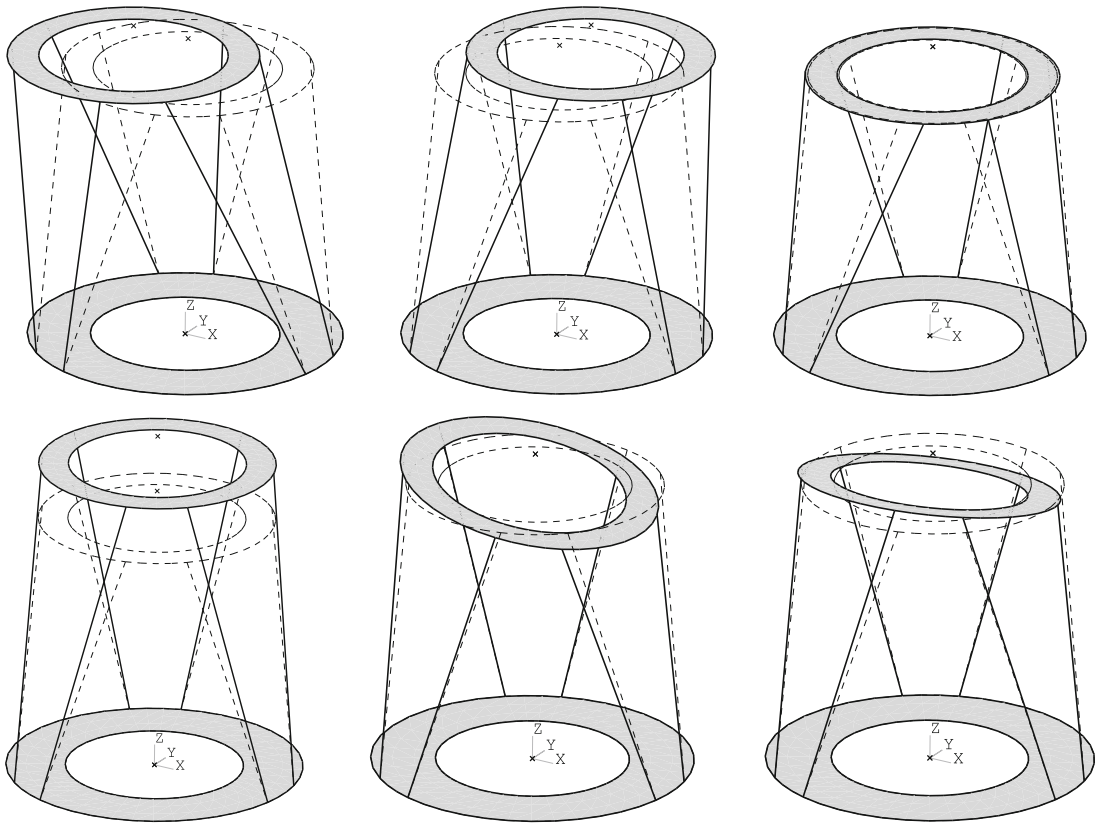


Fig. 3 Six mode shapes of the reference SGSP

$$X_C' = \begin{pmatrix} -26.25 & 0 & 0 & 0 & -0.0019 & 0 \\ 0 & 26.25 & 0 & 0 & 0 & -0.0019 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \quad (39)$$

The eigenvectors (mode shapes) corresponding to the modal matrix presented in (39) are demonstrated in Fig. 3. For the same reference SGSP, with a Cartesian-space approach, exactly the same results are derived in [12].

7 Conclusion

The present work establishes, for the first time, a complete parametric model for the vibrations of SGSPs in the joint space at the neutral configuration. Firstly, the kinematic of SGSPs is taken into account in which the Bryant angles represent the orientation of the platform. Based on the kinematics of the system and the obtained centro-symmetric structure of the inertia matrix, a new and simplified form of the Jacobian matrix with only 10 parameters is parametrically formulated. The joint-space inertia matrix is presented parametrically where its centro-symmetric structure is demonstrated. By solving the eigenvalue problem, the eigenfrequencies and eigenvectors of SGSPs are parametrically given in terms of the design variables. Finally, a numerical calculation is carried out for a reference SGSP in order to demonstrate some results of the presented analytical model.

In comparison with a Cartesian-space approach, generally for asymmetric GSPs a joint-space approach can lead to more complicated calculations but in the case of SGSPs (symmetric GSPs), it is shown that using the joint-space coordinates can be as efficient as using the Cartesian-space coordinates. Furthermore, in this paper two advantages of the joint-space approach are shown. Firstly, it is interestingly observed that, in the presented joint-space approach, the parametric eigenvectors are only dependent on the inertia matrix elements; however, in a Cartesian approach the eigenvectors are functions of the elements of both the inertia and the stiffness matrices. Secondly, since adding the inertia of the struts does not change the structure of the inertia matrix, this method has a good potential to be further extended.

References

1. J.J. Zierer, J.H. Beno, D.A. Weeks, I.M. Soukup, J.M. Good, J.A. Booth, G.J. Hill, M.D. Rafal, Design, testing, and installation of a high-precision hexapod for the Hobby-Eberly Telescope Dark Energy Experiment (HETDEX), Ground-Based and Airborne Telescopes Iv, 8444 (2012).
2. Y.D. Patel, Parallel Manipulators Applications—A Survey, Modern Mechanical Engineering, 02 (2012) 57-64.
3. Y. Ting, Y.S. Chen, H.C. Jar, Modeling and control for a Gough-Stewart platform CNC machine, Journal of Robotic Systems, 21 (2004) 609-623.
4. V.T. Portman, V.S. Chapsky, Y. Shneor, Workspace of parallel kinematics machines with minimum stiffness limits: Collinear stiffness value based approach, Mechanism and Machine Theory, 49 (2012) 67-86.
5. J. Chen, F. Lan, Instantaneous stiffness analysis and simulation for hexapod machines, Simulation Modelling Practice and Theory, 16 (2008) 419-428.
6. B.S. El-Khasawneh, P.M. Ferreira, Computation of stiffness and stiffness bounds for parallel link manipulators, International Journal of Machine Tools & Manufacture, 39 (1999) 321-342.
7. C. Gosselin, Stiffness Mapping for Parallel Manipulators, Ieee Transactions on Robotics and Automation, 6 (1990) 377-382.
8. S. Pedrammehr, M. Mahboubkhah, N. Khani, A study on vibration of Stewart platform-based machine tool table, International Journal of Advanced Manufacturing Technology, 65 (2013) 991-1007.
9. M. Mahboubkhah, M.J. Nategh, S.E. Khadem, A comprehensive study on the free vibration of machine tools' hexapod table, International Journal of Advanced Manufacturing Technology, 40 (2009) 1239-1251.
10. M. Mahboubkhah, M.J. Nategh, S.E. Khadem, Vibration analysis of machine tool's hexapod table, International Journal of Advanced Manufacturing Technology, 38 (2008) 1236-1243.
11. P. Mukherjee, B. Dasgupta, A.K. Mallik, Dynamic stability index and vibration analysis of a flexible Stewart platform, Journal of Sound and Vibration, 307 (2007) 495-512.
12. B. Afzali-Far, P. Lidström, K. Nilsson, Parametric damped vibrations of Gough–Stewart platforms for symmetric configurations, Mechanism and Machine Theory, 80 (2014) 52-69.
13. T.X. Tian, H.Z. Jiang, J.F. He, Z.Z. Tong, Influence of passive joint damping on modal space decoupling for a class of symmetric spatial parallel mechanisms, Applied Mechanics and Materials, 2013, pp. 1152-1157.
14. Z.Z. Tong, J.F. He, H.Z. Jiang, G.R. Duan, Optimal design of a class of generalized symmetric Gough-Stewart parallel manipulators with dynamic isotropy and singularity-free workspace, Robotica, 30 (2012) 305-314.
15. B. Afzali-Far, P. Lidström, K. Nilsson, Analytical Stiffness Optimization of High-Precision Hexapods for Large Optical Telescope Applications, Proceedings of the 25th Nordic Seminar on Computational Mechanics, 2012.
16. H.Z. Jiang, J.F. He, Z.Z. Tong, Characteristics analysis of joint space inverse inertia matrix for the optimal design of a 6-DOF parallel manipulator, Mechanism and Machine Theory, 45 (2010) 722-739.
17. H.M. Lin, J.E. McInroy, Disturbance attenuation in precise hexapod pointing using positive force feedback, Control Engineering Practice, 14 (2006) 1377-1386.
18. Y.X. Chen, J.E. McInroy, Decoupled control of flexure-jointed hexapods using estimated joint-space mass-inertia matrix, Ieee Transactions on Control Systems Technology, 12 (2004) 413-421.
19. J.E. McInroy, J.C. Hamann, Design and control of flexure jointed hexapods, Ieee Transactions on Robotics and Automation, 16 (2000) 372-381.
20. A. Cantoni, P. Butler, Eigenvalues and Eigenvectors of Symmetric Centrosymmetric Matrices, Linear Algebra and Its Applications, 13 (1976) 275-288.

Software Project Planning Using Agile

Jianchao Han and Yan Ma

1 Introduction

Traditional software management technologies can be characterized as linear, sequential processes. Confronting the increasing volatile external and internal environment changes, agile project management is preferred as it offers flexible project management process, changeable and low-cost development mode and better communications with customers. If you have any doubt about the volatile environment changes, you can refer the following samples: Social media has experienced the development from telegraph to email, from email to BBS, from BBS to Blogger, from Blogger to Facebook/twitter or Youtube [1, 2]. Business requirements are no longer the same as those 20 years ago. Take middleware as an example. Middleware technology is widely applied in financial industry, sales, health care and insurance areas. It is the key point to integrate and automate the business processes in distributed organizations to lower business cost and improve efficiency and productivity. The middleware product architecture evolved from 3-tier enterprise applications to Service Oriented Architecture to Software as a server to cloud based services.

Observing this tendency for software requirements to change, more and more software development organizations have adopt Agile project management to solve the new challenges in nowadays' market. Agile methodologies include Scrum, eXtreme Programming (XP), Crystal, Adaptive Software Development, etc. Projects that employ Agile methodologies share the following characters [3]:

- organic teams of from seven to nine members – organizing a project into organic teams implies a minor interaction penalty in terms of communication and cooperation [4];
- guiding vision – a project vision should be translated into a simple statement of project purpose, understood and accepted by all the team members;
- simple rules – team members follow simple rules, but their interactions result in complex behavior that emerges over time;
- free and open access to information;
- light touch management style and adaptive leadership that has the ability to adapt to changes, views the system as fluid system composed of intelligent people, recognizes the limits of external control in establishing order, and adopts humanistic problem-solving approaches.

As English proverb says, "if you fail to plan, then you plan to fail." Project planning is the first task encountered by a project manager when he/she gets involved in a project. As part of the initial phase, project planning covers requirements specification/analysis, procedure/infrastructure definition and efforts assessment. After having identified the reasons why we need Agile project management and the basic characters of projects that employ Agile project management, this paper is organized to show the related research work in Section 2, Exploring the difference between traditional project planning and Agile project planning in Section 3. Section 4 dives into the details and process of Agile software project planning for the adaptability to the increasing turbulent and unpredictable market changes. Section 5 develops the risks assessment and management in Agile software project planning. Section 6 is the conclusion and future work.

2 Related Work

Linda Rising and Norman S. Janoff [5] explored the iterative Agile project planning using Agile methodology Scrum. They summarized the experiences with Scrum from three diverse software development teams at AG Communication Systems. The summary shows that planning tasks must be

J. Han (✉) • Y. Ma
Department of Computer Science, California State University Dominguez Hills, Carson, CA 90747, USA
e-mail: jhan@csudh.edu; yma4@toromail.csudh.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_49,

well quantified and achievable within the sprint time; tasks must be assigned to one individual or the primary individual if it is shared by multiple persons; sprint goal works as an effective tools for keeping people on track and aware of expectation as some people are not good at planning their workload.

Similarly H. Frank Cervon [7] analyzed Scrum model and he organized the Scrum process as the kick off, the sprint planning meeting, the sprint, the daily Scrum and sprint review meeting. The Scrum master and the product owner hold the sprint planning meeting at the beginning of the sprint, which could take up to one day. The spring planning meeting consists of two parts: the product backlog definition followed by the sprint goal determination and the sprint backlog creation.

Arthur English [16] emphasized the planning game in eXtreme Programming (XP) model. It encompasses requirements definition and project planning. End user defines application features with stories instead of UML diagrams or models. Developers prioritize the stories and schedule the most important and difficult for next iteration. Planning and prioritizing runs through the product development cycles.

Martin Molhanec [16] developed a framework for Agile project management and put Agile project planning into the initial phase. The initial phase includes 1) justification (assessment, feasibility study); 2) definition and validation of requirements; 3) definition of initial documents; and 4) definition of project infrastructure.

3 Differences Between Traditional Project Planning and Agile Project Planning

Traditional project planning is represented as "A formal, approved document used to guide both project execution and project control. The primary uses of the project plan are to document planning assumptions and decisions, facilitate communication among stakeholders, and document approved scope, cost, and schedule baselines. A project plan may be summary or detailed." [8] The following contents are commonly included in a traditional project plan:

- Plan description that gives a brief description about what the plan covers;
- Project prerequisites and dependences that must be satisfied before the project can continue;
- Planning assumptions concerning the resource availabilities and technology requirements;
- Project level Gantt chart or bar chart that identifies the management stages;
- Project level activity network;
- Contingency plans that record what to do if unintended events happen.

A traditional project schedule maps two variables:

- Effort required (i.e. from tasks and associated estimates)
- Effort available over time (the velocity or the capacity)

When problems occur, a project that employs traditional project planning usually extends the schedule. Because the scope is the fixed aspect while time and cost can be adjusted.

If problems occur, a project that employs Agile project planning would cut down the scope to stick to the planned time and cost.

Several contents in traditional project plan are not shown in Agile project planning. The Work Breakdown Structure (WBS) [9] is discarded, which looks at the project as a series of fine grain technical tasks whereas Agile sees the project as a series of product increments.

Project level activity planning or resource estimation is also discarded. Only next iteration planning is defined with details. Agile planning believes that it is impossible to make a correct project level estimation at the beginning because some essential information is only available during the project. Agile planning technology keeps changing the iteration plans based on the information gained during the project and feedbacks from customers.

The team is also more involved in Agile project planning. Planning is not just the task of the project manager. However, the project manager is still responsible for getting the planning done. The project manager behaves like a Scrum master in Scrum model, for example.

4 Adapt to the Increasing Turbulent and Unpredictable Market Change

Although there are multiple Agile methodologies such as eXtreme Programming (XP), Scrum, Feature Driven Development (FDD), etc, the project planning can be fairly uniformed. Mike Cohn's Planning Onion beautifully illustrates the levels at which an organization must plan its activity [10]. The similar description in Agile development can also be found in [11].

An Agile team must implement the lower three levels: release plan, Timebox plan and day plan. The lower the plan locates the more details it contains. A release plan covers the activities for a public major release which defines what new features the customers will get and when. A release plan could last from 2 weeks to 6 months (usually 3 months as an average). Each release includes one or more Timeboxes. Each Timebox composes one or more days.

4.1 Release planning

At the beginning of every release, the product owner would hold a release planning meeting where well-prepared user

stories would be prioritized and put into different Timeboxes. The team, together with the product owner and the project manager, estimates the effort size instead of the effort of each user story. Story point is often used in estimating the size of the user story. Story point is an abstract measure based on the size of the functionality. When estimating the story point, the following aspects should be considered [11]: 1) what is involved in the story? 2) how big is the story relative to other stories that the team has already developed? Idea day is commonly used in estimation. eXtreme Programming (XP) first proposed the concept of idea day as the estimating unit for User Stories [12]. Idea day means focusing on work without interruption as if you worked overtime on a weekend: no meeting, no group discussion. An idea day does not equal to a working day which often leads to stakeholders' confusion as they don't know the difference between them. The advantage of using idea day is that it shows the team's velocity clearly without interruptions from external environment.

The story points have to be re-estimated if the team finds the significant flaws in the original understanding. For example, a main entry that looks straightforward actually needs complex processing that the team hadn't been aware of. In this case, the team has to re-estimate the story points based on the new understanding.

To deal with this kind of uncertainty, we introduce buffers into release planning. The buffers are concepts from Eli Goldratt's Critical Chain Project Management approach within the Theory of Constraints [13]. Goldratt mentioned that the critical chain is defined as the longest chain [not path] of dependent tasks, where 'dependent' refers to resources and resource contention across tasks/projects as well as the sequence and logical dependencies of the tasks themselves. This differs from the Critical Path Method [13]. With the concept of critical chain, the three buffers are defined as project buffer, feeding buffers and resource buffers.

- Project buffer is inserted at the end of a project to deal with contingency or uncertainties. As we know, in the Agile project planning, the efforts are estimated based on story points and no space is reserved for contingency or uncertainty. The longer a project lasts the more contingencies there could be. The more uncertainties we have the larger buffer we need. So the Agile project manager usually sets one or more Timeboxes at the end of the project without any user story assigned to them in order to deal with the contingency and uncertainties. Goldratt also recommends that if you have less than 10 user stories, there is no need for the project buffer or the project buffer has to be at least 20 % of the total project time, otherwise, it won't be long enough to handle the contingencies [13].
- Feeding buffers are used for component contingencies. When we are talking about user stories, we assume that each user story is independent from each other so that different developer teams can work on different user stories in parallel. However, in practice, dependencies among the user stories are quite common. The delay to feeding the tasks in critical chain would cause the delay of the whole project. Feeding buffer is the strategy to highlight the dependencies so that the supplying team would take this as priority and deliver things early. The supplying teams have a feeding buffer in the Timebox where they deliver the component. The receiving team has a user story that represents the work to receive the component in the same or later Timebox. If the dependency does not happen on the critical chain, no feeding buffer is needed then.
- Resource buffers are used to ensure the appropriate people, skills and resources are available on critical chain as soon as they are needed. Put the resource buffer before the user story that needs the constrained resources to ensure the resources are not held up by anything else.

4.2 Timebox Planning

In the release planning we slice the whole project into multiple user stories and estimate the size of efforts using story points. We also set the buffers to handle uncertainty and contingencies. Before jumping into the Timebox planning, we prepare acceptance tests for the user stories assigned in this Timebox and sufficient supporting documents to describe each user story adequately. The acceptance tests helps to fully understand the intent of the user stories. The test hasn't to be in detail yet but for each acceptance test it has to answer the question formatted as

"Given < preconditions > when < event > then < result>".

If you have too many acceptance tests for a single user story, it is a sign to slice the user story further.

With the acceptance tests and supporting documents prepared, we can start the Timebox planning. It includes [9]

1) figuring out tasks needed to complete the user stories;
2) estimating the tasks;
3) checking that the team has sufficient capacity during the Timebox to do all the tasks; and
4) adjusting user stories to fit into the Timebox.

When we are estimate the effort of a task we use the concept "idea hour". Similar to the concept "idea day", "idea hour" means the hours that you 100 % focus on the tasks where no meetings, no group discussion, no phone calls. One has four to six idea hours per day. Idea hours help us calculate a developer's velocity without external interruptions. It could be the individual developer estimate the idea hours needed for the tasks. Or it could be a estimation from team.

4.3 Day Planning

Day plans are informal and not written down. They are used during the Timebox to answer three questions:

1) What you have done yesterday?
2) What you are doing today?
3) Any impediment?

The answers are prepared by each team member and discussed in a short-time daily team meeting. In Scrum model, we call it as Scrum meeting. The daily team meeting won't last longer than 15 minutes. It effects as status report and impediments clarification for the whole team. In case there is any emergent issue or a stopping issue, team's efforts would be utilized to solve it. Usually the questioner will get answers like "I have met this before; let me give you some documents offline." Or "I know who has the experience on this. You can contact him/her for further information". It is something that important enough to change the Timebox plan or the release plan, the daily team meeting or the day plan makes sure the issue is exposed as early as possible so that there could be plenty of time to handle it.

5 Risks Planning and Management in Agile Projects

Risks are inherent part of projects. Hence risk management should be integrated into Agile project planning. The difference between risk and issues is that we are sure to know that issues are going to happen while we are not sure whether a risk will happen. The factors of a risk include uncertainty, loss, and duration. An effective risk management plan involves [14]:

1) identifying the risk;
2) analyzing the risk to determine its severity;
3) prioritizing the identified risks per their severities;
4) creating action plans; and
5) continuously monitoring and following the action plan to ensure they mitigate the risks.

The first four actions should be included in the Agile project planning and the last action should be conducted throughout the duration of the project. Risks from two different aspects, tasks and people, will be addressed as follows.

5.1 Risks from Tasks

The risks from tasks should be dissected into the essential drives. Knowing the underlying reasons for the risks naturally provides the action plan for the risks. Usually it is impractical to get a team agreement about the severity of the risks because different people may have different perceptions and understandings. Instead, we look into the essential reasons of the risk; identify the possibility and the severity of the impact to give an object evaluation of the risk severity.

Risks management is integrated into Agile project planning and it does change/impact the iteration plan determination. The whole team is involved into the risk management. Instead of having an official analysis about the financial loss or quality loss, the team is involved in the Agile project planning meeting and discuss and determine the severities in 1 or 2 hours. Preston mentioned the process about the risk management during the general iteration planning meeting [14]:

- Review the product backlog items
- Perform risk identification and analyze the prioritization
- Map the risks identified to the backlog
- Select the backlog items associated with the critical risks as work candidates
- Derive the iteration goal.

Every team member writes down the risks he/she identified on the cards and pin them on the wall. After putting each card into correct groups, team members vote on the risk group by sticking points onto the groups. Once all the team members have voted, project manager selects the top 5 risk groups as the critical risks and determines the backlog items that should be included in the next iteration or Timebox.

Since this is a team work, not an official risk analysis which has quantified with numbers, things could happen as the team missed one or more critical risks. If this does happen, the unidentified critical risks would soon show up in the next iteration. And the project manager should change the iteration plan accordingly.

5.2 Risks from People

Team members are sources of possible risks. Some of the risks can be identified and solved in the planning phase while others, although not directly included in the planning phase, have severely impact to the next iteration planning. Kieran Conboy identified 9 people risks [15]. We consider 5 of them directly or indirectly related with Agile project planning.

Developer fear of skill-deficiency exposure: Developers feared that agile process brings their own deficiencies to light. Whiteboard, which agile team used to communicate design issues or identify and analyze risks, requires direct and constant collaboration and communication. It highlights not only the technical challenges but also communication challenges because the developers must represent themselves in front of the team. The exposure of deficiency results in unhealthy work environment that some developers choose to change the team, leave the organization or become less confident. The solution would be to pair the junior guys with a mentor. The team work would cover individual deficiencies.

Broader skill sets for developers: In Agile environment, a developer is expected more than before. The developer is involved in risk management, plan drafting, customer communication, testing his own work in continuous integration. . .. Trainings are more difficult and cost than ever. Previously developers would receive trainings relevant to their own area only. In Agile environment, the organization needs to send the whole team to all kinds of training such as project management and customer communication. It is very expensive from both resource and money perspective.

Lack of business knowledge: Customers are directly involved in Agile development cycles in User Stories definition and feedback provision. Consequently the developer team has to be involved into on-site customer support. An absence of basic domain knowledge among developers has negative impact to customers. A manager mentioned that "If they don't know the business basic, the customers lose their confidence in overall ability, and their technical strength may be ignored." [15]. Provided that business training could be one of the solutions, hiring people with both technical and business background is another one. Knowing who should never be put in front of the customers is also very important to a successful manager.

Devolved decision-making: There are two folders in this risk. On the one hand developers are sharing much more tasks than before, while on the other hand the project manager might have a feeling of "loss power". Effective team decision-making includes a democratic voting system to ensure that everyone had input on every decision. The project manager is no longer the only decision maker but rather a facilitator. In Scrum model, project manager behaves like a "Scrum Master" to keep team focus on the right item in project planning meeting or daily Scrum meeting.

Implementing Agile-compliant performance evaluation: Traditional performance evaluation focuses on technical skills or technical contribution to the team. In the Agile work model, team members are more involved in team decision making. Especially for the senior developers they might spend a lot of time giving advice, pair with junior peers, and help team in stand-ups and retrospectives. Team cooperation work is hard to be evaluated and rewarded. To make performance evaluation effective and fair enough, 360-degree feedback can be applied, where all the team members evaluate each other to capture voluntary contributions and mentorship.

6 Conclusion and Future Work

Agile project management is an inherent requirement from the increasing turbulent and unpredictable market. Traditional project manager is no longer affordable and inefficient in dealing with frequently changed customer requirements. Various Agile project management methods have been developed in practice and literature. This paper focuses on the application of agile method in software project planning. The paper presents some key questions in Agile project planning and attempts to find the answers for them. The web technology and other IT technologies have been changing software project development and management in recent years. These changes push the urgent need of Agile project management to the stage of modern management technologies. This paper considers the difference between traditional project management and Agile project management, discusses how Agile project planning adapts to the market changes, and investigates how Agile project planning takes in account and resolves risks. The paper analyzes the essential factors in traditional project planning and Agile project planning, proposes that the scope is fixed in traditional project planning while the cost and time are fixed in Agile project planning. Also the paper develops how to make release planning, Timebox planning, and day planning in agile project management model. The combination of these planning methods defines the increment of development and inserts buffers for contingencies and uncertainties, which fully ensure the delivery of most important features to customer while keep alert on possible uncertainties. The possible risks from both technical aspect and people aspect are analyzed and the solutions to each kind of risks are provided.

Agile project planning is not limited to small-size project any more. It becomes a mandatory requirement in many software development organizations. Some of the organizations are international software companies which have multiple offices located in different countries. How to apply Agile project planning to large-scale software development? How to organize project planning meeting for team members across different time zones or physical locations? Addressing these questions would be our future research on Agile project planning.

References

1. Information Content Management, Social Media Timeline, retrieved from http://s2713275.wordpress.com/2012/08/11/social-media-timeline/
2. Nick Grantham, The History Of The Web – An Interactive Lesson, retrieved at http://www.fractuslearning.com/2012/01/30/history-of-the-web/
3. Sanjiv Augustin, Bob Payne, Fred Sencindiver & Susan Woodcock, Agile Project Management: Steering from the Edge, Communications of the ACM 48(12):85-89, 2005.
4. DeMarco, T, The Deadline: A Novel About Project Management, Dorset House, New York, 1997.
5. Linda Rising and Norman S. Janoff, The Scrum Software Development Process for Small Teams, IEEE Software 17(4):26-32, 1999.
6. Arthur English, Extreme Programming, Network World 19(4):60, 2002.

7. H. Frank Cervon, Understanding Agile Project Management Methods Using Scrum, OCLC Systems & Services: International Digital Library Perspectives 27(1):18-22, 2010.
8. Project Manager Institute, Inc., A Guide to the Project Management Body of Knowledge (PMBOK Guide), 2000.
9. Steven Thomas, Agile Project Planning, http://itsadeliverything.com/agile-project-planning, June 2008.
10. Cohn, Mike, Agile Estimating and Planning. Prentice Hall, 2006.
11. Wikipedia, June 2012, Agile Software Development Methodology, retrieved at http://en.wikipedia.org/wiki/File:Agile_Software_Development_methodology.jpg
12. Beck, K, & Fowler, M. 2001. Planning Extreme Programming. Boston: Addison-Wesley.
13. Goldratt, E.M., Critical Chain, North River Press, 1997.
14. Preston G. Smith, etc. 2005, Agile Risks/Agile Rewards, Software Development 13(4):50-53, 2005.
15. Kieran Conboy, etc. 2011. People over Process: Key Challenges in Agile Development, IEEE Software 28(4):48-57, 2010.
16. Martin Molhanec, Agile Project Management Framework, Proc. Of the 33rd International Spring Seminar on Electronics Technology, pp.525-530, 2009.

A Modeling Approach to Support Safety Assurance in the Automotive Domain

Yaping Luo, Mark van den Brand, Luc Engelen, and Martijn Klabbers

1 Introduction

In safety-critical domains, such as the automotive, railway, and avionics domain, manufacturers are expected to deliver continuously safe products. To deal with this, safety-critical systems are often required to undergo a stringent safety assessment procedure to show their compliance to one or more safety standards. Recently, safety standards in the automotive domain, such as ISO 26262 [1], have become widely used for safety assurance. ISO 26262 stimulates the usage of safety cases to demonstrate product safety [3]. A safety case is used for assuring the desired safety and showing the confidence in the claimed safety assurance. Therefore, it must be trustworthy and understandable. If the context of a safety case is not clear, it may affect its correctness. Currently, safety cases are expressed in natural languages, which are unavoidably ambiguous. This sometimes undermines the confidence in safety cases. Moreover, safety assessment or certification is amongst the most expensive and time-consuming tasks of safety engineering for modern systems. It becomes even more challenging when a system evolves and re-assessment is needed. Therefore, finding more efficient and cheaper methods for safety assessment is promising.

Modeling techniques are introduced to support safety assessment. Their contribution is to reduce manual work by creating machine processable and reusable models of the target domain like the automotive or the railway domain. By decreasing the amount of manual work, modeling techniques can support for reducing costs, as well as avoiding human mistakes. Furthermore, modeling techniques can be used to facilitate the development of safety argumentation and to increase the understandability and confidence in the claimed safety assurance [3].

However, there are a number of challenges: First, domain knowledge may be implicit. When domain experts work from their own experiences, the knowledge they have is implicit. It is a challenging task to make this implicit knowledge explicit. Second, a number of users, like safety engineers, safety assessors, can contribute to produce, assess or use the safety case. The larger the number of users of a safety case, the more different interpretations of it. The use of natural languages does not help to reduce the inconsistencies and ambiguities into the safety argumentation. This makes it difficult to avoid misunderstanding and mistakes, and to bridge the gaps between different users of a safety case.

Safety cases are documented in textual or graphical notations. For textual safety cases, the details of reasoning and safety argumentation can be captured in word or excel documents and presented in a purely linear format. The logic and structure of this textual form is not easy to see, which allows for inconsistencies and confusion [15]. This motivates the development of graphical notations, one of which is named Goal Structuring Notation (GSN) [14]. It aims to assist developing well-structured and well-reasoned safety arguments by using some basic elements, like goals, strategies, and evidence. Furthermore, safety argument patterns are introduced to effectively develop successful safety cases. Nevertheless, the statements used inside those GSN elements are still expressed in natural languages. To tackle the aforementioned challenges, GSN needs to be improved.

Based on our previous work, we developed a novel approach that guides system designers in establishing a sound relationship between conceptual models and safety cases. The major contribution of our approach is threefold. First, based on our rule-based approach, conceptual models of safety standards can be extracted, in which domain knowledge is well structured [9]. Second, our approach makes use of Semantics of Business Vocabulary and

Y. Luo (✉) • M. van den Brand • L. Engelen • M. Klabbers
Eindhoven University of Technology, 513, 5600 MB Eindhoven, The Netherlands
e-mail: y.luo2@tue.nl; m.g.j.v.d.brand@tue.nl; l.j.p.engelen@tue.nl; m.d.klabbers@tue.nl

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_50,

Business Rules (SBVR) [12], which enables us to overcome the syntactic inconsistencies and semantic ambiguities involved in natural language expressions. The extracted conceptual models will be described in SBVR format. Later, it could be used to express the safety cases in a clear and controlled way. As far as we know, this is the first proposal to provide SBVR expressions for safety cases. Third, we developed a tool to support safety case construction. It is able to detect the concepts from SBVR vocabulary, and support syntax highlighting and content assistance.

2 Background Information

We describe in this section the basic information used in the remainder of this paper. We give a short description of safety cases, and then we introduce SBVR and identify the applied terminology.

2.1 Safety Case and GSN

A safety case is a structured and reasoning argument, connecting claims to a body of evidence. In ISO 26262, the following definition of a safety case is given: "argument that the safety requirements for an item are complete and satisfied by evidence compiled from work products of the safety activities during development." [1] It is used to show that a system, service or organization will operate as intended for a defined application in a defined environment. In the safety domain, a safety case is a structured argument, supported by evidence, intended to justify that a system is acceptably safe. As noted earlier, a safety case must be represented both correctly and understandably. Thus, the reasoning structure of a safety case should be carefully developed and defined.

Based on these observations, the Goal Structuring Notation has been proposed for representing the argument structure [14]. It provides a number of graphical symbols to assist the development of safety cases. Some GSN notations are shown in Figure 1.

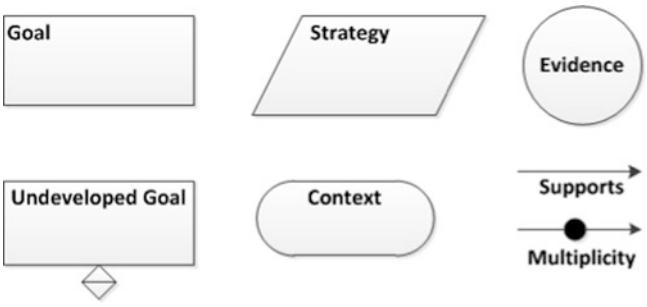


Fig. 1 A simplified illustration of GSN elements

2.2 Semantics of Business Vocabulary and Business Rules

Semantics of Business Vocabulary and Business Rules (SBVR) is a standard business-focused specification proposed by the Object Management Group (OMG) in 2008. It is designed for domain experts to capture business rules in a formal, structured and understandable language. It consists of a highly flexible structure that can capture and define most of the subtle intricacies of natural language. Recently, OMG published a second version (SBVR 1.1). The SBVR specification defines, among others, a metamodel to develop models of business vocabulary (comparable to conceptual models) and business rules (comparable to constraints that should be enforced). The definitions of some main concepts in SBVR specifications are listed as follows [13]:

Vocabulary	set of noun concepts, verb concepts, as well as various specialized concepts such as categorizations.
Concept	unit of knowledge created by a unique combination of characteristics.
Rule	proposition that is a claim of obligation or of necessity.
Business rule	rule with business focus.

In this paper, all SBVR examples are given in SBVR Structured English (SSE), which is introduced in SBVR Annex C. There are four font styles with formal meaning in SSE, which are shown in Table 1. In our implementation, for the font style of *Name*, we use the same font style as *Term*.

There are two key elements of SBVR meanings: business vocabulary and business rules. A business vocabulary defines concepts and their representations (designation, definition or statement). The concepts consist of noun concepts and verb concepts. Business rules provide elements of

Table 1 Font styles and Color with formal meaning in SBVR structured english

SSE Concepts	Font Style	Color	Denotes
Term	Underlined	Green	Noun concepts
Name	Double underlined	Green	Individual concepts
Verb	Italic	Blue	Verb concepts
Keyword	Normal	Orange	Other linguistic symbols used for definitions and statements

guidance on business structure and actions. SBVR defines deontic and alethic modalities for the formulations of guidance. The deontic modal operators describe behavioral (operative rules), which specify expectations of humans or automated systems; for example: for example:**it is obligatory that each functional safety requirement *has* exactly one ASIL** (Automotive Safety Integrity Level). Alethic modal operators enable definitional structural rules, which define features of a model, thus cannot be violated, such as: **it is possible that each functional safety requirement *is derived from* at least one hazard**.

3 Methodology

We propose a model-based approach for safety standard compliance to facilitate the reuse in certification processes. It can be divided into three phases: the conceptual phase, the vocabulary phase, and the modeling phase. First, in the conceptual phase, a conceptual model of the target domain will be built, which can be a conceptual model of a safety standard or a concrete development project. We introduce a rule-based approach for extracting conceptual models from safety standards or existing design documents of projects. After this, we explain how we transform these conceptual models into SBVR vocabulary in the vocabulary phase, where all the domain concepts will be explicitly defined. Finally, in the modeling phase, certification data, by means of safety argumentation, is structured into safety cases represented in GSN. By using the SBVR vocabulary, a link to conceptual models will be added to those safety cases to support assessing and certifying systems. The first two phases are typically done by standard experts, safety assessors or project managers; the last phase is typically done by safety managers or safety engineers.

3.1 Conceptual Phase: From the Domain to Conceptual Model

The conceptual phase is a preparation phase for a safety case. To make the domain knowledge explicit and develop a common understanding, we propose the development of a conceptual model. A conceptual model could represent the main terms and their relations that need to be considered for safety cases. As safety standards are widely used in the certification domain, we presented a rule-based approach [9], called Snowball approach, to facilitate modeling the standards. The approach can be used by modeling experts, but must be validated by the standard or domain experts.

Just like creating a snowman, our Snowball approach involves four steps: First, a basic model with a number of concepts is created from the high-level requirements of a safety standard, such as objectives and goals, or some basic documents of a development project. Second, like rolling a snowball in the snow, the size of basic model becomes bigger and bigger when processing the low-level requirements of the safety standard or other documents of the target project according to the rules. Third, the big snowball is shaped into a "snowman" frame, which is the conceptual model of the standard or the project and preliminary model for practical use. Finally, the snowman frame turns to a real snowman after being validated by domain experts. The snowman here is the conceptual model that we need for the next phase.

3.2 Vocabulary Phase: From Conceptual Model to SBVR Model

The conceptual models that result from our Snowball approach are stored in Eclipse Modeling Framework (EMF) format. To reduce the inconsistencies and ambiguities involved in safety cases development, SBVR vocabulary is introduced for describing conceptual models. Thus a conceptual mapping from EMF concepts to SBVR concepts is defined in Table 2.

As an example, Figure 2 illustrates a simple EMF schema, which is derived from an industrial ISO 26262

Table 2 Conceptual mapping between EMF and SBVR elements

EMF Concepts	SBVR concepts
Class	General concept
Enumeration Literal	Individual concept
Attribute	Characteristic
Association	Association verb concept
Generalization/Enumeration	Categorization

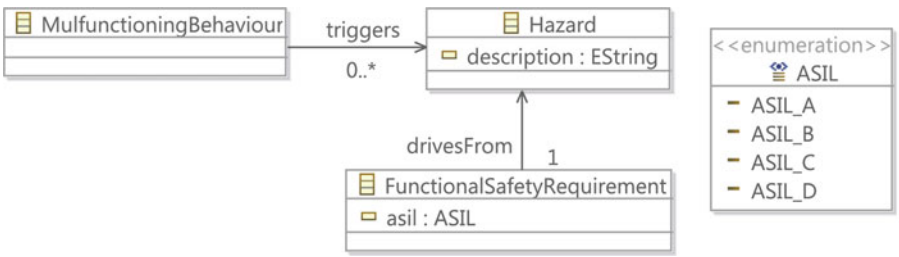


Fig. 2 A part of ISO 26262 metamodel in EMF

metamodels presented in [2]. It represents the relations between *Malfunction Behavior*, *Hazard*, and *Functional Safety Requirement*. Besides it includes an enumeration type, *ASIL*. After we apply the aforementioned mapping rules to this EMF schema, an SBVR representation can be obtained. Some of the results are presented in Table 3. The resulting SBVR representation is an instance of the SBVR metamodel. The main elements we could get from it are:

- Three classes in the EMF model of Figure 2 (*Malfunction Behavior*, *Hazard*, and *Functional Safety Requirement*) are mapped to three instances of *General Concept*.
- One instance of Categorization type represents the enumeration *ASIL*.
- Two associations(*Malfunction Behavior* and *Hazard*) in the EMF model are mapped to two instances of SBVR association. In these SBVR associations, the classes are also identified as verb concept roles.
- Four enumeration literals of *ASIL* are mapped to four individual concepts (*ASIL A*, *ASIL B*, *ASIL C* and *ASIL D*)
- Two instances of *Property Association* are derived from the attributes types in EMF, one for *Hazard* (*Hazard* has *description*), one for *Functional Safety Requirement* (*Functional Safety Requirement* has *ASIL*).

Table 3 Comparison of EMF expression and SBVR expression

EMF expression	SBVR expression
EClass FunctionalSafety Requirement	functional safety requirement Concept type: **general concept** Necessity: **Each functional safety requirement** ***has*** **exactly one ASIL**
EEnumeration ASIL	**ASIL** Concept type: **categorization**
EEnumeration literal ASIL A	**ASIL A** Concept type: **individual concept** General type: **ASIL**
EReference isDrivedFrom	**functional safety requirement** ***is derived from*** **hazard** Concept type: **association** Necessity: **Each functional safety requirement** ***is derived from*** **at least one hazard**
EAttribute description	**description** Concept type: **role**

3.3 Modeling Phase: Integrating SBVR Model to Safety Cases

In the modeling phase, the safety argument will be developed using GSN elements. We propose to use SBVR to express the content of each element. For example, if in our safety case, there is a claim: "All functions are independent". This claim could be expressed as "**All functions *are independent***" by using an SBVR vocabulary, which is created based on the aforementioned SBVR metamodel. Alternatively, we could add a modal operator to make the claim even more explicit. Then, it becomes: "**It is obligatory that all functions *are independent***". In this way, the ambiguity of safety cases can be reduced. Users can always look into the vocabulary to check the definitions of nouns and verbs used in their safety cases to avoid misunderstanding.

4 Results

An example of a safety case for a power window system is presented here. A power window is a window in a car that can be opened and closed by pressing a button. The safety case of the power window provided for this case study is shown in Figure 3.

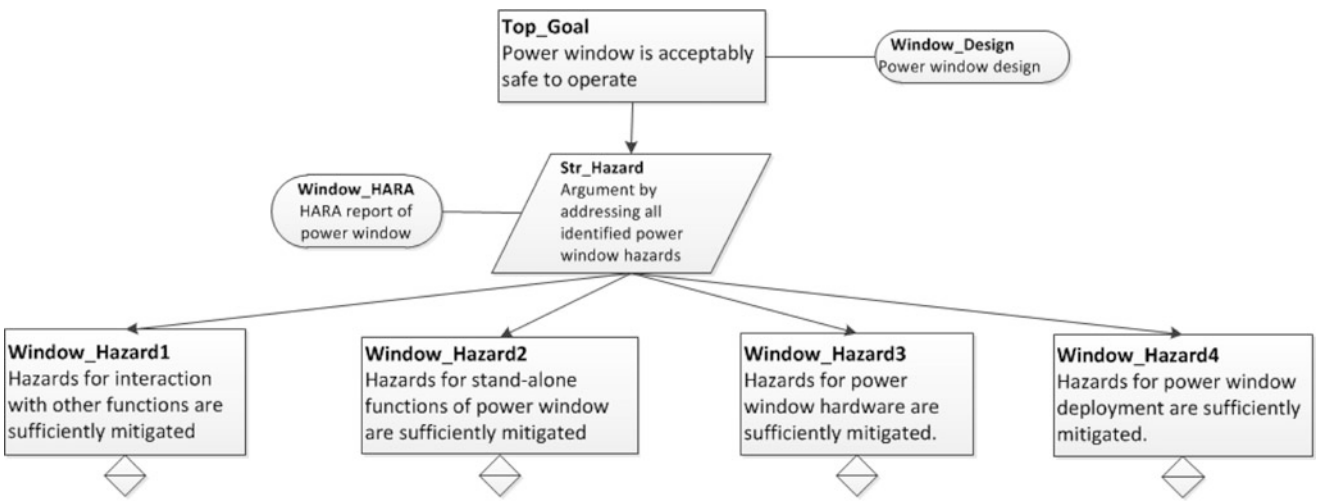


Fig. 3 Extract of the original safety case of power window

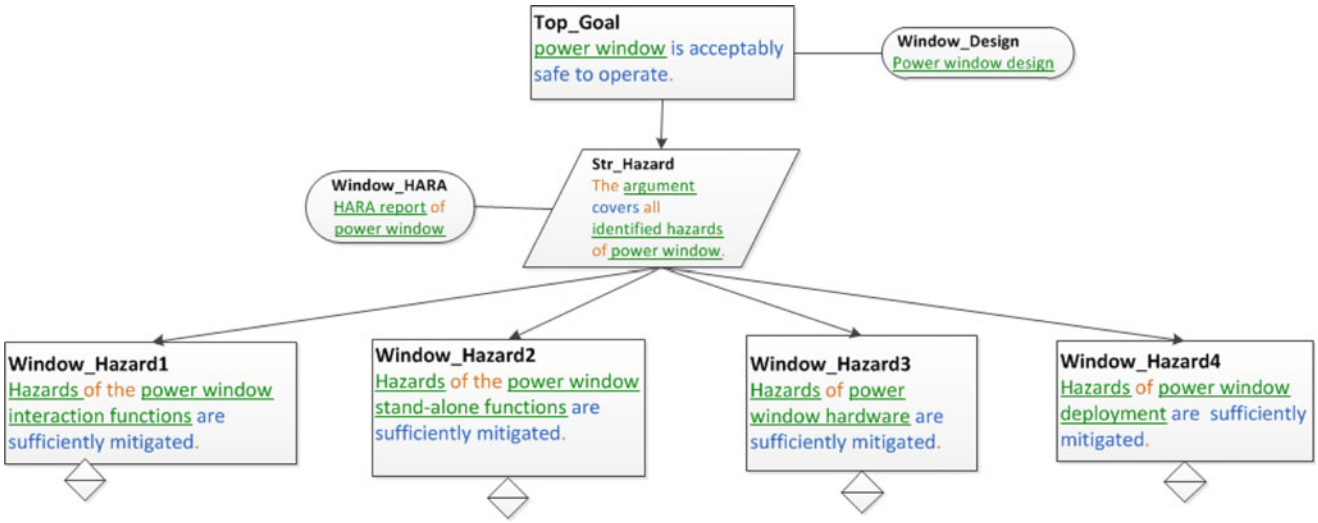


Fig. 4 A part of power window safety case expressed in SBVR

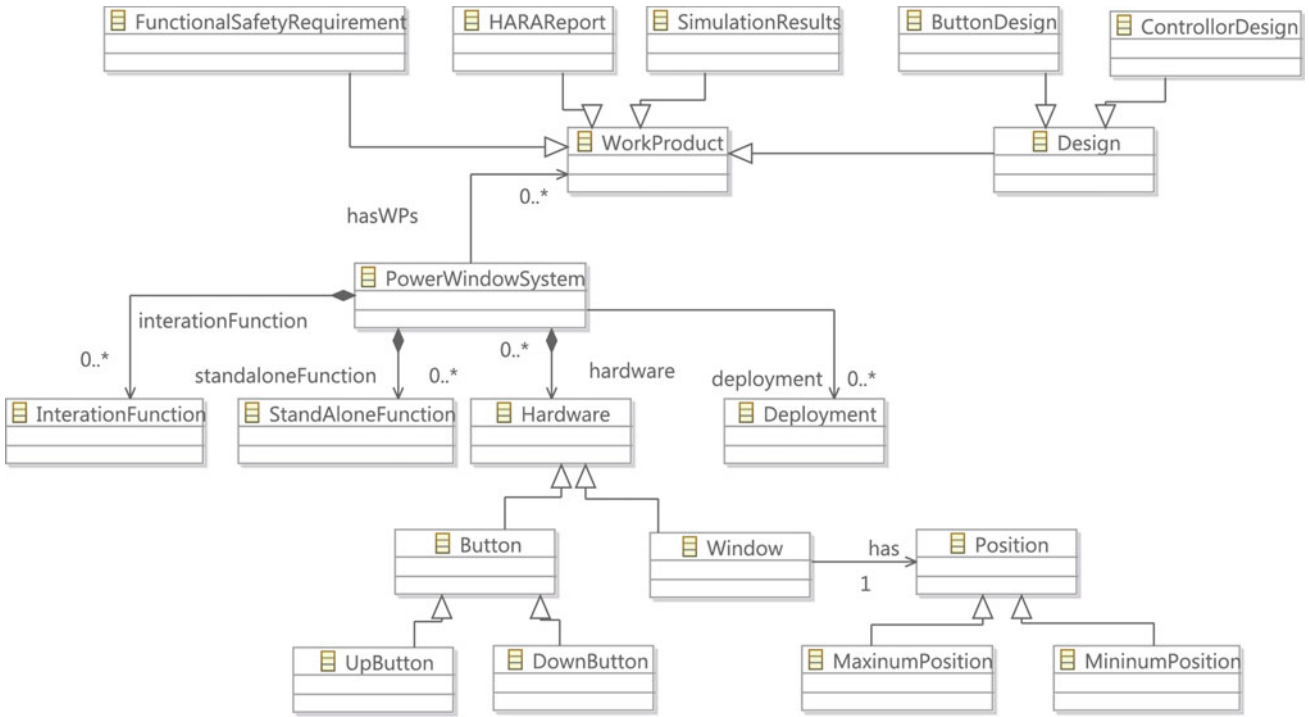


Fig. 5 Conceptual model of power window

Figure 4 illustrates the GSN and SBVR representation of the above safety case. From these two figures, we can see that some descriptions of safety case elements have been modified, for example the description of the "Window hazard1" claim. In the original claim, "interaction with other functions" could mean interaction between functions of the power window or other parts of the power window system. Due to this ambiguity, the confidence in this claim is low. In Figure 4, it is changed to "power window interaction functions", which is defined in our SBVR vocabulary with detailed information. In this way, the ambiguity in the original sentence is reduced, and the confidence of this claim is improved.

After comparing the safety case to the conceptual model of the power window shown in Figure 5, some results are obtained. First, we found that most concepts in the power window conceptual model are used in the safety case. Second, some project-related concepts in the safety case

need to be added by domain experts, such as project managers. In our case, for example, “direction”, “pressed button” and “limit” are introduced into our vocabulary as extra noun concepts for the power window project; “moves in”, “is pressed”, “moves up”, “moves down”, “keeps”, “exceeds” are added as extra verb concepts for the power window project. Finally, some safety related concepts need to be introduced by safety case experts, such as safety assessors or safety managers. In our case, concepts such as “argument”, “identified hazards”, “identified functional hazards”, “functional safety requirements” are added as safety related noun concepts. The concepts “is acceptably safe”, “covers”, and “are sufficiently mitigated” are introduced as safety related verb concepts.

Based on our metamodeling approach, the concepts used in safety cases could be categorized according to different projects or standards. Through the links between conceptual models and safety cases, information like definitions, requirements, and related concepts could be obtained. If safety cases need to be reused in a different context, the concepts could be detected and replaced easily. To facilitate the application of our method we have developed a tool framework, which provides syntax highlighting and content assistance. The input of this tool is an SBVR vocabulary, which includes the keywords for noun concepts and verb concepts used in safety cases.

5 Related Works

Since the SBVR was proposed by OMG, many approaches based on it have been applied to facilitate software engineering. Most of those approaches focus on transforming business rules or requirements written in natural language to SBVR rules through Natural Language Processing [7] [11] [5]. Some researchers worked on generating UML class models or execution models from SBVR models [4] [10]. Moreover, a study of translating UML schemas to SBVR vocabularies has been carried out [6].

Lewis describes the benefits of applying information modeling techniques to the development of an electronic safety case [8]. Furthermore, the relationships within the safety case have been studied. The domain knowledge is implicitly included in the informal relationships within the safety case.

6 Conclusions and Future Work

In this paper, we presented a methodology for safety assurance in the automotive domain. Our approach offers three key benefits:

1. It provides an objective approach to capture the crucial concepts and relations used in a target safety standard, company or project by keeping track of the relationships between the resulting conceptual model and input data.
2. It enables the explicit connection between conceptual models and safety cases to ensure that certification data is built properly and can be reused efficiently.
3. By utilizing SBVR, the content of safety case elements is well-structured and well-controlled. It can reduce mistakes and misunderstanding between the different roles involved in producing, assessing, and using the safety case.

Thus, our method supports improving the clarity and correctness of safety cases, and increasing the confidence in the claimed safety assurance. As on-going research, it has some limitations. First, domain experts are needed for validation of conceptual models. Second, most of the current work is done manually. Finally, the editor for our approach is still under development.

As future work, we plan to explore techniques to diminish the amount of manual work in our approach and to extend our methodology for reuse of certification data across domains. Finally, we aim to cooperate with industrial partners to apply our approach to large scale applications.

Acknowledgements The research leading to these results has received funding from the FP7 programme under grant agreement n° 289011 (OPENCOSS).

References

1. ISO 26262: “Road Vehicles – Functional Safety” (2011)
2. Meta Modeling Approach to Safety Standard for Consumer Devices (2013), http://www.omg.org/news/meetings/tc/agendas/ut/SysA_Slides/taguchi.pdf
3. Safety Case Repository (2013), http://dependability.cs.virginia.edu/info/Safety Cases:Repository
4. Afreen, H., Bajwa, I., Bordbar, B.: SBVR2UML: A Challenging Transformation. In: Frontiers of Information Technology (FIT), 2011. pp. 33–38 (2011)
5. Bajwa, I.S., G. Lee, M., Bordbar, B.: SBVR Business Rules Generation from Natural Language Specification. In: AAAI 2011 Spring Symposium - AI for Business Agility. pp. 2–8. San Francisco, USA (2011)
6. Cabot, J., Pau, R., Raventós, R.: From UML/OCL to {SBVR} specifications: A challenging transformation. Information Systems 35(4), 417–440 (2010)
7. Ceponiene, L., Nemuraite, L., Vedrickas, G.: Semantic Business Rules in Service Oriented Development of Information Systems. In: 15th International Conference on Information and Software Technologies, IT. pp. 404–416 (2009)
8. Lewis, R.: Safety Case Development as an Information Modelling Problem. In: Dale, C., Anderson, T. (eds.) Safety-Critical Systems: Problems, Process and Practice, pp. 183–193. Springer London (2009)
9. Luo, Y., Van den Brand, M., Engelen, L., M. Favaro, J., Klabbers, M., Sartori, G.: Extracting models from iso 26262 for reusable safety

assurance. In: Safe and Secure Software Reuse - 13th International Conference on Software Reuse. vol. 7925, pp. 192–207. Springer Berlin Heidelberg (2013)
10. Nemuraite, L., Skersys, T., Sukys, A., Sinkevicius, E., Ablonskis, L.: VETIS tool for editing and transforming SBVR business vocabularies and business rules into UML&OCL models. In: 16th International Conference on Information and Software Technologies, Kaunas: Kaunas University of Technology. pp. 377–384 (2010)
11. Njonko, P., El Abed, W.: From Natural Language Business Requirements to Executable Models via SBVR. In: Systems and Informatics (ICSAI), 2012 International Conference on. pp. 2453–2457 (2012)
12. OMG: SBVR: Semantics Of Business Vocabulary And Rules (Sepetember 2013), http://www.omg.org/spec/SBVR/1.1
13. Spreeuwenberg, S., Healy, K.A.: SBVR's Approach to Controlled Natural Language. In: Proceedings of the 2009 conference on Controlled natural language. pp. 155–169. CNL'09, Springer-Verlag, Berlin, Heidelberg (2010)
14. T.Kelly: Arguing Safety - A Systematic Approach to Managing Safety Cases. Ph.D. thesis, University Of York (1998)
15. Wilson, S., Kelly, T., McDermid, J.: Safety Case Development: Current Practice, Future Prospects. In: Shaw, R. (ed.) Safety and Reliability of Software Based Systems, pp. 135–156. Springer London (1997)

Dynamic OD transit matrix estimation: formulation and model-building environment

Lídia Montero, Esteve Codina, and Jaume Barceló

1 Motivation

In the context of estimating private transport demand, Origin-to-Destination (OD) trip matrices describe the number of trips between any origin-destination pair of transportation zones in a study area. For private vehicles, route choice models describe how trips select the available paths between origins and destinations and, as a consequence, the number of trips using a given path (or path flow proportions) in private transportation modes. The route choice proportion can vary depending on the time-interval in dynamic models, since the traffic state and the temporal dimension are considered. When a public transportation network is the object of study, OD matrices describe the number of transit trips between OD pairs or OD stops.

While an average OD table for a whole period of interest is acceptable for an urban transportation planning study, OD matrices over consecutive time intervals are required for modeling and/or optimizing dynamic system operations. For all formulations of static traffic or transit assignment models (Florian and Hearn [1]), as well as dynamic models involved in ATIS (Advanced Transport Information Systems) (see Ashok et al. [2]), they assume that a reliable estimate of an OD matrix is available and constitutes an essential input for describing the demand in predicting traffic state evolution over the network. Since OD trips are not directly observable, indirect estimation methods have then been proposed. These are the so-called matrix adjustment methods, whose main modeling hypothesis can be stated for a transit networks as follows: if assigning an OD transit matrix to a network defines the number of passengers in all segments of transit line itineraries sharing a network link, then the same OD transit matrix can be estimated as the inverse of the assignment problem.

ICT (Information and Communication Technologies) sensors can also provide data for estimating dynamic OD transit matrices, though this has received little attention in the literature because of the difficulties in collecting real-time passenger loads. However, with the new technologies, applications for transport planning have been proposed by several authors ([3, 4]), with a particular focus on transit OD inference.

Wong and Tong [5] proposed a maximum entropy estimator, employing the schedule-based approach for dynamic transit assignment. Ren [6] proposed a generalized least squares bi-level approach for estimating time-dependent transit matrices in congested schedule-based transit networks where automatic passenger counts and prior OD matrices are available. The proposal was tested on a toy network, but no recent works from the author have confirmed any offline or online applicability to large scale transit networks.

Real-time transport information systems or traffic management applications have an additional requirement for estimating dynamic OD matrices: a short time response (less than 15 min) for estimating and for forecasting the dynamic matrix and network estate over the next 30 min horizon.

Almost 20 years ago, memory space was expensive and unavailable on ordinary laptop computers. Because of this, linear Kalman filtering (KF) approaches for estimating dynamic trip matrices were considered inefficient, and unable to satisfy the requirements of on-line applicability, as indicated by Wu [7]. A linear KF prototype, coded in MATLAB [9] (Barceló et al [8,10]), has proven to meet the requirements for on-line applications using traditional data collection and new ICT data.

Kostakos *et al* [3] proposed the use of passengers' Bluetooth mobile devices to derive passenger OD matrices in a simplified context. A Bluetooth device set to *discoverable mode* must respond to a discovery request by

L. Montero (✉) • E. Codina • J. Barceló
Dept Statistics and Operations Research-BarcelonaTECH-UPC,
Carrer Jordi Girona 1-3, 08034 Barcelona, Spain
e-mail: lidia.montero@upc.edu; esteve.codina@upc.edu;
Jaume.barcelo@upc.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_51,

transmitting its unique Bluetooth identifier (12 hex digits) and device class (6 hex digits), since a Bluetooth scanner located on vehicle units constantly scans the presence of the various devices it encounters (along with the date and time). Extending this idea, our aim is to explore the possibility of using the experience in private networks to estimate on-line dynamic OD transit matrices. To do this, we use ICT data, counts and travel times, provided by a sample of passengers with Smartphones and BT/Wifi signals are captured by Wi-Fi antennas, which are located in a subset of transit stops. A *prior* historic OD transit matrices estimated off-line is also used.

The remainder of this paper is structured as follows. A section with the previous experience of the authors is presented, followed by a description of the model formulation for estimating dynamic OD matrices in transit networks. The model-building environment for validation is outlined next. The paper concludes with a discussion and concluding statements.

2 Previous experience

In [8], we used counts and travel times provided by Bluetooth devices to estimate dynamic OD matrices for commuters driving their own cars, either on freeways or in urban networks. In that work, we proposed space-state formulations by means of linear Kalman filters (KF). For linear freeways, traditional and ICT data at entry ramps were required. Bluetooth equipped vehicles were identified by antennas (in the following ICT sensors) located on some freeway sections. Equipped vehicles provided accurate travel time data from their entry points up to the freeway points they went through, where they were detected by ICT sensors.

A prevailing basic hypothesis is that equipped and non-equipped vehicles follow common OD patterns. A common dynamic horizon was defined to be 1.5 to 3 hours, and it was divided into time-intervals of between 1 to 5 minutes in length. We adapted an idea proposed by Lin and Chang [11] and proposed our own linear KF formulation that relates state variables and observations. Congestion effects of traffic dispersion were taken into consideration, since time-varying model parameters were derived from travel times provided by the sample of equipped vehicles.

We proposed a flexible formulation suitable for urban networks, and also for freeways as a particular case of a network, where state variables were defined as deviates (Ashok and Ben-Akiva [2]) of the number of OD equipped vehicles for intervals along the *most-likely OD paths* with respect to *a priori* historic values. Dynamic User Equilibrium (DUE) based on the historic time-sliced OD matrix was computed to define the set of *most likely OD paths*. The proportion of OD flows using the selected paths was not taken as input, as it is considered indirectly by some authors through assignment matrices. The measurements were: the link flow counts on traditional detectors, link flow counts of equipped vehicles in a subset of links where ICT sensors were located, and travel times of equipped vehicles between ICT sensors.

We achieved a linear KF formulation dealing with congestion in freeways and networks. Our approach exploited ICT travel time measurements from equipped vehicles in order to estimate discrete travel time distribution (H bins were used for adaptive approximations). Travel times collected from ICT sensors were incorporated into the proposed model [8]. Throughout this paper, they are referred as time-varying model parameters. It was not necessary that vehicles reach their destination, since the measured travel times from any previously crossed ICT sensor updated the discrete travel time approximations at any intermediate sensor that they passed through. Therefore, completed trips were not the only source for updating time-varying model parameters in the program KFX3 implemented in MATLAB [9].

2.1 Proposed framework for time-sliced OD transit estimation

The research experience gained for estimating OD trip matrices and the extensive computational experience [8, 10] of simulating data determined the crucial design factors to be considered. The computational results reported that it is possible to apply a linear KF approach for estimating dynamic OD trip matrices in real-time applications, since a half-hour forecast takes less than 2 min using MATLAB [9].

The conclusions about the key design factors taken into account were:

- The penetration rate of ICT is a critical factor, but not controllable.
- The detection layout defined in terms of the number and location of detectors, which is in practice frequently determined by budget constraints.
- The quality of the historic time-sliced OD matrix used to initialize the KF algorithm and *a priori* covariance matrices involved in the KF process.

Thus, we conclude that, one cannot skip off-line estimation of a reliable historic (time-sliced) OD matrix when defining an on-line framework. So far, this lesson is considered in the framework for estimating the time-sliced OD transit matrices shown in Fig. 1.

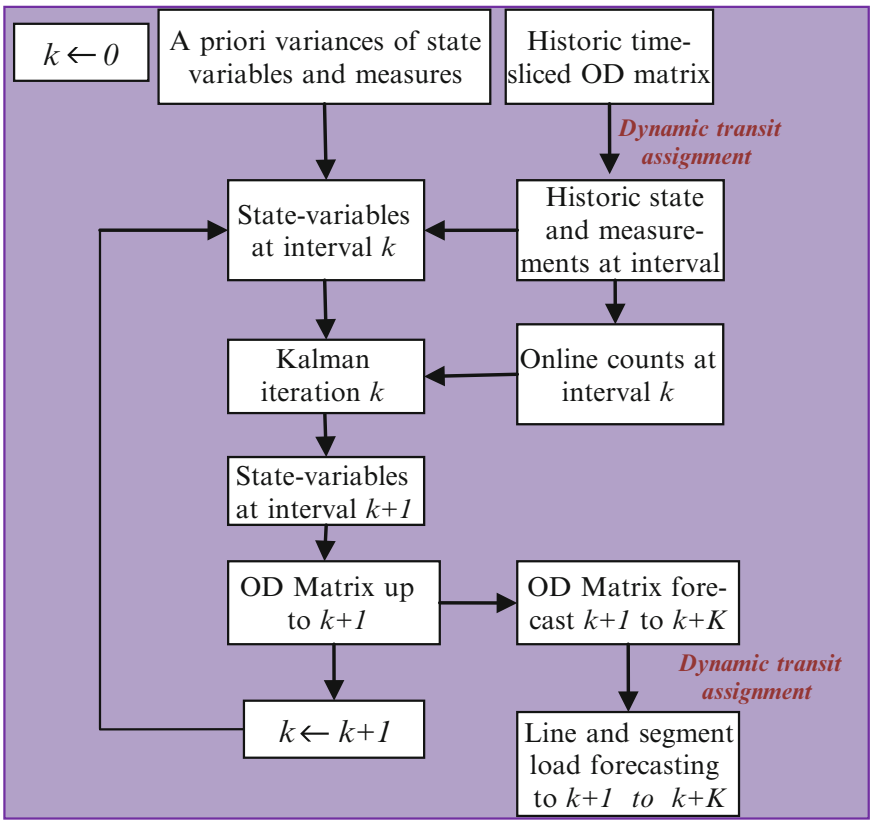


Fig. 1 Framework for Estimating and Predicting in ATIS

Table 1 Definition of model variables

$\tilde{Q}_l(k)$, $\tilde{q}_l(k)$	:	Historic total number of passengers and equipped passengers accessing a transit unit and first detected at stop r, related to centroid i in flow conservation l at time interval k.
$Q_l(k)$, $q_l(k)$	:	Total number of passengers and equipped passengers accessing a transit unit and first detected at stop r, related to centroid i in flow conservation l at time interval k.
$\tilde{y}_q(k)$, $y_q(k)$	:	Historic and actual number of equipped passengers crossing at time interval k either the q sensor or destination sensor s, *for a pair of ICT transit stops* $q = (r,s)$
$G_{ije}(k)$, $\tilde{G}_{ije}(k)$ $g_{ije}(k)$, $\tilde{g}_{ije}(k)$	:	Total number of current $G_{ije}(k)$ and historic $\tilde{G}_{ije}(k)$ passengers, as well as the current $g_{ije}(k)$ and historic $\tilde{g}_{ije}(k)$ equipped passengers accessing centroid i at time interval k and heading towards j using path e.
$\Delta g_{ije}(k)$	:	State variables are deviates of equipped passengers from centroid i during interval k headed towards centroid j using path e with respect to average historic data $\Delta g_{ije}(k) = g_{ije}(k) - \tilde{g}_{ije}(k)$.
$z(k)$, $\tilde{z}(k)$	:	The current and historic measurements of equipped passengers during interval k, a column vector of dimension Q (counts) plus L (flow conservation equations).
$u_{rs}^{h}(k)$	:	Fraction of equipped passengers that require h time intervals to reach sensor s at time interval k *from sensor* r during time interval $[(k-h-1)\Delta t, \quad (k-h)\Delta t]$. Time-varying model parameters.
$\bar{t}_{rs}(k)$	:	Average measured travel time for equipped passengers that reach sensor s at time interval k *from sensor* r

3 Formulation of the model

Notation is defined in Table 1. Some additional aspects to the data model and formulation that have to be considered are:

- The demand matrix for the period of study is assumed to be divided into several time-slices, accounting for different proportions of the total number of transit trips in the time horizon.
- The approach assumes an extended space state variable for M + 1 sequential time intervals of equal length Δt (between 5 and 10 min), in order to consider non-instantaneous travel times. M should guarantee traversal of the network.
- Bluetooth antennas are ICT sensors assumed to be located at (some) transit-stops.
- OD paths are those involved in optimal strategies for OD pairs for the period of study. They can be computed by any transportation planning software including a strategy-

based equilibrium transit assignment using historic demand. We do not have a strategy-based dynamic transit assignment tool available, so we considered EMME [12] to define state-variables in our tests. The description of paths involved in optimal strategies from centroid i to j going through ICT sensors (pairs of ICT sensors) can be systematically programmed in any language (although this is not trivial), from the EMME output to the input files for the MATLAB data model.

The total number of origin and/or destination centroids (related to transportation zones) is I; the total number of ICT sensors is P, located either at bus-stops or at segments in the inner network; and the total number of paths corresponding to optimal transit strategies from the historic OD transit assignment for the period of study is K. Q is the number of pairs of ICT sensors *(r,s) plus* individual sensor counts q to be considered. We estimate OD transit trips between OD pairs, not between transit-stops.

The state variables $\Delta g_{ije}(k)$ are assumed to be stochastic in nature, and OD path flow deviates at the current time k are related to the OD path flow deviates of previous time intervals by implementing an autoregressive model of order r $<<$M; the state equations are:

$$\mathbf{\Delta g}(k+1) = \sum_{w=1}^{r} \mathbf{D}(w)\mathbf{\Delta g}(k-w+1) + \mathbf{w}(k), \qquad (1)$$

where $\mathbf{w}(k)$ is zero mean with diagonal covariance matrix $\mathbf{W_k}$, and $\mathbf{D}(w)$ are transition matrices which describe how previous OD deviates $\mathbf{\Delta g_{ije}}(k\text{-}w+1)$ affect current flows $\mathbf{\Delta g_{ije}}(k+1)$ for $w = 1,\ldots,r$. In our tests, we assume simple random walks to provide the most flexible framework for state variables, if no convergence problems are detected. Thus, our first trial is $r = 1$, and $\mathbf{D}(\mathbf{w})$ becomes the identity matrix. The relationship between the state variables and the measurements involves *time-varying model parameters* (congestion–dependent, since they are updated from sample travel times provided by equipped passengers) in a linear transformation that considers:

- The number of equipped passengers first detected in the system at equipped transit-stops r, related to origin zones through explicit flow conservation equations l during time intervals in $k, \ldots, k\text{-}M$, $q_l(k)$.
- H $<$ M *time*-varying *model parameters* in form of *fraction matrices*, $[u_{rs}^{h}(k)]$, where the H adaptive fractions are updated from measures provided by ICT sensors. Direct samples of travel times allow the updating of discrete approximations of travel time distributions.

At time interval k, the values of the observations are determined by those of the state variables at time intervals $k, k\text{-}1, \ldots, k\text{-}M$.

$$\mathbf{\Delta z}(k) = \mathbf{F}(k)\mathbf{\Delta g}(k) + \mathbf{v}(k), \qquad (2)$$

where $\mathbf{v}(k)$ is white Gaussian noise with covariance matrix $\mathbf{R_k}$. $\mathbf{F}(k)$ maps the state vector $\mathbf{\Delta g}(k)$ onto the current blocks of measurements at time interval k: counts of equipped passengers at ICT sensors, accounting for time lags and congestion effects.

The solution should provide estimations of the OD transit matrix between OD pairs for each time interval up to the *k-th* interval, once observations of equipped passengers at the transit stops equipped with Wi-Fi antennas up to the *k-th* interval are available. KF prediction of OD trips for ICT equipped passenger up to some intervals ahead must be considered and expanded upon in accordance with historic profiles (for day-type and time-period), in order to feed a dynamic transit assignment tool that will provide the forecasted travel times, line loads and boardings/alightings at transit-stops in the short-future. We consider a 30 min forecasting horizon.

4 Model Building environment

The formulation has been programmed as a MATLAB prototype (named KFX3T). Correct codification has been verified, but the approach needs to be validated against either actual data, which was not available, or by simulation, the chosen option.

In the research undertaken within the framework of the European Union COST Action MULTITUDE, a common evaluation and benchmarking platform was developed for estimating dynamic OD trip matrices [13]. The main goal of that platform was to provide a testbed in which a number of algorithms can be linked and tested under the same conditions in a common network model.

The MULTITUDE Platform consists of a Mesoscopic Traffic Simulator (AIMSUN [14]), one dynamic OD estimation algorithm coded in MATLAB [9], and a MATLAB function (AIMSUN.m). The MATLAB function allows dynamic communication between the estimation algorithm and the Aimsun model in order to execute a traffic simulation run within the OD estimation algorithm. It does this by creating and executing a batch file, which launches the Aimsun executable and a python script with the relevant information. When the simulation is finished, the MATLAB function collects all the outputs and produces goodness of fit indicators. A case study consisting of an old AIMSUN model for Vitoria-Gasteiz, a medium-size city in Spain, was provided to MULTITUDE partners.

We chose to include the KFX3T OD transit estimation tool and Vitoria case study in the MULTITUDE platform for validation purposes. Vitoria's model contains the description of a bus network. AIMSUN does not include public transport under the scope of the *transit-assignment* to the transit network, neither under a strategy-based nor a scheduled-based

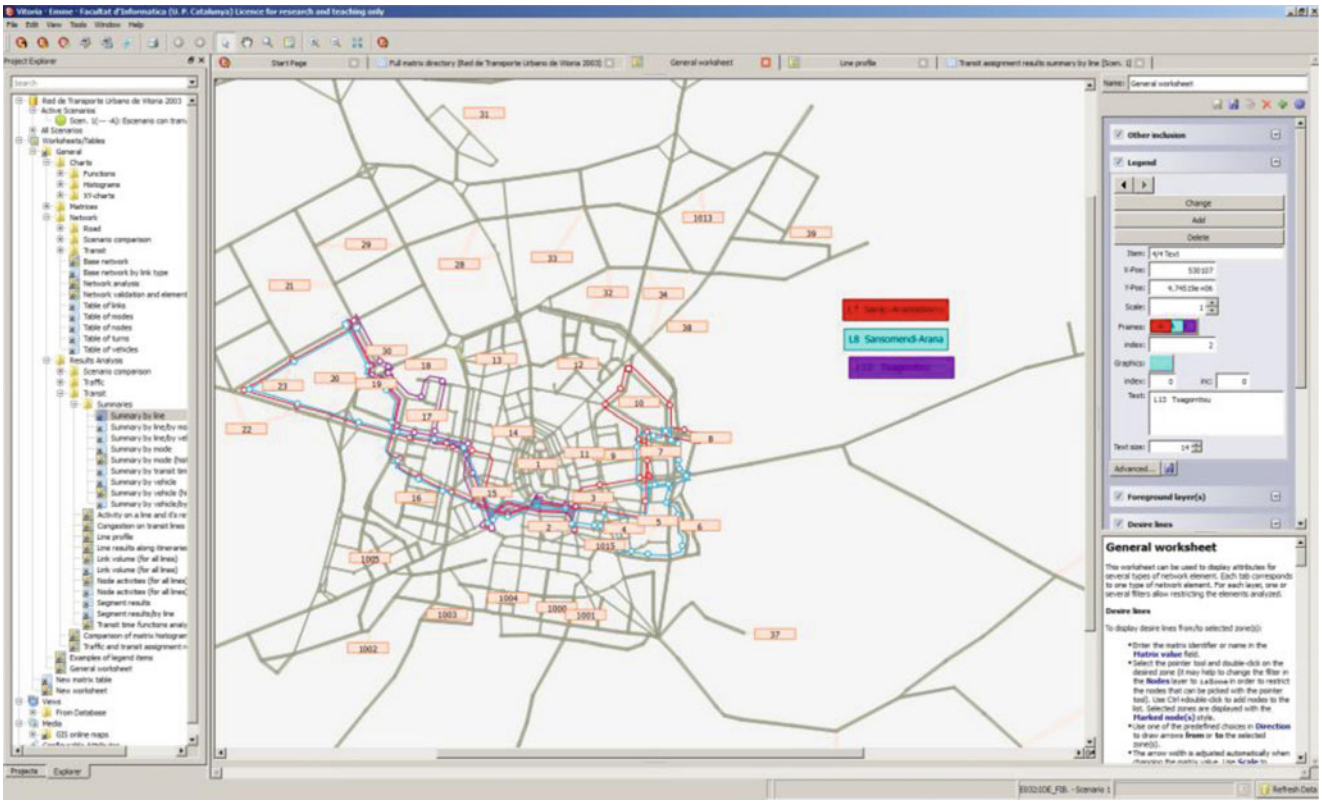


Fig. 2 Emme Model for Vitoria: BT located at transit-stops for lines L7, L8 and L13

behavioral hypothesis. We decided to consider EMME [12] instead of AIMSUN and we selected an EMME model that matches the MULTITUDE network. The total number of public transit trips was 47816.3 for a 2003 working day within 340 non-null cells, and an average value of 140 trips (max 1186 trips).We fixed ICT equipped transit-stops for three of the most important bus lines: L7, L8 and L13. The model and equipped transit lines are shown in Fig. 2**.** After selecting OD pairs with captured flows greater than 10, these lines have a total load (working day) of 20765 transit trips in 151 OD pairs (according to the additional options transit assignment results in EMME).

At this point, we modified the MULTITUDE platform and substitute the AIMSUN model for EMME in order to provide:

- A graphic description (centroids, nodes, links and connectors) and transit lines (stops, headways and itineraries).
- A Transport Zoning System and historic demand matrix for the period of study.
- The location of ICT sensors at transit-stops (identified by an additional segment attribute @nBT, in EMME terminology, and an additional line attribute, @sline, to identify lines affected by BT-equipped stops, either totally or partially).
- OD paths involved in optimal strategies for the historic transit demand according to an *Extended Transit Assignment*, where OD flows are split between connectors in the origin zone using a *logit* model (scale parameter 0.5) and are subject to an OD path proportion greater than 5 %.

A strategy file is saved by EMME and the *Extended Analysis tool* reports path-based details on paths extracted from strategies in a text file. This text file contains the path description included in the optimal strategies restricted to OD pairs and whose OD flow could be captured by some ICT sensors in accordance with historic demand and transit assignment. The path file has a complex format and is post-processed using python script to generate the desired information in the data model of KFX3T. For each ICT-equipped transit-stop, the captured optimal transit OD paths are identified. For each pair of ICT transit-stops, the paths and subpaths connecting both sensors are identified. Report files related to OD pairs, nodes, links, etc. were converted to worksheet and *.csv* format files, in order to be read by the KFX3T model building procedures. This led to:

- A network workbook containing worksheets for: centroids, nodes and links.
- A demand workbook containing worksheets for: OD pairs, Entrances, Exits and OD paths involved in optimal strategies.

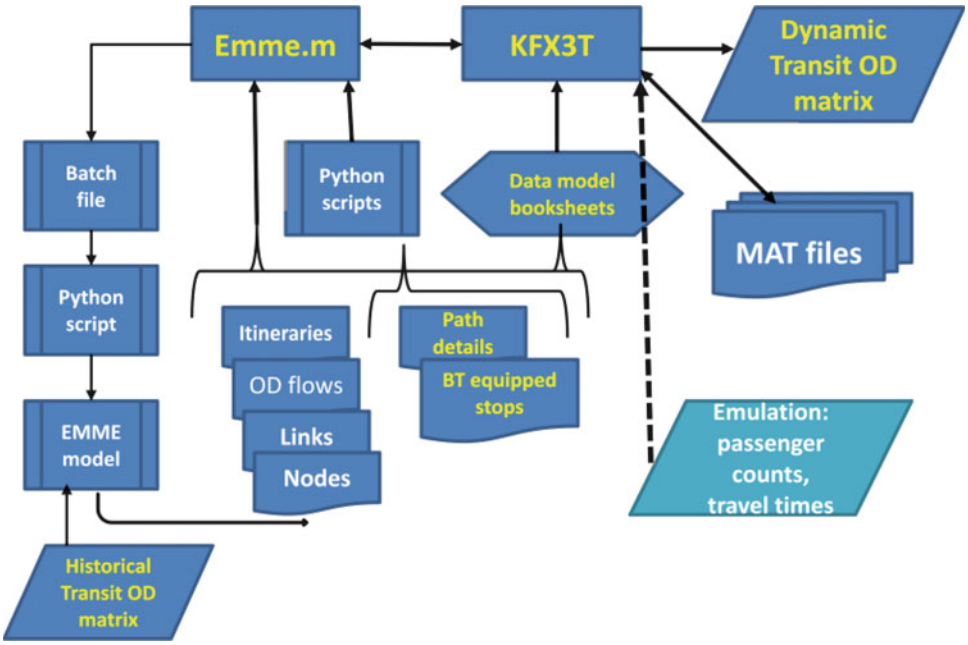


Fig. 3 Flowchart with the main elements of KFX3T in the MATLAB application

– A measurement workbook containing worksheets: Measures, SODMeasures, ActiveMeas, CapODPaths and Global (parameters).

The KFX3T internal data model is divided into MAT files that are loaded as needed into the program. They are: *Global.mat, Tuning.mat, Graph.mat, Demand.mat* and *Measures.mat*. Additionally, two additional MAT files have been included for internal use in order to simplify the access to some critical structures. These two files are: *AccDem.mat* (related to accessing OD pairs and paths) and *AccMes.mat* (related to OD paths captured by each defined sensor and pairs of ICT sensors).

Fig. 3 shows the flowcharts between the dynamic transit estimation KFX3F tool and the source of data for model building (MAT files). We have completed development of all software pieces, except for the simulation tool that emulates passenger counts at stops and travel times between pairs of ICT sensors. Once available, the estimation and forecasting tool (KFX3F) could be validated using Vitoria's network.

5 Conclusions

This paper has given an account of how to estimate Dynamic OD matrices for passengers in a real-time context, since they constitute the basic input for dynamic models involved in Advanced Traffic Management and Information Systems. The purpose of the present paper is to show the model-building process for a linear Kalman filter formulation developed by the authors. The framework implements a MATLAB program (KFX3T) which takes advantage of the EMME transportation planning environment through the definition of paths related to optimal strategy-based transit assignment.

Finally, the software prototype, KFX3T, has been validated in a limited way. Firstly, using a toy network and simulated data. Secondly, the mean squared error (MSE), which summarizes accuracy on average; is the default goodness of fit for testing the convergence. However, it would be interesting to assess the performance results of different measures in the future. Further experimental evidence is needed with a medium-size network. Computational results of a similar approach applied to private transport networks by the authors support the on-line applicability.

Acknowledgments The research was funded by TRA2011-27791-C03-02 of the Spanish R+D National Programs, and it benefited from EU COST Action TU0903 MULTITUDE.

Reference

1. Florian, M. Hearn, D.: Network Equilibrium Models and Algorithms. In: Chapter 6 in M.O. Ball et al., Eds. Handbooks in OR and MS, Vol.8, Elsevier Science B.V (1995).
2. Ashok, K. Ben-Akiva, M.: Alternative Approaches for Real-Time Estimation and Prediction of Time-Dependent Origin-Destination Flows. In: Transportation Science vol.34 no.1, pp. 21-36 (2000).
3. Kostakos, V., Camacho, T., Mantero, C. Wireless detection of end-to-end passenger trips on public transport buses. In Proceedings of the IEEE Conference on Intelligent Transportation Systems, Funchal, Portugal pp. 1795-1800 (2010).
4. Wang, W., Attanucci, J. P. and Wilson, N.H.M., Bus *Passenger Origin-Destination Estimation and Related Analyses Using*

Automated Data Collection Systems. In: Journal of Public Transportation, vol.14, no.4, pp.131-150 (2011).
5. Wong, S.C., Tong, C.O. Estimation of time-dependent Origin-Destination matrices for Transit networks. In: Transportation Research B vol.32, no.1, pp.35-48 (1998). Elsevier Science Ltd.
6. Ren, H.L. Origin-Destination Demands Estimation in Congested Dynamic Transit Networks. In Proceeding 14th International Conference on Management Science and Engineering, Harbin (2007).
7. Wu, J. In: A real-time origin-destination matrix updating algorithm for on-line applications. In: Transportation Research B vol. 31-5 pp.381-396 (1997).
8. Barceló, J., Montero, L., Bullejos, M., Serch, O., Carmona, C.: A Kalman Filter Approach for the Estimation of Time Dependent OD Matrices Exploiting Bluetooth Traffic Data Collection. In: GITS Journal of Intelligent Transportation Systems, vol. 17(2) pp.123-141 (2013).
9. MatLab (R2012a) by MathWorks, Inc. (www.mathworks.com) (2012).
10. Barceló, J., Montero, L., Bullejos, M., Linares, M.P., Serch, O. : Robustness And Computational Efficiency Of A Kalman Filter Estimator of Time Dependent OD Matrices. Transportation Research Record 2344-04 pp.31-39 (2013).
11. Lin, P., Chang, G. A generalized model and solution algorithm for estimation of the dynamic freeway origin-destination matrix. In: Transportation Research B vol.41 (2007) pp.554-572.
12. EMME 4.0.8: User's Manual. INRO, Montreal (www.inrosoftware.com) (2013).
13. Antoniou, C., Ciuffo, B., Montero, L., Casas, J., Barceló, J., Cipriani, E., Djukic, T., Marzano, V., Nigro, M., Bullejos, M., Perarnau, J., Breen, M., Punzo, V. and Toledo, T. A framework for the benchmarking of OD estimation and prediction algorithms. Presented at 93rd Transportation Research Board Annual Meeting (2014), Washingtong, D.C.
14. AIMSUN 7.0.4 Microscopic Simulator, TSS- Transport Simulation Systems, Barcelona (www.aimsun.com) (2013).

Microstrip Spiral Resonator for the UWB Chipless RFID Tag

A. K. M. Z. Hossain, S. M. A. Motakabber, and M. I. Ibrahimy

1 Introduction

Radio Frequency Identification (RFID) implies a brunch of the Automatic Identification system which uses the radio waves for identification [1]. In recent years there is tremendous flow in the research and development of RFID systems for replacing the conventional optical barcode identification systems. Though the optical barcode is low in cost but has many disadvantages compared to RFID system, such as risk of tampering; need line-of-sight to read, sensitive to wear and tear etc. [2, 3]. A reader (interrogator) and the tag (transponder) are the major parts of an RFID system. The reader and tag utilize a radio wave interface to communicate with each other. Based on the on board power supply there are three types of tag; active tag has on board battery power for its operation, semi-active tag has also on board battery powered but it harvest reader's signal energy to transmit the interrogation signal and the passive tag don't has on board battery and fully dependents on the reader's interrogation signal energy to communicate [4]. The passive tag is the simplest and cheapest one among the others tags. The chipless tags are inherently passive type tag in addition these types of tags do not contain any silicon chip, as a result these are at the forefront of finding low-cost solution. The RFID system has several dedicated ISM radio frequency bands for communicating with the tag and reader [5]. The recent need of high speed item tracking in the field of toll collection and surveillance system attracting the researchers to choose the Ultra Wide Band (UWB) frequency range (3.1 GHz-10.7 GHz) defined by FCC [6]. This communication band is still under research and not yet fully deployed commercially. In recent years, several techniques are proposed to detect the UWB chipless tags: Surface Acoustic Wave (SAW) tag, Continuous Wavelet Transform (CWT) based tag and Microstrip Resonator tags [6]-[8]. All the techniques are based on the backscattered modulation, where the tag antenna receives the interrogation signal from the reader, changes specific properties of the signal and reflects back to the reader for identification.

A microstrip resonator tag has different types of planar spiral resonators such as rectangular spiral, archimedean/linear spiral, logarithmic spiral, triangular spiral, octagonal spiral and so on. Based on the number of data bit a series of resonators (one resonator for one bit) is used close to the transmission line to represent a unique ID of the tag. Each resonator acts as a tuned circuit inside the tag. In other words it can be said that the tag works as a multi frequency bandstop filter. In [9] a microstrip rectangular resonator based UWB tag is presented. Figure-1 shows a basic three bits chipless RFID tag structure with microstrip rectangular resonators and transmission line. Two different shapes of microstrip resonator (rectangular and circular shapes) are designed at same resonance frequency (2.5GHz) to observe the effect in space reduction.

2 Design of Circular Resonator

The spiral resonators give a parasitic effect while placed close with a transmission line. If different resonators are designed for different frequencies, the tag gives band stop responses in different frequencies. A planar circular spiral resonator and its equivalent circuit are shown in figure-2. The resonance frequency (f_{res}) can be expressed by the equation (1). Where, L_T and C_T are total distributed inductance and capacitance respectively of the spiral resonator.

The distributed capacitance can be calculated from the equation (2) described in [10] -[12]. Where, C_0 is the total distributed capacitance of one turn spiral, r_m is the mean radius and R_{in} and R_{out} is the inner and outer radius of the spiral. The distributed inductance (L_T) can be calculated from equation (3) which is well described in [13] and [14].

A.K.M.Z. Hossain (✉) • S.M.A. Motakabber • M.I. Ibrahimy
Department of Electrical and Computer Engineering, Faculty of Engineering International Islamic University Malaysia, 53100 Gombak, Kuala Lumpur, Malaysia

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_52,

Where, L_s is called as the self-inductance of the spiral and M is the mutual inductance between turns of the spiral. After calculating the distributed capacitance and inductance with the help of equation (1), the resonance frequency can be estimated for any spiral resonator.

$$f_{res} = \frac{1}{2\pi\sqrt{L_T C_T}} \tag{1}$$

$$C_T = \frac{C_0(R_{in} + R_{out})}{2r_m} \tag{2}$$

$$L_T = \sum_{(i=1)}^{n} L_s + 2[\sum M^{+-} \sum M^{-} + \sum_{(j=1)}^{j} Mi,j+1 \pm \sum_{k=1}^{k} M_{k,k+1}] \tag{3}$$

3 Results and discussion

The above stated theories give a good approximation for the lumped circuit of the spiral. But while designing the layout of the spiral circuit the coupled line estimations are needed which are well described in [15] and [16]. One planar rectangular spiral (PRS) resonator and one planar circular spiral (PCS) are designed for the resonance frequency at 2.5 GHz. The dimensions are listed in the table-1 (all the parametric values are in millimeters). Where, W_L is the width of the 50 Ω transmission line, S is the spacing between two spiral conductors; S_L is the separation of the spiral from the 50 Ω line, L_R is the length of the PRS, W_R is the width of the rectangular spiral, W_C is the width of the conductor of the spirals and R is the outer radius of the PCS.

Both PRS and CRS are simulated using CST MWS. Figure-3(a) and 3(b) show the simulation model of PRS and PCS respectively where the both spiral have 2.5 turns. The simulation bandwidth 1GHz to 4GHz is taken in the CST MWS. The line per wavelength of the simulator, 14 is chosen to create a dense mesh volume to give higher accuracy. The spirals are simulated on a TLX-0 substrate having a dielectric constant, $\varepsilon_r = 2.45$, height of the substrate is 0.787 mm. The spirals are modeled as PEC (perfect electric conductor) in the simulator and the thickness of the copper is taken 35 μm. The simulation results of the S-parameter are shown in the figure-4.

From figure-4, it can be observed that both spirals are resonating in the same frequency (2.5 GHz) and giving a band stop response (less than -10 dB transmission). Referring to the table-1, it is also observed that the occupied area of the PCS resonator is less compared with the PRS resonator.

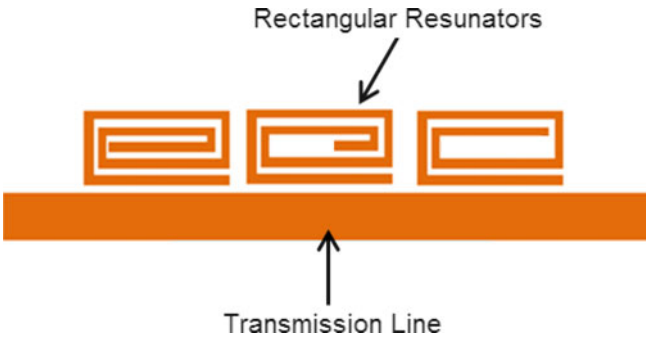


Fig. 1 Basic structure of a multiple-rectangular spiral resonator of a chipless tag

Table 1 Parameters of the spiral resonators

	W_L	S_L	S	L_R	W_R	W_C	R	Area
Rectangular	2.26	0.2	0.3	6.85	6.3	0.8	N/A	43.15
Circular	2.26	0.015	0.3	N/A	N/A	0.8	3.52	38.92

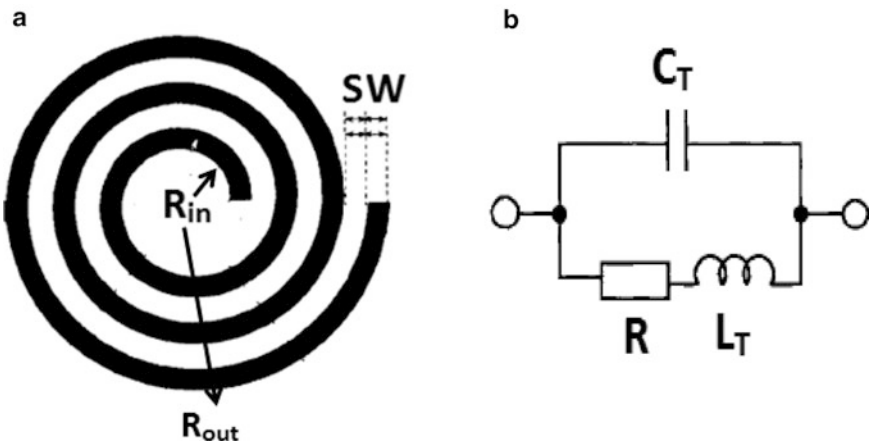


Fig. 2 (a) Geometry of a planar circular spiral and (b) Equivalent Circuit of the spiral

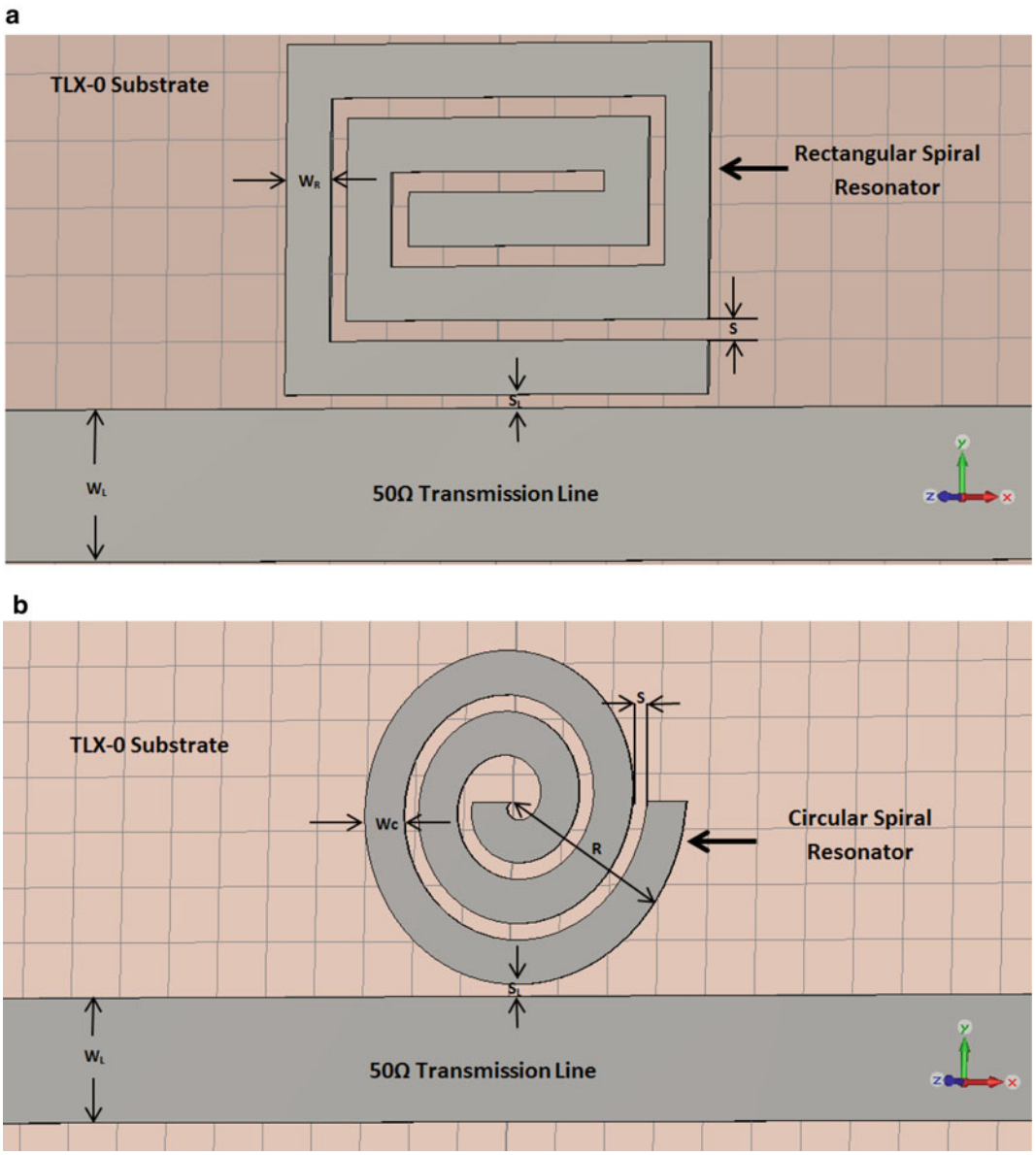


Fig. 3 Layouts of the (a) PRS and (b) CRS in CST MWS

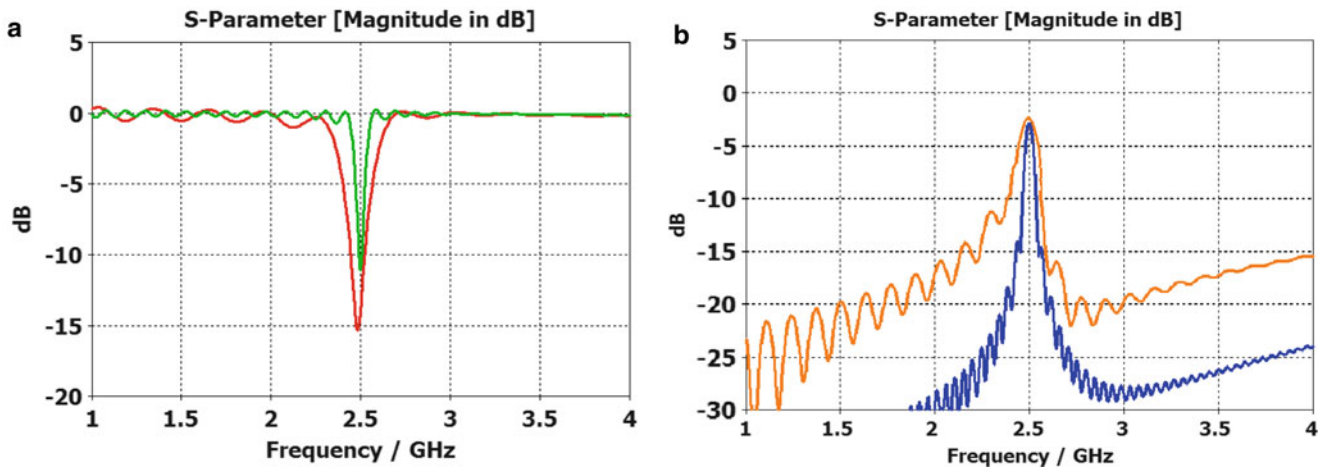


Fig. 4 S-parameter: (a) S21 and (b) S11 response for circular and rectangular spiral resonator

4 Conclusion

Two different spirals are simulated using EM simulator CST MWS for the same resonance frequency. The results of S-parameter are studied and good responses are found at 2.5GHz for both of the resonators. In terms of the area, it is clear that a reduction of 9.8 % is obtained while implementing the PCS instead of PRS esonator. It can be concluded that more resonator as well as number of data bits are accommodated in the chipless tag by using PCS instead of PRS resonator.

Acknowledgment This research has been supported by the Ministry of Higher Education of Malaysia through the Fundamental Research Grant Scheme FRGS13-027-0268.

References

1. Finkenzeller, K. (1999). RFID handbook: radio-frequency identification fundamentals and applications (pp. 151-158). New York: Wiley.
2. Want, R. (2006). An introduction to RFID technology. *Pervasive Computing, IEEE*, *5*(1), 25-33.
3. Uddin, M. J., Ibrahimy, M. I., Reaz, M. B. I., & Nordin, A. N. (2009). Design and application of radio frequency identification systems. *European Journal of Scientific Research*, *33*(3), 438-453.
4. Ngai, E. W. T., Moon, K. K., Riggins, F. J., & Yi, C. Y. (2008). RFID research: An academic literature review (1995–2005) and future research directions. *International Journal of Production Economics*, *112*(2), 510-520.
5. Nambiar, A. N. (2009, October). RFID technology: A review of its applications. In *Proceedings of the world congress on engineering and computer science* (Vol. 2, pp. 20-22).
6. Motakabber, S. M. A., Ibrahimy, M. I., & Alam, A. H. M. (2013, August). Development of a position detection technique for UWB chipless RFID tagged object. In *Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on* (pp. 735-738). IEEE.
7. Plessky, V. P., & Reindl, L. M. (2010). Review on SAW RFID tags. *Ultrasonics, Ferroelectrics and Frequency Control, IEEE Transactions on*, *57*(3), 654-668.
8. Lazaro, A., Ramos, A., Girbau, D., & Villarino, R. (2011). Chipless UWB RFID tag detection using continuous wavelet transform. *Antennas and Wireless Propagation Letters, IEEE*, *10*, 520-523.
9. Preradovic, S., & Karmakar, N. C. (2010). Chipless RFID: bar code of the future. *Microwave Magazine, IEEE*, *11*(7), 87-97.
10. Preradovic, S., Balbin, I., Karmakar, N. C., & Swiegers, G. (2008, April). A novel chipless RFID system based on planar multiresonators for barcode replacement. In *RFID, 2008 IEEE International Conference on* (pp. 289-296). IEEE.
11. Uddin, M. J., Nordin, A. N., Ibrahimy, M. I., Reaz, M. B. I., Zulkifli, T. Z. A., & Hasan, M. A. (2009, April). Design and simulation of RF-CMOS spiral inductors for ISM band RFID reader circuits. In *Microelectronics and Electron Devices, 2009. WMED 2009. IEEE Workshop on* (pp. 1-4). IEEE.
12. Jiang, Z., Excell, P. S., & Hejazi, Z. M. (1997). Calculation of distributed capacitances of spiral resonators. *Microwave Theory and Techniques, IEEE Transactions on*, *45*(1), 139-142.
13. Ellstein, D., Wang, B., & Teo, K. H. (2012, October). Accurate models for spiral resonators. In *Microwave Conference (EuMC), 2012 42nd European* (pp. 787-790). IEEE.
14. Hejazi, Z. M., Excell, P. S., & Jiang, Z. (1998). Accurate distributed inductance of spiral resonators. *Microwave and Guided Wave Letters, IEEE*, *8*(4), 164-166.
15. Schmückle, F. J. (1993). The method of lines for the analysis of rectangular spiral inductors. *IEEE Transactions on Microwave Theory and Tech*, *41*(6/7).
16. Preradovic, S., & Karmakar, N. C. (2012). Spiral Resonators. In *Multiresonator-Based Chipless RFID* (pp. 25-51). Springer US

An Evaluation of Intrusion Detection System on Jubatus

Tadashi Ogino

1 Introduction

As the internet is spreading in our daily life and in the business scene, the attacks using the internet are increasing. Although the technology of defense from such attacks is making progress rapidly, the attacks are becoming smarter.

There are mainly two different approaches for network intrusion detection technology [1]. One is signature-based technology, and another is anomaly detection technology.

With signature-based technology, the system has set of attack patterns and compares them with actual transferred data. When the data match the attack patterns, it means the data is an attack. This system can detect all the data in the attack set, but cannot detect new attacks which are not included in the attack data set.

The anomaly detection system has a normal behavior pattern profile about the defense system. When coming data is different from the normal pattern, it is assumed as an attack. This system can detect new attacks. The problem is that this system has a possibility of alarming for normal date as attacks.

Majority of intrusion detection studies had been about signature-based technology. The anomaly detect technology has been getting focus recently. Studies in the area of machine learning, big-data analysis and so on, have been applied to anomaly detection studies. As a result, anomaly detect systems with feasible performance are being developed.

The objective of our research is to build an intrusion detection system combining current up-to-date technologies. As a preliminary study, we evaluate the performance of anomaly detection algorithm "LOF" [2] on the online machine learning framework "Jubatus" [3].

T. Ogino (✉)
Department of Information and Communication Systems Engineering, Okinawa National College of Technology, Okinawa, Japan
e-mail: ogino@okinawa-ct.ac.jp

2 System Overview

The basic processing of our system is shown in Figure 1. The system collects the system logs, i.e. traffic log, manipulation log and so on. All the collected logs are statically analyzed and learned as normal data. In this period, we estimate the system is not attacked. The system can detect new attacks after this learning period.

After a suitable period, the system starts to analyze the collected data. When the system detects the outlier, it analyzes the data in more detail. The outlier is not necessarily the intrusion. When new tools or new services are introduced in the user system, it might produce new traffic or usage patterns. So it is necessary to check if the outliers are results of normal usage or attacks. We suppose these tasks are executed by the system administrators. When the outliers are intrusions from outside, they are blocked. On the other hand, when the outliers are results from proper but new procedures, they are permitted and learned as normal data. After some proper amounts of those data are learned, they are not classified as outlier data.

This feedback step of registering the proper but new procedures as normal data is necessary, since a lot of new services appear in a short time recently.

The specific objectives of this system are as follows.

- The system can detect new un-known intrusions.
- The system can detect intrusions in a real-time.
- Not only the communication from/to outside, but also irregular communications/operations in the local area network can be detected.
- The detected incidents can be investigated by system administrators and decided if they are actual intrusions or normal usage.

The total system configurations are shown in Figure 2. Each component has the following functions.

– **External Communication Point**: The communication points between external networks i.e. internet and internal systems (LAN). For the purpose of BCP enhancement,

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_53,

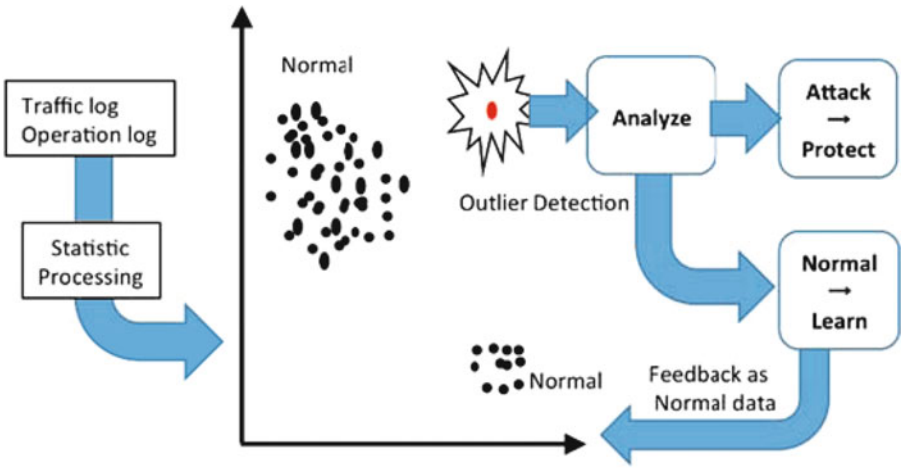


Fig. 1. Overview of System Processing

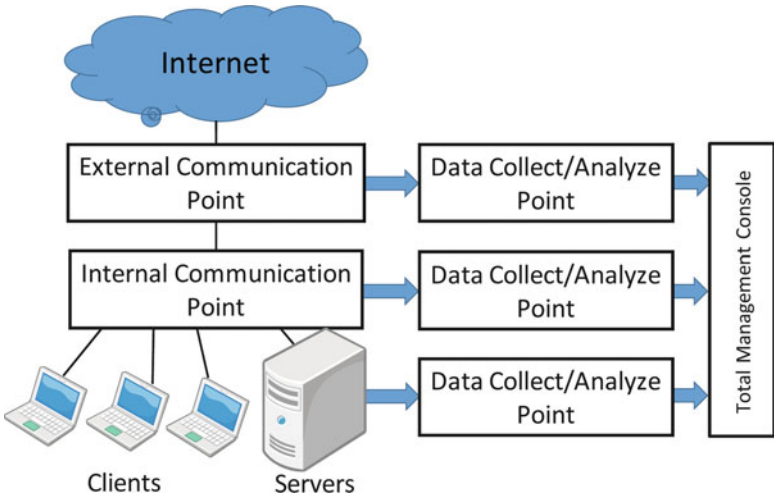


Fig. 2. System Configuration.

more than two external communication points might be equipped. The traffic logs from/to internet are collected from these points.

- **Internal Communication Point**: The communication points between local clients and servers. Usually these points are built by routers/switches. The traffic logs in the local system are collected from these points.
- **Servers**: File servers, print servers, application servers, etc. Server operation logs are collected.
- **Clients**: Each client in the system. Client operation logs are collected.
- **Data Collect/Analyze Point**: The Data Collect/Analyze Points collect all the traffic logs and operation logs, analyze all the data and find the abnormal data. In the case of servers/clients, data analysis function can be executed by themselves.
- **Total Management Console**:All the found suspicious incidents are gathered. The system administrators can invest and manage all the incidents.

This paper describes the evaluation of collecting traffic data at External Communication Point and detecting anomaly data.

3 Detection Method

This chapter explains how our system analyzes the traffic data and finds anomaly data in real time.

3.1 Anomaly Detection

A couple of studies have been conducted on anomaly detection from huge data sets. Most of them come from extensive studies for clustering. The main objectives of such studies are to find anomalies to remove them from the data set, since they are “noises” for clustering purpose. There are less studies about anomalies in order to process anomalies as

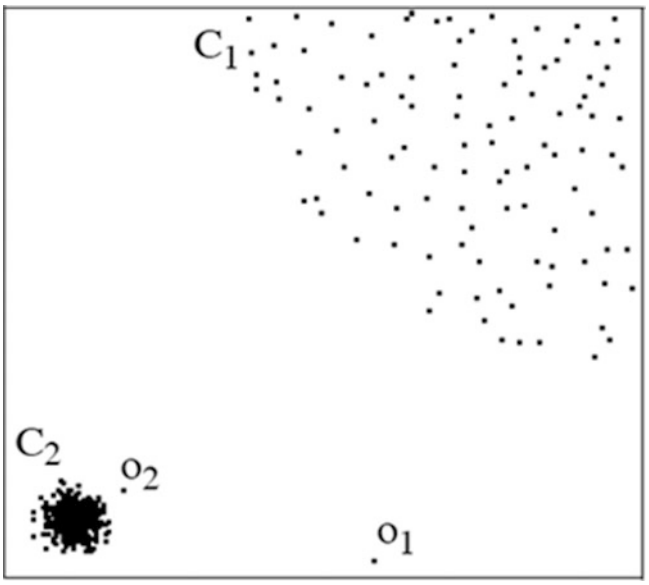


Fig. 3. 2-d Dataset.

main subject. In this paper, we use LOF (Local Outliner Factor) algorithm [2] from such 'outlier-oriented' studies. It is reported that LOF is superior to other outlier detect algorithm for detecting network intrusion [4]. Another reason to use LOF is that machine learning framework Jubatus has LOF algorithm for its standard repertoire. Jubatus will be explained in 3.3.

3.2 LOF

There are a couple of definitions for outliers. Hawkins gave the definition of outliers as "an observation that derives so much from other observations as to arouse suspicion that it was generated by a different mechanism." [5]

Most of the cases, this definition is enough for detecting outliers. But, in the case of Figure 3 [2], though o_1 can be detected as outliner, o_2 can not be detected. LOF is designed to find such cases.

The details of LOF algorithm are explained in the original paper.

Some examples of detecting cyber attacks using outlier detect algorithm are as follows:

- Example 1: When the time period of packet transmission is too short, there are possibilities of DoS attacks.
- Example 2: When the length of the packet is too long, there are possibilities of BufferOverFlow attacks.

Even in the case of new unknown attacks, since there may exist some different parameters for normal traffic, the possibility to be detected as outliers might be high.

3.3 Jubatus

The amount of network traffic is increasing dramatically. In order to analyze those data in real time, it would be better to apply 'big data' analysis technology to manipulate huge data.

One of the most famous big data analysis tools is Hadoop [6]. Hadoop is an open source software tool. It is used in the wide areas such as recommendation, web search, text mining etc. Hadoop is usually used under batch processing. It is a little bit difficult to use it for real time analysis.

We use Jubatus [3] for our real time analysis platform. Jubatus is a distributed machine learning platform developed by NTT and PFI. It is developed for the purpose of real-time, deep analysis in a distributed environment.

4 Experiment

For the purpose of preliminary study, we evaluate the performance of LOF algorithm running in Jubatus framework with KDD Cup 99 data [7]. Since Jubatus includes LOF algorithm as standard function, KDD Cup 99 data can be analyzed with a simple task. But as far as we know, the performance of the system has not been reported yet. This paper is to report it.

4.1 Evaluation System

We use Ubuntu server for evaluation system. As the future plan includes the distributed system evaluation, the evaluation system is built on a virtual machine using KVM on Linux. Out PC server has 2 E5-2650 (2GHz) CPUs and 128GB Memory.

4.2 Evaluation Data

We use KDD Cup '99 data as the experiment traffic data. KDD Cup is a Data Mining and Knowledge Discovery competition organized by ACM. In 1999, the main topic was Network Intrusion Detector. The competition data of KDD Cup 99 can be used as the evaluation data for network intrusion detection system. The data simulates the typical U. S. Air Force LAN. The raw data was 4GB of compressed TCPdump format from seven weeks of network traffic. This data was processed to 5 million connection records. Each connection record is labeled as either a normal, or an attack. An attack data has its attack type such as buffer_overflow, guess_passwd etc. Each record consists of 41 columns and

Table 1. Execution Time Result

Seconds

number of data		Type: all	Type: skip method ratio=10	Type: skip method ratio=100	Type: skip method ratio=1000
100%	4,898,431	>96 hours	—	—	3836
10%	494,021	>96 hours	2588	434	411
1%	49,402	460	40	37	39
0.1%	4,940	4.5	5.2	5.3	5.2

— not tried

the record size is around 100 bytes. They also supplies 10 % data, which has the same data distribution and 500 thousand records.

Jubatus has LOF processing function. We use this function with no modification. KDD Cup 99 data is also used with no modification. It is treated as 41 dimensional data.

5 Results

5.1 System Processing

The experimental system processing is as follows.

1. First 1000 data are treated as training data.
2. When the number of data exceeds 1000, the majority of the data is not treated as learning data and just calculates the LOF score
3. Just a fraction of the data is learned as training data

The calculation time of LOF score depends on the cumulative learned data so far. When all the data are learned, the execution time increases and cannot get the result in a reasonable time. On Jubatus, it is possible just to calculate the LOF score of one record and doesn't learn the data. This means the number of learned data doesn't increase and the time to calculate LOF score doesn't increase as the processing proceeds. There is no meaning for the magic number 1000. This number will be tuned in the future study.

The reason to add some small part of the following data as training data is that the network traffic is changing as time elapsed. The system has to continue to learn the traffic change.

We call this method as 'skip method'. The ratio of training data vs. calculated data is called skip ratio.

Using this skip method, the KDD Cup 99 analysis finished in a reasonable time with skip ratio 10, 100, 1000.

5.2 Evaluation of Skip Method

Other than KDD Cup 99 full/10 % data, we use the following data for evaluation.

- 1 % data: first 10 % data of KDD Cup 10 % data.
- 0.1 % data: first 1 % data of KDD Cup 10 % data.

The distributions of these data are different from that of the original KDD Cup data.Table 1 shows the execution time of skip method.

When the learned data exceeds tens of thousands, the total execution time increases dramatically. Under this condition, we need at least 10 for skip ratio to get results in a reasonable experimental time.

5.3 Detection Rate

Table 2 shows the detection rate for each attack type. This result is from 10 % data and skip rate is 100. The detection rate for normal label is calculated with records classified as normal, i.e. LOF equals 1 or infinity. The detection rate for attack labels is calculated with records classified as anomaly.

The ratio of correctly classified attack records ranges from 0 % to 80 %. The total ratio of correctly classified attack records ranges from 8.6 % to 25 %. The previous study [8] shows the ratio of correctly classified is around 50 ~ 60 %, although the definition and method of detection is different. This means our system needs more tunings.

The ratio of correctly classified attack records ranges from 0 % to 80 %. The total ratio of correctly classified attack records ranges from 8.6 % to 25 %. The previous study [8] shows the ratio of correctly classified is around 50 ~ 60 %, although the definition and method of detection is different. This means our system needs more tunings.

Table 2. Detection rate for each attacks

label	number of records	skip method					
		ratio = 10		ratio = 100		ratio = 1000	
		N	%	N	%	N	%
normal	97,278	85,578	88.0%	89,914	92.4%	96,480	99.2%
back	2,203	4	0.2%	0	0.0%	0	0.0%
buffer_overflow	30	0	0.0%	0	0.0%	0	0.0%
ftp_write	8	0	0.0%	0	0.0%	0	0.0%
guess_passwd	53	0	0.0%	0	0.0%	0	0.0%
imap	117	0	0.0%	0	0.0%	0	0.0%
ipsweep	1,247	495	39.7%	20	1.6%	0	0.0%
land	21	2	9.5%	1	4.8%	0	0.0%
loadmodule	9	0	0.0%	0	0.0%	0	0.0%
multihop	7	0	0.0%	0	0.0%	0	0.0%
neptune	107,201	96,115	89.7%	81,967	76.5%	21,895	20.4%
nmap	231	8	3.5%	0	0.0%	0	0.0%
perl	3	0	0.0%	0	0.0%	0	0.0%
phf	4	0	0.0%	0	0.0%	0	0.0%
pod	264	69	26.1%	0	0.0%	0	0.0%
portsweep	1,040	364	35.0%	460	44.2%	11	1.1%
rootkit	10	2	20.0%	0	0.0%	0	0.0%
satan	1,589	1,135	71.4%	706	44.4%	410	25.8%
smurf	280,790	3,703	1.3%	15,761	5.6%	11,705	4.2%
spy	2	0	0.0%	0	0.0%	0	0.0%
teardrop	979	96	9.8%	0	0.0%	0	0.0%
warezclient	1,020	334	32.7%	0	0.0%	0	0.0%
warezmaster	20	0	0.0%	0	0.0%	0	0.0%
all attacks	396,848	102,327	25.8%	98,915	24.9%	34,021	8.6%
total	494,021	187,905	38.0%	188,829	38.2%	130,501	26.4%

6 Discussions

6.1 Execution Time

Our target is to detect cyber attacks in a real time. KDD Cup '99 data consists of 7weeks traffic data. Considering that our system needs administrator operations, the evaluation result that analyzing time for 10 % data is around a couple of minutes shows our target can be achieved with suitable system designs.

In order to improve system performance, we have some ideas now under evaluating.

- **Parameter Tuning**: Jubatus and LOF algorithm have some tuning parameters. As we haven't tuned such parameters this time, tuning might improve system performance.
- **Distributed Configuration** : As Jubatus is designed to run in a distributed environment with multiple servers, this can improve system performance with ease.
- **Algorithm Tuning** : In this experiment, we learned that the number of learned data affects the system performance dramatically. We haven't made any intelligent effort to decrease the number of learned data except just skipping the records. Now we have a couple of ideas to decrease the learned data. We are evaluating those ideas.

6.2 Detection Rate

As shown in the previous chapter, the detection rate of this experimental system is very low. We have some candidates to increase the detection rate.

At first, we examined the LOF score carefully and found "inf" value was included over 40 % (Table 3). After we checked the data whose LOF value is "inf", we found that even if there was no same vector in the learned data, Jubatus returns "inf". This is because Jubatus implementation uses 'euclid_lsh' for finding the nearest-neighbor data. When

Table 3: LOF score actually returned.

	inf	1.0	others	Total
normal	94,887	595	1,796	97,278
attack	116,625	247,363	32,755	396,743
Total	211,512	247,958	34,551	494,021

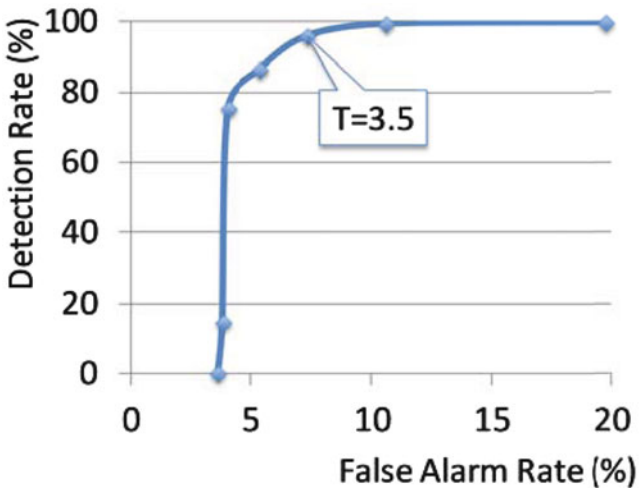


Fig. 4. Detection Rate vs False Alarm Rate.

there are no vectors in the same hash bucket, Jubatus quits finding near vectors and returned 'inf' value for the data. This happens for both normal data and attack data and we can not decide if 'inf' data is normal data or anomaly data. In order to decrease this effect, we should increase Jubatus LOF parameter 'bin_width', which means the size of hash buckets.

7 Improved System

From the previous discussion, we changed the experiment system as follows:

- **Change parameters** : We changed 'bin_width' parameters from original 10 to 10000.
- **Classification** : We set the threshold value T more than 1. If LOF value is less than T, the data is assumed as normal data and if LOF value is over T, the data is assumed as an attack data. The detection rate changes according to T value.

Figure 4 shows the detection rate and the false alarm rate. If we choose suitable T value, we got better results than the original skip method. For example, when T = 3.5, the (total) detection rate is 95.9 % and the false alarm rate is 7.33 %. We are investing and improving this system in more detail.

8 Future Works

We need to invest the following issues.

- **Automatic Parameter Tuning**
  The performance of the improved system might be increased by more parameter tuning.We need to study how to find the best threshold value T automatically.
- **Better Learning Algorithm**
  The number of learned data affects the total system performance. We need to find a better learning algorithm to continue to learn and not to decrease performance.
- **Administration Issue**
  We need to develop the administrator interface.

9 Conclusion

In this paper, we propose the cyber attack detection system using LOF algorithm running on Jubatus platform. Our evaluation shows the performance of the system is good enough for building real time detection system. The detection rate can be improved with appropriate parameters. According to the investigation of the result, we need more studies to achieve our goal. We intend to continue this research in the future.

References

1. Chandola, V., Banerjee, A. & Kumar, V.: Anomaly Detection: A Survey. In:ACM Computing Surveys, July.41(3). (2009)
2. Breunig, M. M., Kriegel, H.-P., Ng, R. T. & Sander, J.,: LOF: Identifying Density-Based Local Outliers. In: Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data (2000)
3. Jubatus, http://jubat.us/
4. Lazarevic, A. et al., : A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection. In: Proceedings of SIAM Conference on Data Mining (2003)
5. Hawkins, D. M.: Identification of Outliers. London: Chapman and Hall. (1980)
6. hadoop, http://hadoop.apache.org/
7. KDDCup1999, https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
8. Aleksandar, L. et al.: A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection. In: Proceedings of the Third SIAM International Conference on Data Mining. (2003)

System of Conceptual Design Based on Energy-Informational Model

Viktorya Zaripova and Irina Petrova

1 Introduction

As part of the science of engineering design a large number of design methods have been developed. Many methodologies have similar objectives, structure and inherit each other's design theory principles. Last research review is presented in [1]. Important conclusions that have been made in this work identified the following trends in engineering design:

1) the development of decision making intelligent systems of engineering design, based on a vast and complex knowledge bases
2) the development of ontology and semantic interoperability as the tool for engineers co-working.

In [2] there is a graph showing the value of design decisions at various stages of the product realization cycle and the availability of tools of human-computer interaction for designers (Figure 1).

As follows from this figure, only few tools are available to help designers make best decisions early in the product realization cycle, where they provide the greatest benefit.

In [1] 324 sources have been analyzed, and in [2] 80 journals and conference proceedings as well as about 20 R&D projects have been examined. This allows arguments to the following conclusions:

1) the creation of intelligent tools to support human-designer in the early stages of design is still relevant and useful task;
2) the creation of such tools need in creation and using of vast knowledge bases and ontological methods for fast and relevant search. It will provide engineers to work together in real or virtual space.

V. Zaripova (✉) • I. Petrova
State Autonomous Educational Institution of Astrakhan Region of Higher Professional Education, «Astrakhan Civil Engineering Institute», Astrakhan, Russia
e-mail: vtempus2@gmail.com; irapet1949@gmail.com

In [3] the classification of AI-based models of innovative design is discussed. It is shown that the methods using a systematic approach and knowledge base of the physical effects of different nature are the most effective ones. Additionally, this article lists several scientific schools involved in research related to the systematization and modeling of knowledge in natural sciences in order to develop intelligent systems for engineering design. The scientific school of Prof. Zaripov (Russia) is also mentioned there [4, 5, 6].

Therefore, in this paper we present A System for Conceptual Design Based on Energy-Informational Model (elaborated on base of Prof. Zaripov scientific school), consider the design of a knowledge base on the physical effects, which was elaborated on base of the domain ontology.

2 Theoretical Bases of Energy-Information Method of Analysis and Synthesis of Technical Solutions

2.1 Provisions of Non-Equilibrium Thermodynamics as the Theoretical Basis of Energy-Information Model of Chains of Different Physical Nature

Analysis of different systematic approaches to developing of knowledge bases for conceptual design [3], [12] showed that to systematize the knowledge we need in a method that combines:

- mathematical modeling of processes in the technical device (invariant to the physical nature of these processes)
- the possibility of operating with the physical effects and phenomena that are beyond the strict framework of the model,
- the possibility of structural description of the physical principle of the device operation.

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_54,

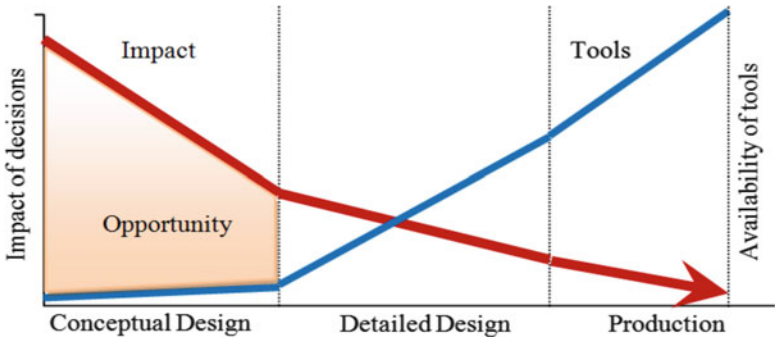


Fig. 1 Few tools are available to help designers make best decisions early in the product realization cycle, where they provide the greatest benefit.

Therefore, the basis of non-equilibrium thermodynamics was chosen for developing a conceptual model for systematization of knowledge on physical phenomena and effects for the synthesis of new technical devices (L. Onsager, S.R. de Groot and P. Mazur, I. Prigogine) [7, 8], as they allow you to obtain a complete system of transport equations and other laws, without opening their molecular mechanism. Additionally, you must expand the system of physical quantities and parameters used in the non-equilibrium thermodynamics like to the theory of electrical circuits, in order to use its methods of analysis and synthesis.

There are a number of phenomenological laws describing the irreversible processes in the form of direct proportional relationships. For example, Fourier's law of proportionality of the heat flux to the temperature gradient, or Fick's law which relates the diffusive flux to the concentration, Ohm's law of proportionality of electric current to the potential gradient, Newton's law of proportionality of the force of internal friction to the velocity gradient, the law of proportionality of the chemical reaction rate of the chemical potential gradient.

When two or more of these phenomena occur simultaneously, they are superposed on each other and cause the appearance of a new effect. For example, thermoelectricity occurs after imposing of thermal and electrical conductivity, thermal diffusion (Soret effect) - occurs after imposing of diffusion and heat conduction.

Thus, irreversible phenomena can be attributed to the following factors: temperature gradient, concentration gradient, the gradient of the electric potential, chemical potential, etc. In thermodynamics of non-equilibrium processes all of these factors are called thermodynamic forces and denoted by $(i = 1, 2, 3 \ldots n)$. These forces cause the known irreversible phenomena: heat flux, diffusion current, chemical reactions, etc. All these are called fluxes and denoted by Ji $(i = 1, 2, 3 \ldots n)$.

In the most general case any force can cause any flux. For example, the diffusive flux can be caused by the presence of a concentration gradient or temperature gradient or electrical potential gradient. Therefore, any irreversible phenomenon can be written as a phenomenological relation:

$$J_i = \sum_{k=1}^{n} L_{ik} \cdot \chi_k \ (i = 1, 2, 3 \ldots, n), \tag{1}$$

Odds L_{ik} $(i = 1, 2, 3, \ldots n)$ are called phenomenological coefficients or transfer coefficients.

According to the expression (1) each flux is a linear function of all system thermodynamic forces. L_{ii} describe simple processes (electrical conductivity, thermal conductivity, diffusion, etc.). L_{ik} coefficients associated with superimposed phenomena (thermal diffusion, electrodiffusion, etc.) when $i \neq k$.

Fundamental Onsager reciprocal relation was derived for the phenomenological coefficients in the thermodynamics of nonequilibrium processes. It argues that the L_{ik} coefficient matrix is symmetric, i.e. cross ratios are equal:

$$L_{ik} = L_{ki}, \tag{2}$$

if the generalized fluxes J_i and forces χ_i appropriately selected.

Choosing of fluxes and forces is made by using the entropy balance equation:

$$dS = d_eS + d_iS, \tag{3}$$

where d_eS - the entropy change due to its entering to the system from the environment $\left(d_eS = \frac{dQ}{dT}\right)$, dQ – is a heat, send to the system from the environment);

d_iS - entropy change that occurs in the system (for reversible processes $d_iS = 0$ for irreversible - $d_iS > 0$).

Common forces, flows and coefficients for the physical effects of different physical nature can be defined via equations (1-3).

2.2 Energy-information model of circuits (EIMC) of various physical nature

Within EIMC the analysis and synthesis of the technical device expose therein the physical phenomenon of the particular nature (mechanical, thermal, electric, etc.) and the

corresponding constructive elements of these phenomena. To describe these phenomena EIMC introduces the following concepts:

Axiom 1 *Circuit* of certain physical nature is the idealized material medium having certain geometrical dimensions and characterized by its physical constants inherent only to phenomena of given physical nature.

Axiom 2 *Values of circuit* of same physical nature vary in a wide range and is characterized an external influence on a circuit of a given physical nature and its corresponding reaction

Axiom 3 *Circuit parameters* characterize the relative unchangeability of a material medium in which physical processes occurs. Parameters are defined by their geometrical dimensions, physical and chemical properties of materials.

The most simple energy-information model uses the following values: P - action momentum; Q - reaction charge; U - action force; I - reaction rate; as well as parameters: R - resistance; $G = 1/R$ - conductance; C - capacitance; $W = 1/C$ - rigidity; L - inductance; $D = 1/L$ - deductance.

Axiom 4 *EIMC criteria* - a system of equations that reflect the links between values and parameters, and used to identify specific values and parameters in the circuits of different physical nature.

The quantities and parameters of EIMC are interrelated by six criteria (the most simple case):

- first criterion(energy) : $U \cdot I = N\ (\mathrm{N} - \text{power})$, (4)
- second criterion (invariable) : $I \cdot L = P$ or the derivative criterion : $P \cdot D = I$; (5)
- third criterion (invariable) : $U \cdot C = Q$ or the derivative criterion : $Q \cdot W = U$; (6)
- forth criterion (invariable) : $I \cdot R = U$ or the derivative criterion : $U \cdot G = I$; (7)
- fifth criterion (variable) : $U = dP/dt$ or the derivative criterion : $\int Udt = P$; (8)
- sixth criterion (variable) : $I = dQ/dt$ or the derivative criterion : $\int Idt = Q$; (9)

The authors identified the system of values (analogs of physical ones) and parameters (analogs of physical ones) to describe processes in circuits of different physical nature.

Axiom 5 Communication between circuits of different physical nature is going by means of physical and technical effects. *Physical&technical effect (PTE)* is the objectively existing causal link, reflecting the dependence between physical quantities, which could not be described through the only EIMC criteria.

Analytical expressions for the physical and technical effects (PTE) coefficients and the numerical values of these coefficients, as well as performance of technical constructions on base of these PTE determined from the results of theoretical and experimental researches in the field of physics and technology and available in various sources of scientific - technical information.

One should note the following features. One physical phenomenon could be provided as several PTE depending on what and in what proportion quantities and parameters of different physical nature are involved in the description of a physical phenomenon. For each effect input and output values should be clearly indicated.

3 Ontology for Modeling Design Knowledge on base of EIMC

All the variety of interactions between quantities and parameters can be described as a complex graph via energy-information models for describing of chains of different physical nature and parametric structural diagrams (Fig. 2). The figure shows the graph of physical and technical effects and intracircuit dependencies on n chains: mechanical, magnetic, electrical circuit and the i-th physical nature.

Ontology is often represented in the form of a semantic network graph with nodes reflecting concepts or individual objects, and the arcs reflecting the relationships or associations of these concepts [9].

From the viewpoint of ontological approach any physical effect (PTE), connecting two chains of i-th and j-th physical nature or parameter of the chain of i-th physical nature may be represented by tuples of type:

$$PTE = \left\{H_{PTE}, B_{iin}, B_{j_{out}}, K, K_0, KM_{PTE}, D_{i_{in}}, D_{i_{out}}, EX^N_{n|1}\right\} \quad (10)$$

$$\Pi_i = \left\{H_{\Pi}, B_{iin}, B_{j_{out}}, \Pi, \Pi_0, \Pi M_{\Pi}, D_{i_{in}}, D_{i_{out}}, EX^N_{n|1}\right\} \quad (11)$$

Tuples can be divided into 2 groups, where the first group is the description of the physical and technical effect:

H_{PTE}, H_{Π} - PTE or parameter name, text value,

$B_{i_{in}}$ - the i-th physical quantity,

$B_{i_{out}}$ - the j-th physical quantity,

K - PTE coefficient, reflects the dependence of the input and output value (the simplest case - a linear one $B_{j_{out}} = K_{ij} \cdot B_{i_{in}}$),

Π – parameter of the chain of i-th physical nature ($B_{i_{out}} = \Pi_i \cdot B_{i_{in}}$),

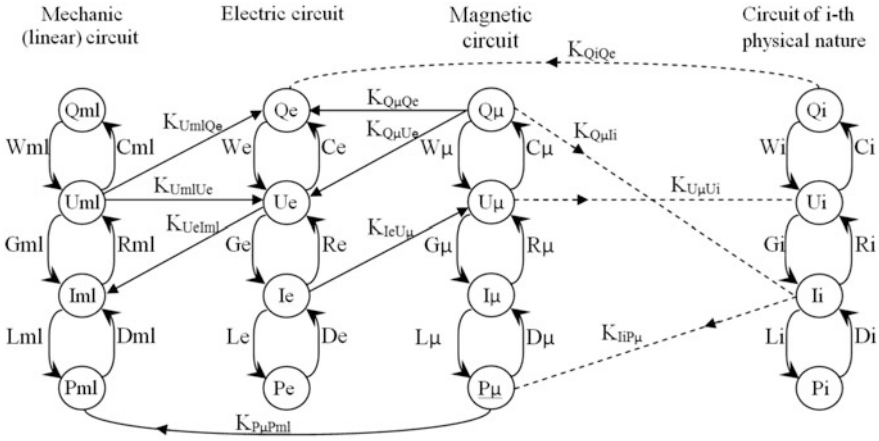


Fig. 2 Complex graph for n circuits of different physical nature.

K_0 or Π_0 - text variables, description of the K_{ij} coefficient or parameter and its physical formula through known physical constants, material parameters and its geometric dimension,

KM_{PTE} - mathematical model of PTE, which specifies the factors that influence the functional link of the physical input and output quantities, such as the influence of the fields (the value 1 or 0),

ΠM_{Π} - mathematical model of parameter (the value 1 or 0),

$D_{i_{in}}, D_{j_{out}}$ - variation range of input and output values. In order to ensure efficiency of the circuit it is necessary to observe the rules of crossing the ranges of output values for each of the previous effect and the input ones of each subsequent effect in the circuit: $D^n_{i_{in}} \cap D^{n-1}_{i_{out}}$ и $D^n_{j_{out}} \cap D^{n+1}_{j_{in}}$.

The second group - is a set of performance characteristics (with values from 0 to 10). The set is determined and filled by the group of subject area experts:

$EX^N_{n|1}$ - variables to calculate the performance of new synthesized physical operation principle. If we know values of operational characteristics of at least one type for all known PTEs in the synthesized circuit it is possible to calculate this operational characteristic for the entire device (entire circuit).

A complete correlation of output and input quantities of each in-circuit pair of PTEs is necessary and sufficient condition for the synthesis of the operation principle of technical device:

$$TD = \left(PTE_{i_1 j_1}, PTE_{i_2 j_2}, \ldots, PTE_{i_\Pi j_\Pi} \middle| \; PTE_{i_K j_K} \in DB \wedge j_k = i_{k+1} \wedge Q_{j_k\ out} = Q_{j_k\ in} \middle|_{k=1}^{\Pi}\right) \quad (12)$$

Thus technical device will be workable only if the corresponding quantities ranges overlap each other $D^n_{i\ in} \cap D^{n-1}_{i\ out}$

Operational characteristics of a technical device are computable only if each PTE in the circuit has calculated performance for the characteristic:

$$H_{nTD} = f(H_{nTD}, (\forall PTE \in TD)(\exists H_{nTD})) \quad (13)$$

The minimum set of values necessary for the successful synthesis procedure could be determined based on the PTE logical model and the above expressions

$$PTE = \{H_{PTE}, \mathrm{B_i}\, in, \mathrm{B_j}\, out, 1, 0, 0, (-\infty, +\infty), (-\infty, +\infty), \{0\}, 0, 0, 0, 0, 0\}$$

The sample PTE passport is shown in the Table 1.

4 Morphological synthesis of technical devices

Once the variety of operation principles is generated, an assessment of their performance is made and the best solutions are selected, engineer could assess the previous experience in designing of such devices and make some improvements. In order to do this he uses morphological approach.

Morphological investigation of the structure includes stage of analysis of array of known technical realizations of this structure as well as creation of structure's morphological matrix; and then goes the stage of new solutions

Table 1 The PTE sample passport

The reverse (converse) piezoelectric effect

(Name of the PTE)

(Parametric block diagram of the PTE)

$$Q_{ml} = K_{U_eQ_{ml}} \cdot U_e$$

(The PTE's formula, according to EIMC)

$$\frac{\Delta l}{L} = \varepsilon = d \cdot E_e \qquad \varepsilon = \frac{Q_{ml}}{L}; \qquad E_e = \frac{U_e}{L}; \qquad Q_{ml} = d \cdot U_e$$

(Phisical formulas of PTE description)

$$K_{UeQ_{ml}} = d \ [\text{m/V}]$$

(PTE coefficient, reflects the dependence of the input and output value)

E_e - electric field strength [V/m]

$\varepsilon = \Delta l / L$ - strain (relative length change),

L - the thickness of the quartz plate [m]

U_e – the electric potential difference [V]

$Q_{ml} = \Delta l$ – the displacement of the faces quartz plate [m]

d - the piezoelectric coefficient [C/N]=[m/V]

(Description of simbol in the formulas)

$K_{U_eQ_{ml}} = d_{11} = 2{,}31 \cdot 10^{-12}$ [m/V] (quartz)

$d_{33} = 16 \cdot 10^{-12}$ [m/V] $(LiNbO_3)$

$d_{33} = 100 \cdot 10^{-12}$ [m/V] $(BaTiO_3)$

(Numerical values of physical constants and properties of materials)

Sensitivity: $4 \cdot 10^{-10}$ [м/В]
Price: 5
Reliability: $1 \cdot 10^{-5}$ [1/час]
Error: 5 %
Nonlineriaty: 3 %

Range: 0 ÷100 [В]
Speed: $1 \cdot 10^{-6}$ [с]
Losses: 0
Ecology: $1 \cdot 10^{-6}$ [Кг/час]
Weight: 0,01 [Кг]

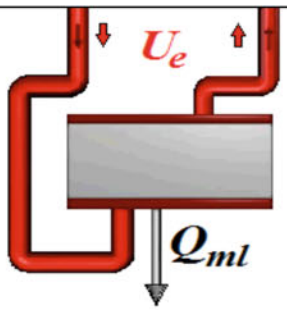
(PTE technical characteristics)

S. Garcia, E. Kunitz, K. Sampson Piezoelectric effect and its applications // Retrieved 2013-01-24 from http://www.utwente.nl/ewi/tst/education/elbach/mandt/extra/background/piezo01.pdf

Jan Tichý et al. Fundamentals of Piezoelectric Sensorics // DOI 10.1007/978-3-540-68427-5 // Springer-Verlag Berlin Heidelberg 2010

Gautschi, G. (2002). Piezoelectric sensorics. Springer Berlin, Heidelberg, New York.

(Bibliografy of PTE)



(Figure of PTE technical realization)

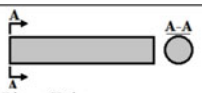
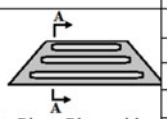
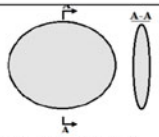
The reverse piezoelectric effect: the internal generation of a mechanical strain resulting from an applied electrical field

The amount of deformation of the crystal varies linearly with changes in the field value.

Inverse piezoelectric effect is used to measure large or rapidly changing pressures, particularly for the study of ultrasonic vibrations; voltage that appears on the faces of the deformed crystal is amplified and then supplied to the measuring instruments.

(Description of PTE features)

Table 2 The fragment of the morphological matrix of a physico-technical effect

Feature (sign)	Values of features								
Shapes of the Piezo compo-nents	 The Piezo Rectangle Plate	2 3 2 2 2 2	 The Piezo Bimorph Rectangle Plate	5 1 5 3 3 3	 Piezo Tube	1 4 3 5 5 5			
	 The Piezo Plate with variable section on height	4 2 4 4 4 4	 The disk-shaped piezo component	3 5 1 1 1 1	Sensitivity Price Liability Accuracy Non-linearity Range				
Conduc-ting electrode s	 Thick-film elec-trodes are applied to the piezo ceramic by screen printing technology.	2 1 1 1 1 2	 Thin-film electrodes are applied to the ceramic using modern PVD processes (sput-tering).	1 2 2 2 2 1	Sensitivity Price Liability Accuracy Non-linearity Range				
Piezo-electric materials	Natural monocrys-talline materials: Quartz, Tourmaline and Rochelle salt	1 3 3 1 1 1	Monolithic piezoelec-tric ceramics: $BaTiO_3$, Zirconate titanate (PZT),	3 1 1 3 3 3	Sensitivity Price Liability Accuracy Non-linearity Range				

Choice of one or several optimal technical solutions according to totality of their performance data.

synthesis according to morphological matrix and choice of optimal solutions according to field-performance data [10]. The authors have suggested to elaborate morphological matrixes not for a device in whole but for specific parameters or physico-technical effects, which make up the principle of operation.

The totality of existing and imaginary constructive realizations of each parameter or PTE make a morphological set. Finiteness of the set is stipulated by finiteness of human knowledge and by restrictions made by input circuit. Morphological set of each circuit parameter or PTE is being in evolution state. Any new constructive realization, new technology or new materials replenish this set. Advantage of using such matrix is that it enables to formalize the process of searching the best constructive realization of PTE on morphological set. An example of morphological matrix of a physico-technical effect is given in Table 2.

Stage of selecting of constructive realizations from morphological matrix is called a morphological synthesis, it includes the following actions:

Evaluation of all variants available in morphological matrix according to totality of performance data.

Choice of one or several optimal technical solutions according to totality of their performance data.

5 Comparing TIPS, SAPB and EIMC

In this article, we have discussed only the basic principles of the conceptual design based on energy-information model. Architecture of automated system for synthesis of new technical solutions is given in [5, 6, 11].

This section presents results of the systematic comparative analysis of the proposed conceptual design methodology with the one discussed in [12], presenting comparative analysis of the theory of Inventive Problem Solving and the systematic approach of Pahl and Beitz. Results are shown in Table 3. The aspects for the comparison have been selected to cover the task clarification and conceptual design stages of the design process.

Table 3 TIPS vs SAPB vs EIMC.

Aspect	TIPS	SAPB	EIMC
Scope	Emphasis on Inventive tasks and challenges of components design	Entire design process. Simple and difficult problems of systems design	Emphasis on Inventive tasks and challenges of components design
Task clarification	Laws of engineering systems evolution	General procedures	Laws of nonequilibrium thermodynamics, Onsager's theory.
Problem formulation	Identification of physical contradiction	Abstraction of essential problem	Setting of input and output values, performance coefficients
Systematic methods for solutions generating	Functions coupled to physical effects and examples of Standards Principles	Functions coupled to physical effects. Design catalogues	1) PTE catalogues are added with a set of performance characteristics, that allows you to organize the choice of solutions to aggregate performance 2) Each PTE has morphological matrix of variety of its technical implementations. It allows you to consistently improve the founded solution
Solution space	Focused - only "promising" directions are followed Minimal change of system	Large - "all" possible solutions considered	Solutions are ranked according to the aggregate performance, it allows you to select the best solutions
Product models	S-Field model	Design specification Function structure Concept (organ structure) Component structure	EIMC - energy-informational model of chains of different physical nature
Knowledge bases	Effects Patents Principles Laws of engineering systems evolution	Effects Design catalogues Engineering knowledge	PTE and morphological matrices for each PTE. Patents are grouped in a narrow class of technical devices and after grouped by the identified methods to improve performance
Learning time	Long time to learn	Short time to learn	Average time to learn
Computer support	Commercial	Research prototypes	Software Intellect-Pro - research prototypes

6 Conclusion

Analysis of the table shows that there are several differences of EIMC from other theories:

1) A description of each physical effect is formalized into a passport and morpho-logical matrix of possible technical implementations of this effect. It increases the number of synthesized technical solutions in several times.
2) The synthesis results can be ranged further due to preliminary expert assessment of performance characteristics. Thus each result can be assessed on the whole vector of performance characteristics, which is impossible in TIPS and SAPB.
3) Unlike TIPS EIMC has rigorous physical principles of nonequilibrium thermo-dynamics as the basis.

Thus, we can assume that these three methodologies can complement each other and lead to better solutions.

References

1. Chandrasegaran S.K. et al.: The evolution, challenges, and future of knowledge representation in product design systems. Computer-Aided Design 45, 204–228 (2013)
2. Wang L. et al.: Collaborative conceptual design—state of the art and future trends. Computer-Aided Design 34 981-996 (2002)
3. Sushkov V., Alberts, L. and Mars N.:. Innovative engineering design based on sharable physical knowledge. In: Artificial Intelligence in Design'96: Proceeding of the Interna-tional Conference Artificial Intelligence in Design, pp. 723–742. (1996)
4. Zaripov, M. F.: 1988, Energy-Informational Method of Scientific and Engineering Creativ-ity, VNIIPI, Moscow (In Russian).
5. Zaripov M., Petrova I., Zaripova V. : Project of creation of knowledge base on physical and technological effects, In: Joint IMEKO TC-1 & XXXIV MKM Conference Education in Measurements and Instrumentation - Challenges of New Technologies, Proceedings of TC-1 Symposium, vol.I, pp.171-176 (2002)
6. Petrova I., Zaripova V. : Systems of teaching engineering work on base of internet tech-nologies. International J. Information Technologies and Knowledge Vol.1, pp. 89-95 (2007)

7. Onsager L. : Reciprocal relations in irreversible processes I. Physical review 37, pp. 405-426 (1931)
8. De Groot S. R. and Mazur P. : Non-Equilibrium Thermodynamics. North-Holland Publish-ing Company, Amsterdam (1962)
9. Huhns M, Singh M. : Ontologies for agents. Internet Computing. IEEE 1(6): 81–3 (1997)
10. Zwicky F.: Discovery Invention, Research Through the Morphological Approach. McMil-lan, New York (1969)
11. Zaripova V. : Elaboration of automated system for supporting and training of creative thinking and designing for engineers (INTELLECT - PRO). In: 2-nd Advanced Research in Scientific Areas ARSA, (2012) (http://www.arsa-conf.com/archive/?vid=1&aid=2&kid=60101-45)
12. Malmqvist J. et al.: A comparative analysis of the theory of inventive problem solving and the systematic approach of Pahl and Beitz. In: The 1996 ASME Design Engineering Tech-nical Conferences and Computers in Engineering, pp.1-11 (1996)

An Algorithm for Multi-Source Geographic Data System

Chiang-Sheng Lee, Hsine-Jen Tsai, and Yin-Yih Chang

1 Introduction

Geographic data is very diverse and dynamic. It is becoming a critical part of computing applications. Traditionally, geographic data is captured, stored and displayed in a single data source. Over the last decades, there has been an exponential increase in the amount of geographic data stored and available from multiple data sources. System designers need to develop integrated systems that allow users to access and manage information from several scattered geographic data sources at the same time to come up with a satisfactory answer. One reason for such need has been that environments for data access have changed from centralized data systems into multiple, distributed data sources. Another more recent cause for the attention of integration technologies is the emergence of Geographical Information System and its needs for accessing data repositories, application and legacy source that located across the organization intranet or on the Internet [Smart et al 2010].

The geographic data may be unstructured or semi-structured, and usually there is no a regular schema to describe them. As the amount of geographic data grows, the problem of integration among multiple data sources becomes a critical issue in developing distributed geographic systems. Many researchers have been interested in resolving the heterogeneity between geographic data sources, and different solutions have been proposed [Janowicz 2008, Ghulam 2010, Tsai 2011].

While interoperability of distributed information system gains much attention and studies, there is comparison little written about to gain a better performance of data integration of multiple data sources and it has being studied by many researchers in different applications. In [Song 2007], they focus on box covering algorithms that cover a computer network with the minimum possible number of boxes. They demonstrated that such covering problems can be mapped to the well-known graph coloring problem and argued that the algorithms presented provide a solution close to optimal. [Hert et al 1996] presented an online terrain-covering algorithm for a robot moving in an unknown three-dimensional underwater environment. They showed that the path length of their algorithm is shorter and to be linear in the size of the description of the boundary of the area. The focus of our study is to develop an algorithm that covers a region with the minimum possible number of geographic data, i.e. maps.

2 The System

This study is based on a map integration system which embeds in a large distributed data environment. A data server resides on top of scattered database systems and allows applications to access data from remote databases. In order to test our algorithm, we implement a system that would mimic that of a distributed data environment. Figure 1 provides a simple illustration of the map integration system in a distributed data environment.

While there are different format of a geographic data, the focus in our study is in the form of maps. The basic structure of the data in our system is a bounding box which is based on a R-tree structure [Guttman, 1984].

3 The Process

The process starts with the server receiving the request from users/applications. The request, again, is in the form of a bounding box. We assume that there is no data source in the

C.-S. Lee (✉)
Department of Industrial Management, National Taiwan University of Science and Technology, Taipei City, Taiwan, R.O.C
e-mail: cslee@mail.ntust.edu.tw

H.-J. Tsai • Y.-Y. Chang
Department of Information Management, Fu-Jen Catholic University, New Taipei City, Taiwan, R.O.C.
e-mail: tsai.fju@gmail.com; 042833@mail.fju.edu.tw

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_55,

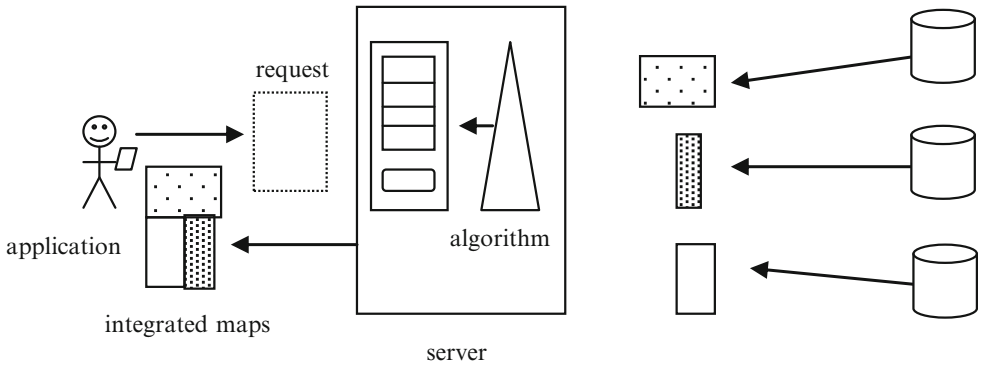


Fig. 1 The Map Integration System

distributed environment which can provide with a map that can cover the entire bounding box of the request. In another word, to response to the request, multiple geographic data need to be retrieve from the scattered data sources. The server of the system makes use of its database to locate data sources that can provide maps overlapping part of the requested map area.

The server then makes use of its algorithm to decide the covering sequence of those geographic data. By processing those geographic data one at a time, the region of the request area is partitioned into uncovered portion and covered portion. The process continues until either the collection of overlapped geographic data is empty or the requested map area is totally covered.

Not only accessing correct geographic data, but also performing the integration within limited time needs to be considered in a distributed geographic environment. The motivation of our algorithm is to increase the performance of the integration system by finding a set of possible minimum number of geographic data (i.e. maps). The algorithm not only finds the possible minimum set of geographic data but decides the covering sequence [Tsai et al 2013]. The following section gives more detailed descriptions of our algorithm.

4 The Algorithm and an Example

The objective of our algorithm is to find the minimum number of geographic data to cover the bounding box of the incoming request. A bounding box is an area defined by two longitudes and two latitudes and specified by the set of coordinates which represents the right-bottom and left-top of the bounding box. The right-bottom point is the minimum longitude and maximum latitude. The left-top point is specified by the maximum longitude and minimum latitude of the bounding box. Given a set of geographic data that partially overlap the request bounding box and a list of geographic data whose bounding boxes overlap part of the request bounding box, the algorithm starts the search by attempting to cover from the left-top corner of the request bounding box to the right-bottom corner of the bounding box.

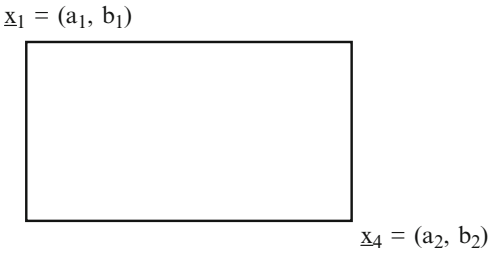


Fig. 2 A request bounding box with (a_1, b_1) and (a_2, b_2)

To achieve above goal, our algorithm is to decide the next adopted graphical data that will cover the current corner points as many as possible and repeat the same scheme until all corner points covered. Before introducing the main algorithm, we define the following notations.

1. Let P be the set of corner points of the region that has not be covered by the located geographic data. Initial values for P is the set of four corner points of the request bounding box. For example, $P = \{x_1 = (a_1, b_1), x_2 = (a_2, b_1), x_3 = (a_1, b_2), x_4 = (a_2, b_2)\}$ where x_1 and x_4 are the left-top and right-bottom corner points of the request bounding box, respectively. That is, a_1 is the minimum latitude and b_1 is the maximum longitude of the bounding box, a_2 and b_2 are the maximum latitude and minimum longitude, respectively. (see Fig.2)
2. Let $S_i = (s_{i1}, s_{i2}, s_{i3}, s_{i4})$ be the set of four longitude/latitude coordinates of the bounding box with (s_{i1}, s_{i2}) for the left-top corner point and (s_{i3}, s_{i4}) for the right-bottom corner point.

The algorithm includes the following steps:

Step1: Find the corner-point set P for the request area and set k=0 at the first time

Step2: Let ni be the number of corner points covered by the i'th geographic data from the database. Under the same maximum value of ni, put the last geographic data found into the list.

Step3: Set k=k+1 and execute the following two steps.

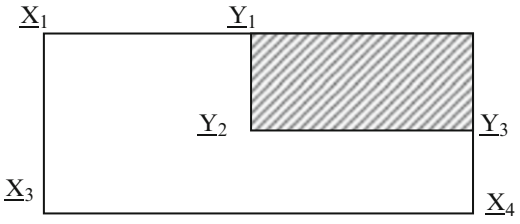


Fig. 3 The shaded graphical data after the first loop of the algorithm

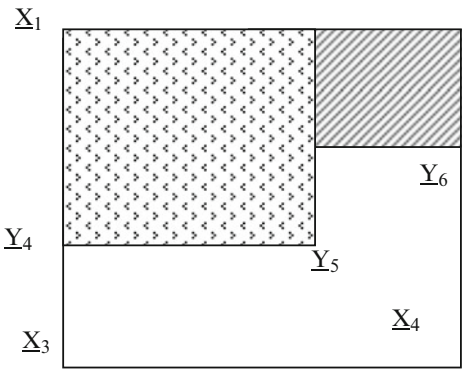


Fig. 4 The shaded graphical data after the second loop of the algorithm

(i) delete the corner points covered by the previous geographic data from the set P, and
(ii) add the new corner points to P if they are generated by the same geographic data.

If the set P is empty after Step3, then the algorithm stops. Otherwise, it goes to Step2.

The following example illustrates our algorithm's scheme. Let the bounding box of the request is specified by (a1, b1) and (a2, b2), the initial values are defined as follows:

$P = \{\underline{x}_1 = (a_1, b_1), \underline{x}_2 = (a_2, b_1), \underline{x}_3 = (a_1, b_2), \underline{x}_4 = (a_2, b_2)\}$ and $k = 0$.

Figure 3 shows the covered status of the request bounding box after algorithm finds geographic data that covers the request bounding box with maximum value of $n = 1$ and is shaded under the diagonal line. The set P has new corner points which are $P = \{ \underline{Y}_1, \underline{Y}_2, \underline{Y}_3, \underline{X}_1, \underline{X}_3, \underline{X}_4\}$

For one more example for this algorithm, Figure 4 shows that the algorithm finds the maximum value of n=3 and the second graphical data is also shaded. The new corner-point set P would be P = { Y3, Y4, Y5, X6, X3, X4} and K = 2. Since set P is not empty, the algorithm will continue until P becomes vacant.

5 Conclusion

In this work, we have demonstrated an approach to search mechanism in the context of geographic data integration. Our contribution has been the proposal of an algorithm to retrieve multiple geographic data to response a user's request of a geographic map covering a certain region. This approach leads to a process that ensures the possible minimum geographic data are located. The three steps of the approach allow such a property to be satisfied. The first step keeps the corner points of the bounding box of the request to ensure the total area of the request will be covered. By checking the corner points set the second step locates the geographic map that overlaps uncovered area and updates the corner points of the uncovered area. Since the geographic map located covers the maximum number of corner points the algorithm ensures a possible minimum number of geographic maps are located. The third step then updates the parameters of the algorithm and ensures the algorithm halts. As future work, we will consider a more complex algorithm that utilizes an optimization function.

Reference

1. P.Smart, C.Jones, and F.Twaroch, (2010) "Multi-source toponym data integration and mediation for a meta-gazetteer services", GIScience, LNCS 6292, 234-248.
2. K.Janowicz, M.Wilkes and M.AndlutzN, (2008) "Similarity-Based Information Retrieval and Its Role within Spatial Data Infrastructures", Geographic Information Science, pp. 151-167.
3. Ali Mohammad Ghulam, (2010) "A Framework for Creating Global Schema Using Global Views from Distributed Heterogeneous Relational Databases in Multi-database System", Global Journal of Computer Science and Technology Vol. 10 Issue 1 (Ver 1.0), pp. 31-35
4. A. Guttman, (1984) "A dynamic index structure for spatial searching", SIGMOD '84 Proceedings of the ACM SIGMOD international conference on Management of data, pp. 47 - 57
5. S. Hert, S. Tiwari, V. Lumelsky, (1996) "A terrain-covering algorithm for an AUV", Autonomous Robots 3, pp. 91-119.
6. Chaoming Song[1], Lazaros K Gallos[1], Shlomo Havlin[2], and Hernán A Makse, (2007) "How to calculate the fractal dimension of a complex network: the box covering algorithm", Journal of Statistical Mechanics: Theory and Experiment, pp.03006
7. Hsine-Jen Tsai, (2011) "A spatial mediator model for integrating heterogeneous spatial data", Dissertation, Iowa State University under the guidance of Dr. Les Miller
8. H-J Tsai, C-S. Lee and L. Miller. (2013) "A Search Mechanism for Geographic Information Processing System", In Proceedings of the Institute of Industrial Engineers Asian (IIE Asian) Conference, pp. 945 – 952.

Methodology and Platform for Business Process Optimization

Adam Grzech, Krzysztof Juszczyszyn, and Paweł Świątek

1 Introduction

The subject of the proposed solution relates generally to the issue of the knowledge-based business process management using advanced computational techniques in the tasks of business process optimization. The issue is addressed in two terms. The first is the methodology, which is devoted to elaborate methods and tool allowing, mostly automatically, translation of business process descriptions into proper and adequate mathematical models, and further into domain-specific optimization tasks. The second is the framework, which is devoted to compose, mostly automatically, service-oriented application solving the discovered, domain-specific optimization tasks.

Due to the independence of business rules management tools from the described processes, the same solution can be applied in different domain-specific management processes to solve different tasks and to address different purposes. These tools typically allow the simulation of the decision-making support system which gives the possibility to verify the effects of changes in management processes in the event of a changes of business rules, requirements and available resources. Such systems are often integrated with models of business processes, corporate databases, and internal and external computer communication systems. Currently, one of the key unresolved issues and in the area of business computing is the problem of formulating a universal, adaptable model of business rules. At the same time the number of decision-making problems and the optimization of operational (in particular in the area of transport systems) is huge, and the weight of their decisions - important for the efficiency of enterprises providing transportation services.

A. Grzech (✉) • K. Juszczyszyn • P. Świątek
Wrocław University of Technology, 50-370, Wybrzeże Wyspiańskiego 27 Wrocław, Poland
e-mail: Adam.Grzech@pwr.wroc.pl; Krzysztof.Juszczyszyn@pwr.wroc.pl; Pawel.Swiatek@pwr.wroc.pl

The main objective of the discussed approach is to propose flexible Methodology of selection of requirement analysis, modeling and data processing methods for analysis, planning and optimization purposes in management information systems in the field of transport. The general idea of the Methodology is based on assumption that many business process optimization tasks are based on the very similar mathematical models and that the possible, formulated based on the mathematical models, optimization tasks may be solved using the same scope of algorithms.

A key research need addressed in the framework of the Methodology is the development and integration:

- methods of description, modeling, processing, optimization and presentation of business processes,
- algorithms for solving real-world unconstrained and constrained optimization problems,
- methods of flexible and adaptive composing adaptive of service-oriented making decision support systems in the context of a universal platform, which, in particular, will be used as a generator of domain-specific information systems.

Innovation research and anticipated results of the research involves a systematic approach to integrate the three, usually (as demonstrated by analysis of the state-of-the-art) separated research areas related to the design, construction and implementation of integrated decision-making support systems. At the same time, it is planned to achieve innovative results in the form of the development of original techniques, methods and algorithms within each of these areas.

Achieving these expected outcomes requires the implementation of a number of research tasks relating to, inter alia, development of methods for the automatic composition algorithms to complex data and information processing, languages describing complex services with respect to their non-functional parameters and methods of translating business processes descriptions to the requirements addressing complex services supporting decision-making information systems. This is because of the necessity of individually

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_56,

designed algorithms to solve optimization problems in domain-specific decision-making support information systems, and the lack of methods for rapid prototyping of algorithms which solve unique optimization problems determined by specific business processes.

Known results on the task of developing methods for characterization, modeling, processing and optimization of resource management systems, passenger transport by road and rail point to numerous limitations and incompleteness of previously developed methods, including:

- The requirements (service level agreement) analysis at the same time in natural language and in a form that may be automatically (or almost automatic) processed in order to compare business processes and to obtain proper mathematical models of decision-making support systems necessary to formulate adequate optimization tasks.
- Obtaining, scalable and open platform to generate domain-specific making-decision support information systems requires a tool to assist in the verification of models of business processes through incremental verification of the model and assessing the current [1–4].
- In many cases, existing models are not sufficiently precise (due to the use of simplifications) reflect the reality in the management of resources (including vehicles), scheduling problems (determination of timetables) and other related optimization problems (e.g. the inclusion of additional resources planning the location of bus stops, routes passes) take into account the additional assumptions and requirements (e.g. availability of vehicles associated with failures, reducing their efficiency, variable number of passengers, passengers who are disabled, etc.) [5, 6, 8, 9]. Due to the fact that, in reality, parameters of such problems may be changed, modeling must take into account the issues related to the resources availability and prediction of resources availability.
- Build effective mathematical models requires industrial research in the field of modeling and description of the processes occurring in the tasks related to the management of resources in transport systems, among others, to develop a list of the processes of business organizations to model business processes, obtaining data about the resources of the organization (transport, conditions of transport processes) as well as contextual information (used for infrastructure, the business environment of the organization) [7].

The issue at work are associated with the project, which aims to provide the following objectives:

- development, implementation and qualitative and quantitative analysis of the effectiveness of methods of description, processing, comparison and optimization of resource management in the enterprise;
- development of methodologies for flexibility in the choice of methods of analysis, planning and optimization of enterprise resource planning systems in a domain-specific information system;
- develop a validation environment for the verification of the correctness and effectiveness of methodologies;
- quantitative and qualitative evaluation of the effectiveness of developed methodologies;
- develop platform for optimization of business processes in integrated information systems;
- verification methodology applicability, and its implementation as a Platform to automate manufacturing processes, optimize and maintain the real domain-specific information systems.

2 The problem of transport planning and Framework architecture

The terms used to in the proposed Methodology, issues of transport planning and optimization are formulated as follows (Fig. 1). The given business process, being an organization-specific set of rules, allows to identify requirements (criteria and constraints) necessary to formulate mathematical model and optimization problem (for example, set a new timetable). The given business process owner expects the optimal solution of the optimization tasks.

Form of the resulting solution is determined by the means of transport, which are at the disposal of the organization, the available communication infrastructure, as well as the location of resources (people, goods, etc.) and relevant to the task on their transportation needs (Fig. 1). In this situation, an appropriate solution of the given optimization task depends on adequate representation the information system of knowledge about the requirements and conditions under which the transport processes are performed as well as on the appropriate definition of data sources and the use of dedicated optimization algorithms.

If additional requirement is the possibility of rapid prototyping solutions, dedicated to the organization and to provide reusability of selected components of the Platform (in particular implementing optimization algorithms), this leads to the selection of a suitable service architecture for the implementation of the Platform. In particular, the implementation of the components of the Platform as a Service allows to use attempts and improve approaches known in existing service-oriented systems:

- Languages description of services - to describe the services used are XML-based languages, in accordance with the recommendations of the World Wide Web Consortium (www.w3c.com), such as WSDL or OWL -S. It is also used domain- ontologies (dictionaries), allowing the description of the functionality and interfaces of services.

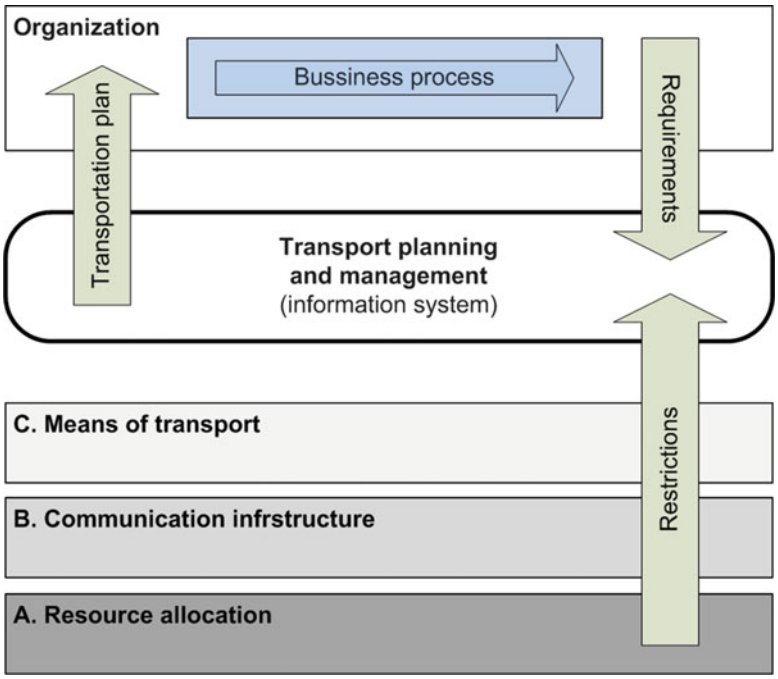


Fig. 1 Transport planning and optimization

- Mechanisms for complex service composition - a composition of services currently requires the most operator intervention and is done in semi - automatic mode [10], and there are not available tools supporting the operator in the tasks of generating descriptions of services and management environments to enable the composition [11], in particular - tools integrating service composition and service execution engines.
- Automatic translation of business processes descriptions - a significant problem of direct translation of the business processes representation into the demands for services which implement the functions of the process is solved in part; usually for selected languages and with a limited range of applications [12]. There are not available systems being both complex solutions and offering required level of openness and flexibility, allowing them to be easily adapted to business process optimization needs [13].
- Mechanisms of adaptation and integration – the proposed Framework of services for process optimization purposes is expected to be the Framework where the above mentioned issues and proposed solutions are integrated in gain to obtain reconfiguration abilities in case of changes resulting from the changes in ongoing business processes (organization providing transport services).

The choice of service-oriented architecture makes possible to utilize widely used dedicated domain-specific dictionaries (ontologies), allowing for a consistent description of software components, ensure the compliance of the messages between them and enabling composition of services in complex processes based on pre-defined requirements.

3 Methodology of flexible choice of methods of analysis, planning and optimization

The basic element that allows for the integration of the platform components, the use of appropriate data sources, selection of planning and optimization algorithms, and above all - an analysis of the organization's business process in a manner consistent with the objectives of enabling the use of an information system (frameworks) are ontologies - subject specific dictionaries containing terms and relationships that describe Platform components and the reality of the business process organization.

Among them are:

- dictionary of algorithms and problems;
- dictionary data types;
- glossary platform components (services);
- domain- and organization-specific dictionaries (terms specific to the organization, used to describe its business processes);
- glossary requirements (criteria and limitations for assessing solutions to problems).

The proposed Methodology involves the use of dictionaries available at the Platform for description - in the framework of properly conducted business process identification, which is implemented in the organization - of the transport process and its determinants, together with the performance and evaluation as well as constraints (Fig. 2).

Described – in terms of available dictionaries - business process may then be associated with proper mathematical models necessary to formulate adequate, recognized at the

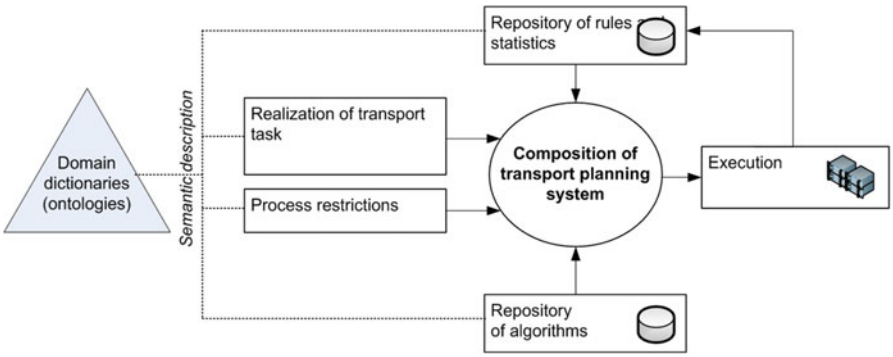


Fig. 2 Methodology of analysis, planning and optimization of transport processes.

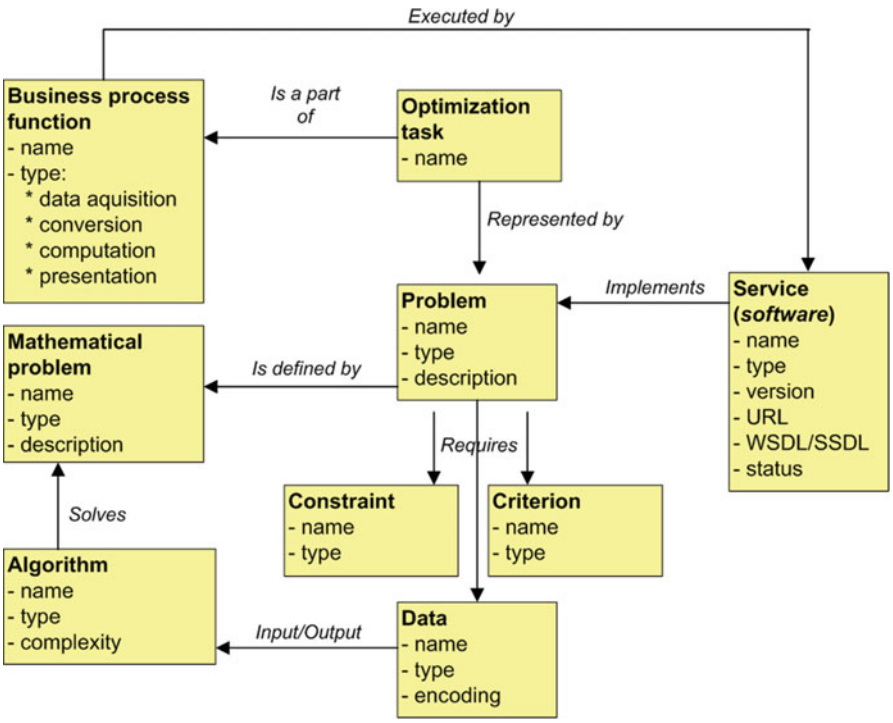


Fig. 3 Simplified structure of the Platform's dictionaries

business process analysis stage, optimization tasks. The latter may be further associated with proper algorithms – in form of services being a software components - that allow to solve the identified optimization task under additional constraints (time, accuracy, computational and memory complexity, etc.). Such software is a complex service, which consists of elementary services, among are the following services:

- computational,
- data access,
- data processing,
- data communication,
- user interface.

The above mentioned complex service is delivered by being part of the Platform runtime environment and is subject to monitoring and assessing the quality and efficiency of execution.

Linking the task of business process model identification with the selection of proper software components, available at the Platform in Software-as-a-Service (SaaS) mode, for the implementation of the process optimization is possible thanks to the original structure of dictionaries (ontologies) is presented in simplified form in Fig. 3. The proposed structure of dictionaries is based on experience collected during previous works relating to service-oriented systems for Scientific Workflows information systems [14, 16, 17].

Computing services, available at the discussed Platform, are devoted to solve optimization, analysis and planning problems, each of which corresponds to at least one mathematical problem (which may be formulated in different ways). The mathematical problem is solved by an appropriate algorithm, while maintaining the nature of the problem arising from constraints and criteria. An important solution to the application of the methodology is to conduct the

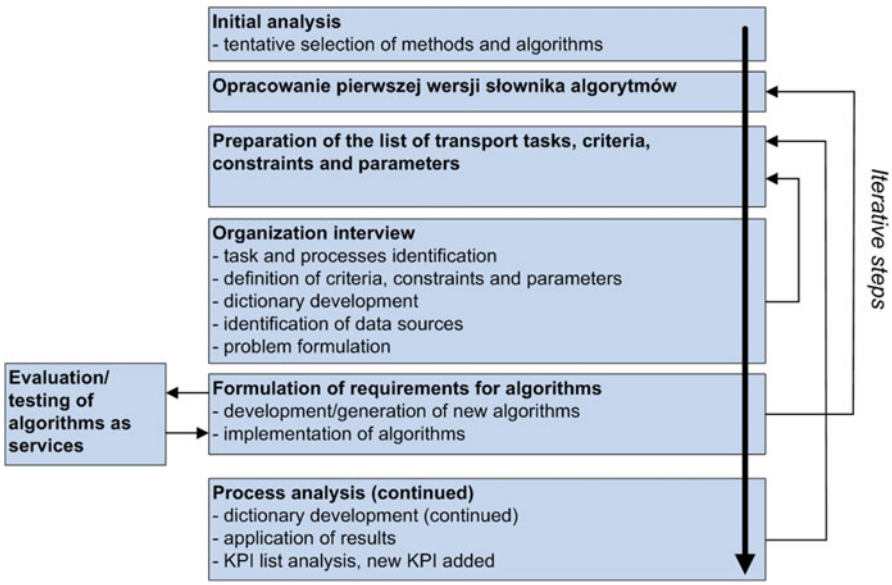


Fig. 4 Key steps of transport process' optimization in our Methodology

business intelligence organizations so as to identify the transport process optimization problems and the resulting actual computational problems. Such a task is then regarded as one of the activities of the business process is implemented by the appropriate service computing.

Because it is not expected that implemented anytime computing components at the Platform will exactly match the needs of the organization and its specific (in terms of mathematical models, optimization problems and computational efficiency) optimization problems, an iterative approach is assumed in the proposed Methodology. Iterative approach gain is to deliver as soon as possible solution for recognized and identified optimization process and the obtained results are the starting point for further, if required, deeper investigation of the given business process and environment within which the process is performed. The further investigation is to extend business process description and dictionaries in gain to recover detailed characteristics of the business process than represented in mathematical models and optimization tasks formulation (Fig. 4).

Before the first contact with the organization carries out transport processes takes place initial analysis of the field and the first version of the specification of domain vocabularies (problems, algorithms, criteria, constraints and generic data types). Business intelligence is to identify optimization tasks and problems of the organization and describe its initial business processes in a manner consistent with the existing dictionaries. An important element of the interview is to identify the sources of data in organization information systems, and describe the methods available to them in order as soon as possible to use in the implementation of optimization and planning (e.g., planning timetables, network storage, etc.). The interview allows the formulation of requirements (in terms of the criteria, constraints, and an input - output data on algorithms for solving identified and formulated optimization problems. They can already be implemented as components of the Platform, or there may be a need for their design and implementation. In parallel, there is replenishment of domain dictionaries for the necessary concepts and relationships. Due to identify and describe appropriate data sources, it is possible to obtain computational solution of optimization tasks occurring in the investigated business processes within the given organization and to verify whether such a solution is suitable for the business process owner.

4 Conclusions

The proposed approach, involving the innovative use of a Platform which offers dedicated services along with domain-specific dictionaries and Methodology for selection analysis, planning and optimization methods for domain-specific business processes in the field of transport. The Platform is an integrated approach to the business process modeling, optimization problems formulation and software prototyping.

The Methodology is implemented as logic of Platforms for business processes optimization, the functionality of which should lead to increased efficiency and lower costs of each stage of the software life cycle (i.e. phases: requirements analysis, design, implementation, testing, deployment, maintenance and adaptation of information systems, cost reduction the various stages of manufacture of domain information systems).

The proposed Methodology allows to speed up development and reduce the cost of software development at all stages of the software life cycle: requirements analysis, design, implementation, validation, implementation, maintenance and product development.

It should be emphasized that special attention is paid to the possibility of demonstrating the results of the optimization process at an early stage and their use in subsequent phases of software development.

Taking into account that for the forthcoming Platform for all the above results will be used together, as part developed Methodology, a very important factor is the occurrence of innovation and the use of synergies between them. The proposed Platform will also have the functionality of the unknown in the current market solutions in the field of manufacturing of complex information systems, and allowing, among other things:

- automatic planning, adaptation and optimization of the logical architecture of domain information systems
- automatic adaptation of the modular system for the purpose of supporting the implementation of processes for which the system was not designed,
- automatic composition of software components available functionality defined in the domain-specific business processes.

Acknowledgments The research presented in this paper was partially supported by the European Union within the European Regional Development Fund program no. POIG.01.03.01-02-079/12.

References

1. Marouane K., Houari S., Mounir B.: Model Transformation as an Optimization Problem. MoDELS 2008, LNCS 5301, 159–173, (2008)
2. Lin Y.: A Model Transformation Approach To Automated Model Evolution. The University of Alabama at Birmingham (2007).
3. Mens T., Van Gorp P.: A Taxonomy of Model Transformation. Electronic Notes in Theoretical Computer Science 152, 125–142 (2006)
4. Tisi M., Jouault F., Fraternali P. Ceri S., Bezivin J.: On the Use of Higher-Order Model Transformations. Proceedings of the 5th European Conference on Model Driven Architecture - Foundations and Applications, 18–33 (2009)
5. Ceder A.: Public-transport vehicle scheduling with multi vehicle type, Transportation Research Part C: Emerging Technologies, 19, 485-497 (2011)
6. Khisty, C.J., Lall, B.K.: Transportation Engineering: An Introduction. Prentice Hall (2003)
7. Mens T., Van Gorp P.: A Taxonomy of Model Transformation. Electronic Notes in Theoretical Computer Science 152, 125–142 (2006)
8. Rudek, A., Rudek, R.: A note on optimization in deteriorating systems using scheduling problems with the aging effect and resource allocation models, Computers & Mathematics with Applications, 62, 1870–1878 (2011).
9. Zäpfel G., Bögl M., Multi-period vehicle routing and crew scheduling with outsourcing options. International Journal of Production Economics, 113, 980–996 (2008)
10. Agarwal, V., et. al.: Synthy: A system for end to end composition of web services, World Wide Web Conference 3(4), pp. 311-339 (2005).
11. Ponnekanti, S. R., Fox, A.: SWORD: A developer toolkit for Web service composition, 11th World Wide Web Conference, pp. 97-103 (2002).
12. Hackmann, G., Gill, C., Roman, G.: Extending BPEL for interoperable pervasive computing. IEEE International Conference on Pervasive Services, pp. 204–213, (2007).
13. Swiatek P., Stelmach P., Prusiewicz A., Juszczyszyn K.: Service Composition in Knowledge-based SOA Systems. New Generation Comput. 30(2-3): 165-188 (2012).
14. Grzech A., Juszczyszyn K., Kołaczek G., Kwiatkowski J., Sobecki J., Świątek P., Wasilewski A., Specifications and Deployment of SOA Business Applications within a Configurable Framework Provided as a Service, Studies in Computational Intelligence, 2013, Springer, pp. 7 – 71.
15. XuboFei, Shiyong Lu: A Dataflow-Based Scientific Workflow Composition Framework. IEEE T. Services Computing 5(1): 45-58 (2012).
16. Diamantini C., Potena D., Storti E., Ontology-Driven KDD Process Composition, Advances in Intelligent Data Analysis VIII, Lecture Notes in Computer Science Volume 5772, 2009, pp 285-296.
17. Oliveira, D., et.al., Ontology-based Semi-automatic Workflow Composition, 2012, JIDM 3(1): pp.61-72.

Review and Refined Architectures for Monitoring, Information Exchange, and Control of Interconnected Distributed Resources

Y. V. Pavan Kumar and Bhimasingu Ravikumar

1 Introduction

The world wide gap in the demand and supply of the electricity for the current needs and future projections is a major cause of concerns for the entire humanity. The rising concern on the greenhouse gas emissions, shortage of the fossil fuels, and rapidly growing economies in the developing world are forcing for a major research and development in the area of the alternative and renewable energy sources [1]. Smart grids and micro grids are some of the new concepts that are developed recently [2]. Various models are under development which leads to the growth of the new and smart EPS architectures.

IEEE 1547.3™-2007 guide is primarily concerned with MIC between the DR unit controller and the outside world. However, the concepts and methods should be helpful to designers and implementers of communication systems for loads, EMS, SCADA, EPS, equipment protection, and revenue metering. In general, the architecture for MIC of DRs is intended to facilitate the interoperability of DRs and help DR stake holders to implement MIC to support technical and business operations of DR and transactions among the stakeholders. The architecture design is primarily to represent the possible implantation of such a system and different views of the architecture and topologies [3]. Also, the new DRs introduces a set of systems, controllers to the existing automation system and both these existing and new systems should communicate with each other for safe and reliable EPS operation. IEEE 1547.3™-2007 guide provides use case methodology and examples (DR unit dispatch, scheduling, maintenance, ancillary services, and reactive supply) rather than addressing the economic or technical viability of specific types of DRs. Hence, this paper uses the fundamental concepts of this guideline and proposes refined system architectures for integrating DRs to the existing EPS in different views for MIC.

Y.V.P. Kumar (✉) • B. Ravikumar
Department of Electrical Engineering, Indian Institute of Technology Hyderabad (IITH), Hyderabad, India
e-mail: ee14resch01008@iith.ac.in; ravikumar@iith.ac.in

2 Conventional Architectures of the System

2.1 Introduction to Fundamental System Organization

Architecture of a system constitutes the combination of hardware and software components connected in a specific pattern to meet the objective of the business. IEEE 1471-2000 defines the architecture as "The fundamental organization of a system, embodied in its components, their relationships to each other and the environment, and the principles governing its design and evolution" as shown in Fig. 1. Different stakeholders may have different views of the architecture of the system depending on their particular concerns. The success of the system depends critically on its ability to meet the wide range of requirements placed on it. Many of these are addressed by the overall structure and behavior of the system, i.e., its architecture. IEEE 1471 provides a conceptual framework for identifying stakeholder concerns and then responding to them with appropriate architectural models. The key concept is that of an architectural "view point", which define a set of stakeholder concerns and the modeling and analysis techniques that can be used to address those concerns.

IEEE 1471 is gaining considerable momentum in the software industry with an increasing number of prominent standards and tools demonstrating adoption and alignment with the recommended practice. Software architecture and IEEE 1471 in particular clarifies what is being designed and why it matters. Sometimes "architecture" is used to describe the high level design of a particular component of a system. However, it is the end-to-end structure and behavior of the system that is the primary concern of the architect. In considering this, the architect deals with high level abstractions of the system that represent structures and behaviors that

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_57,

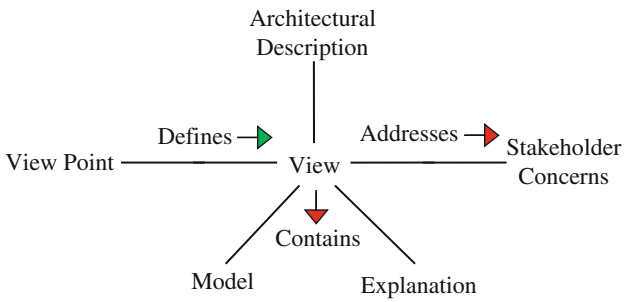


Fig. 1 IEEE 1471 - Fundamental Organization of a System

may not be readily apparent to someone responsible for the implementation of one particular component of the system. The idea of architectural "views" is also not entirely new, although previous approaches typically prescribed a fixed set of views to be applied to all systems regardless of the nature of the system or the specific concerns of stakeholders.

2.2 Architecture Modeling

Much of the intellectual effort in software design is involved with the creation of models. The key reference on which this series of whitepapers is based is IEEE standard 1471-2000, IEEE Recommended Practice for Architectural Description of Software-Intensive Systems, in which architecture is described using a set of models, organized into a set of views [4]. There are a variety of models and the choice of models used in any development effort is directed by the nature of the project, the application domain, the stage of development, the ability of those involved to create, comprehend, and the availability or otherwise of tool support. In particular, models that are suited to one domain may be ill-suited to another. Important aspects of a model are given as follows.

- Communication: Models provide a common language that can be used to communicate about a particular aspect of the system.
- Design: Some models capture a system's design; that is, the model contains the information needed to then proceed to create that system (or part of it), without including all the specific details of how exactly that takes place.
- Specification: A model can be used to specify particular properties of the system. The complete system exhibits the same properties as the model.
- Prediction: A model can allow us to predict certain properties of the completed system relatively cheap.
- Simulation: Some models support simulation of some aspect of the system. That is, a program executes the representation expressed in the model, in order to predict (by computation rather than analytically) properties or output of the system.
- Verification: Models with a formal mathematical basis can be verified against certain criteria. A state machine can be checked for absence of deadlock, for example.

2.3 Models in Software Design

As software systems are so complex, it often creates models that provide us with intellectual traction, by allowing us to focus on the things that are important for a particular purpose, while ignoring those that are not. In software systems development, models are typically not mathematical, but structural or behavioral. For an example, the interconnection structure between components in a software system, without having to know about the directory structure of the source code tree. Or, it can be examined that the sequence of events that takes place between these components in response to a user action, without needing to know the details of the transport protocol. The more widely used modeling language in software systems design is Unified Modeling Language (UML). It includes a number of notations for expressing object-oriented models: class diagram, sequence diagram, collaboration diagram, state charts, use-case diagrams etc [5-8]. Petri nets, flow charts, de Marco's Data Flow Diagrams etc are models that have been applied in software design for decades. More specialized models are used in specific areas of software design, such as task models for user interface design and entity-relationship models for database design.

Fig. 2 provides a reference overview of IEEE 1547.3 guidelines for MIC of DR interconnection. The diagram identifies the components that participate in processes of interest. These components are the subjects, or actors, of the process descriptions. The upper ovals represent the roles of stakeholders who may need to exchange information with the DR system about its interconnection with that area EPS. The DR units are represented by the hexagons. There may be one or many DR units at a site, but there will be at least one DR controller that performs a monitoring and control function. The DR controller has the intelligence with which to collaborate with stakeholders and site equipment. The DR units and controllers can be installed in a variety of configurations. DR units and DR controllers may be packaged together or separately, depending on the business strategy of the manufacturers and the requirements of the clients. The bottom small oval represents a load. Some loads may be facilities with facility EMS controllers to optimize their operations. A building facility controller is represented as a rectangle and labeled building EMS. The building EMS represents the intelligent component and need to collaborate with DR operator to ensure appropriate interaction between the DR controller and building EMS. In

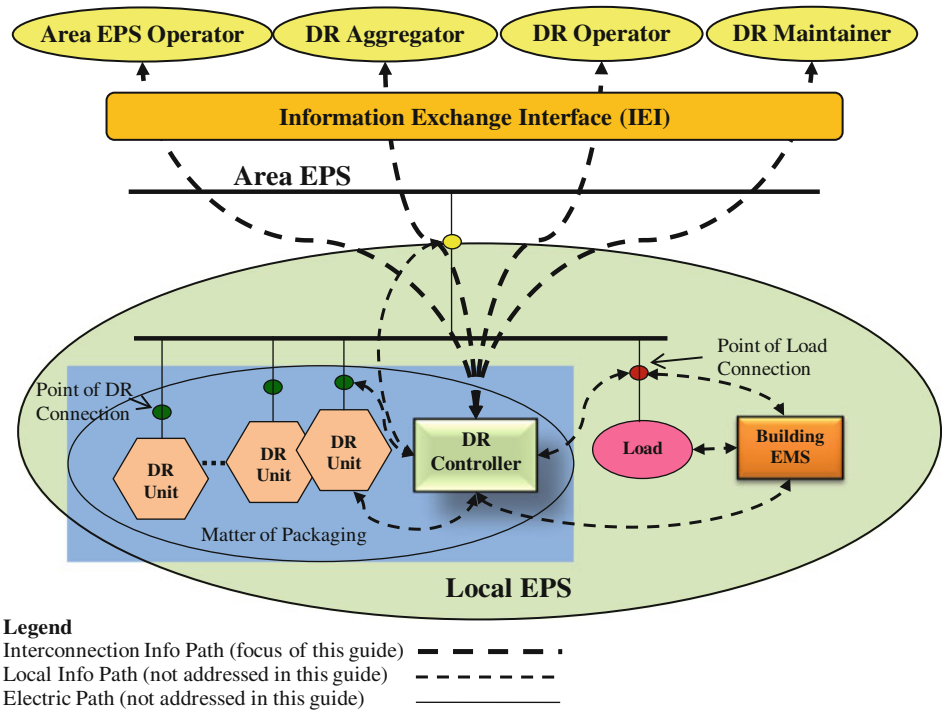


Fig. 2 IEEE 1547.3 - Schematic for Information Exchange.

Fig. 2, solid lines represent electrical connections and dashed lines represent communication paths. The communication paths are further sub-divided into those that are relevant to the guidelines and those that are acknowledged as important in an installation but are not in the focus of IEEE 1547.3. The IEI could be an actual single point of interface for all remote information flows to and from the device, or it could be an abstraction that represents information flows by multiple, but coordinated, physical media. IEI is information exchange counterpart of point of common coupling in electrical system.

2.4 Domain-Specific Models

In some cases, models may be based on a more rigorous foundation than is typically available for general-purpose software development. These foundations are often mathematical in nature. For example, a system that processes discrete-time signals may be specified using the language of signal theory-time and frequency domain, filter theory, Z-transforms etc. The presence of a formal language to describe the domain leads to models that can guide and support software development [9].

Real-time systems development is an example of a domain in which considerable work has gone into using models to express important system characteristics. State machine formations are commonly used to model reactive systems, in which external inputs trigger activity and a change of state. While state machines are commonly thought of as being expressed diagrammatically, there are also textual languages for programming reactive systems, such as the synchronous-reactive family of languages. A formal underlying model can provide an automated tool with enough information to "reconstruct" the information elided from the model. This can only work if the domain is sufficiently constrained that an automated tool is able to do this reconstruction. Many of the software systems don't meet these criteria. Often, there is no formal model at all. What formal underpinning is there for a user interface, for example? Other times, while it might be possible to create a formal model, the specific problem domain is not sufficiently well-studied for one to exist, let alone for there to be tools to express or synthesize code from them. And enterprise-scale systems such as SCADA are not sufficiently constrained to support this kind of modeling.

Even in the presence of a formal model, other concerns may limit the utility of those models beyond the understanding and simulation phases. Performance is one: code generated from models usually has poorer performance than hand-written code. Scalability is another: when you need to partition a program across multiple processes and compute nodes, more knowledge is needed than can be reconstructed from an abstract model. Finally, it is worth remembering that there is a fundamental difference between formal models and their realization in a software system [9-11]. Formal models are declarative-that they say what a thing is. Computer programs are imperative-that they say how you do something. Much of the work in software

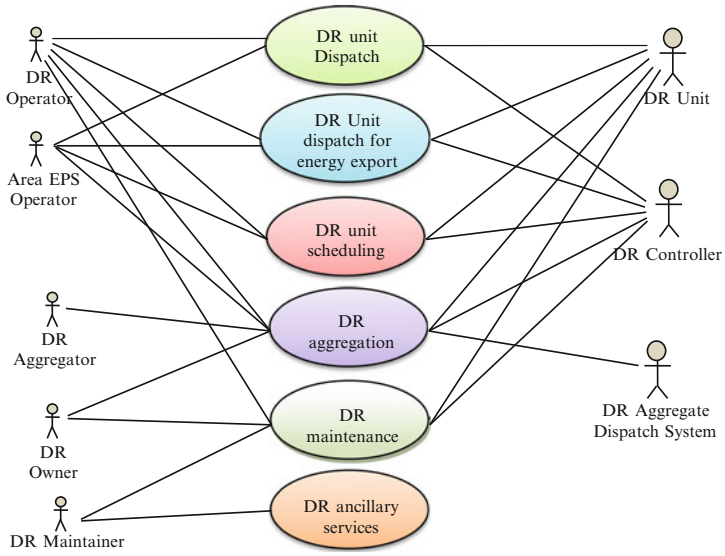


Fig. 3 Distributed Resource System Architectural View

systems development is all about figuring out imperative representation of the system's function.

3 Proposed Architectures of the System

The architecture is in actuality a set of models that uses to support the design of the system. It typically also includes a lot of other supporting information, such as scenarios, constraints, technology evaluation etc. Architectural design models have, in general, a key difference to software design models. In a UML class diagram, each class has an exact correspondence in the source code: a class in C++ or some other object-oriented programming language. The attributes and operations of the UML class correspond to data and function members of the C++ class. In architectural models, however, there is generally no such obvious connection between the model and the system implementation. The choice of exactly how the architecture is modeled is very fluid, and there is no industry consensus on exactly how to go about this.

The close connection of object-oriented design models to implementation code could be seen as an advantage, as it makes easier to learn how to create UML models. It does, however, also have the significant disadvantage that there are now two artifacts that represent much of the same information; the design models, and the code. As a result, the models and the code tend to drift out of sync over time. One solution to this dilemma is round-trip engineering, in which the code is annotated with information that enables the modeling tool to reconstruct changes to the model when changes are made to the code. Another is to attempt to use the model as the sole design artifact, and generate the code from the model. This approach is currently successful only in restricted application domains. Architectural models have no such close correspondence to code. In the architectural view shown in Fig. 3, processes and threads are modeled. But the correspondence to code is much looser than UML.

This creates a separation of concerns between architectural modeling and software development that enhances the value of the architectural models. That is, the architectural models capture and highlight information that is not made explicit elsewhere. They provide the common language for architectural design that is otherwise missing from the high-level aspects of system design. This, along with the relative brevity of architectural models, is a compelling argument for focusing on architectural models in preference to object-oriented design models. The high level of abstraction present in many architectural models can, on the other hand, lead to them being criticized as being vague. This is understandable, since they are typically attempting to model a complex system by the time enough information is removed to make the model intellectually tractable, the models can indeed seem to be vague. Another source of difficulty is lack of clarity about what exactly should be described in an architectural model. A generally accepted approach to architectural modeling that mitigates these difficulties is the creation and use of multiple views of a software system [11]. The concept of view is formalized in IEEE 1471, which describes how to construct a systematic framework for architectural modeling. According to this, the description of an architecture hinges around the concept of viewpoint, which specifies a collection of specific concerns of interest to a specified set of stakeholders, and the ways in which those concerns are expressed and addressed. A viewpoint includes description and usage of models in that viewpoint. The proposed view points of the system are divided into requirements, conceptual, concurrent, and network views as given below.

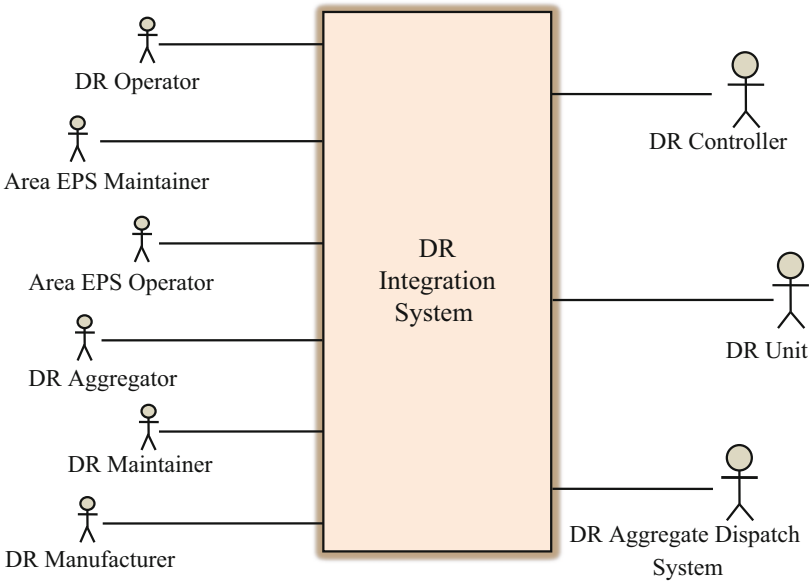


Fig. 4 Distributed Resource System Architecture - Requirements View

3.1 Requirements View

The requirements view expresses the system requirements and context in a manner that highlights and focuses attention on the architecturally-significant requirements as shown in Fig. 4. Specific use cases, quality attributes, and architectural risks are called out and explored during the process of creating this view. This captures and refines the significant requirements that serve to shape the architecture, in order that the architecture can be tested against them, and vice versa. Often, non-functional requirements are said to shape the architecture of a system. That is true to a large extent, as specific requirements in areas such performance, capacity, scalability, security, and reliability, can have significant impact on the system structure along with other constraints such as limitations on cost, feasibility, development skills etc. These constraints are also captured in a requirements view. But ultimately, architecture must meet the needs of its stakeholders. The requirements viewpoint therefore considers the user situation and the goals of the stakeholders as primary drivers.

3.2 Conceptual View

The conceptual view expresses the system architecture in terms of its functional capabilities. A completed conceptual view ensures that all key functionality has been considered by the architect team. This view describes the system in terms of conceptual components and key information flows between these components as shown in Fig. 5. Conceptual components are defined to be a set of domain-level responsibilities. This ensures that the conceptual model remains rooted in the problem domain from which the requirements are drawn. Conceptual components represent coherent chunks of functionality and they are connected by primary information flows through system. Modeling these two aspects of the system is sufficient to reveal the high-level structure of the system and its architectural styles. External interfaces are also a key part of the conceptual model. These interfaces help to illustrate the ways in which the system interacts with its environment. At higher level, this view establishes a basic decomposition of the system into subsystems. This high level model gives most basic description of the system used to organize the way of system architecture development.

3.3 Concurrency View

The concurrency view expresses the run-time structure of the system in terms of concurrently executing components as shown in Fig. 6. The models in this view range from high-level, coarse-grained subsystems, down to detailed threading models. A thorough treatment of the system architecture in concurrency view ensures that key issues in DRs such as, performance, robustness, and scalability.

3.4 Network View

The overall positioning relative to a typical modern software development life cycle (SDLC), the Rational Unified Process, is shown in network view as shown in Fig. 7. In this diagram, the horizontal line represents progress through the system development lifecycle, while the height of each line represents the intensity of work put into each view at that time in the

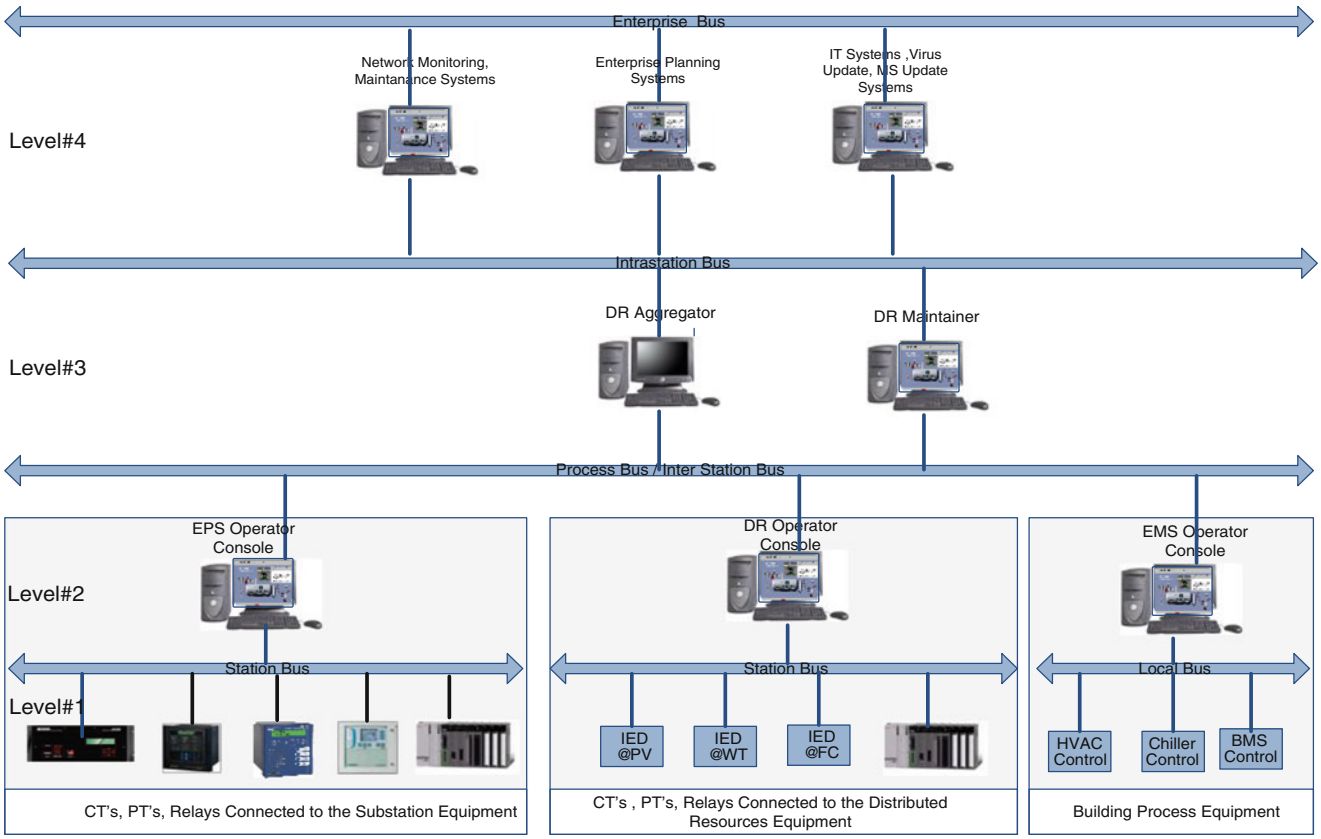


Fig. 5 Distributed Resource System Architecture - Conceptual View

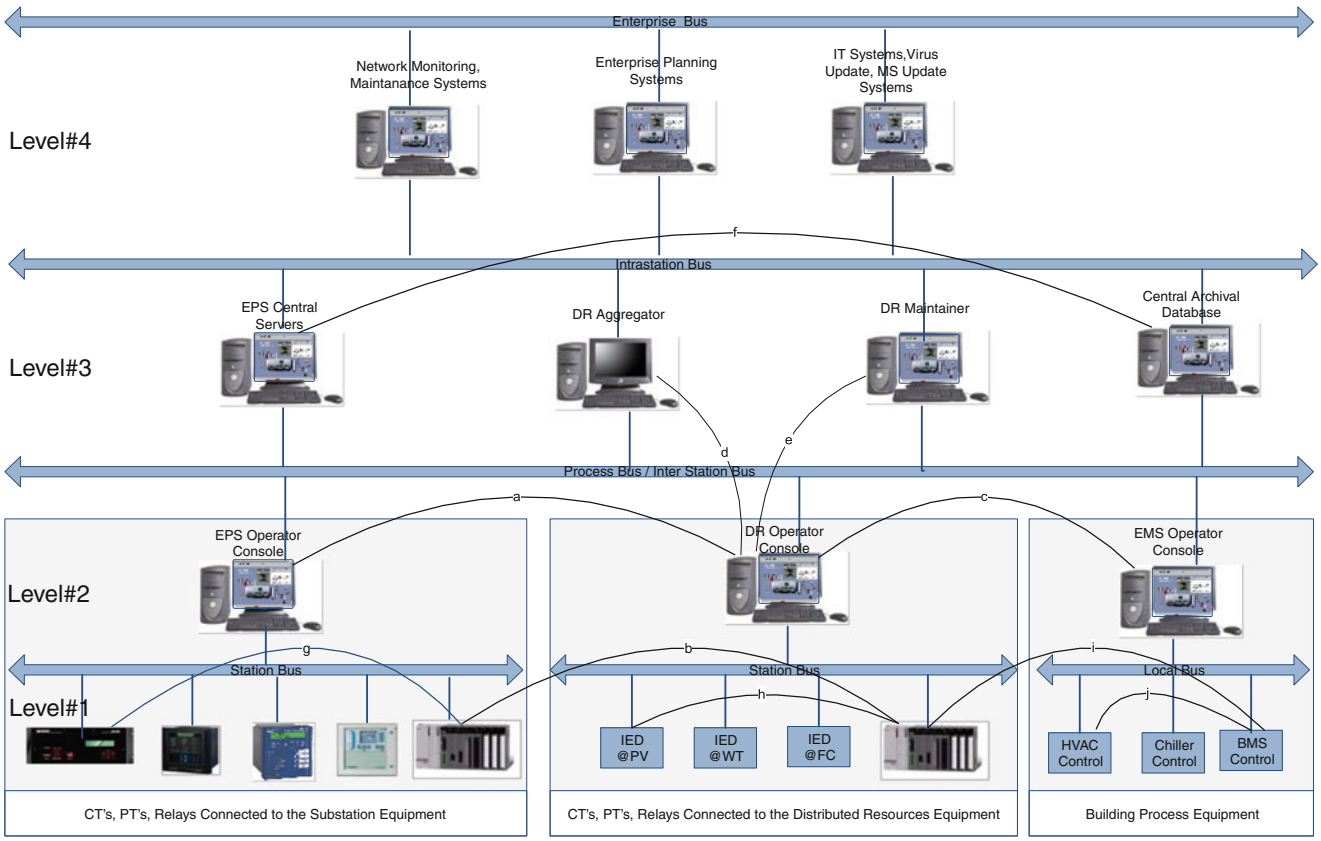


Fig. 6 Distributed Resource System Architecture - Concurrency View

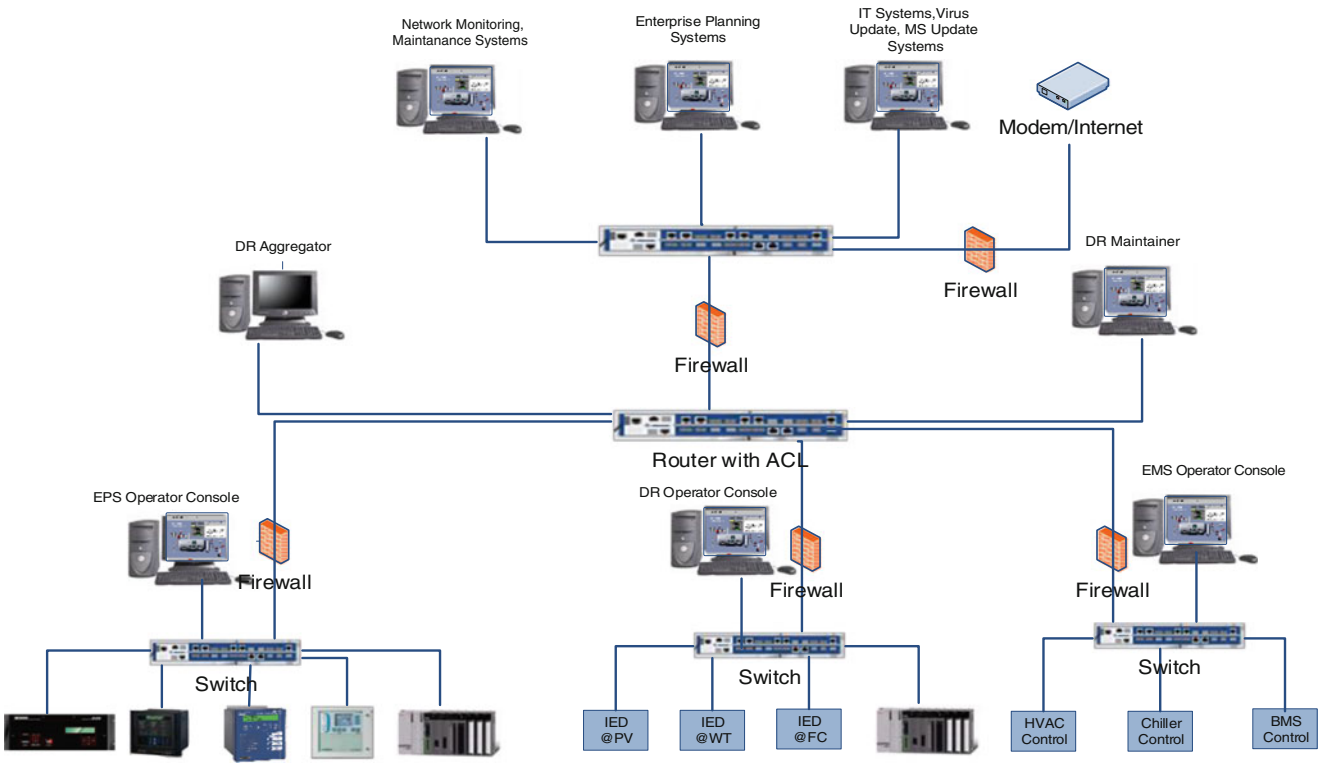


Fig. 7 Distributed Resource System Architecture - Network View

lifecycle. A large number of models can potentially be employed in these viewpoints, but each centers around one or a small number of models that give the viewpoint its essential character. This core model is supplemented by other models which can be brought into play when needed [12-13].

4 Conclusion

In this paper, the topology and architecture of a system meant for integrating the DRs in to the EPS is derived and presented in different ways. This review and refinement is done based on the guidelines of IEEE 1547.3™-2007. This will be useful to understand the design of DR technology topologies in a lucid manner. This helps in selecting best topology with respect to the application to achieve optimum utilization of resources and economy with the design of wide control and communication schemes.

References

1. H. Von, J. Schmid, Martin Faulstich, "Renewable Energies and Energy Efficiency," Kassel University Press, Band.14, Vol. 14, 2009.
2. Juan Carlos Vasquez Quintero, Josep M Guerrero, "Decentralized Control Management Applied to Power DGs in Microgrid-Renewable Energy Integration towards the smart grid", Lambert Academic Publishing, Germany, 2010.
3. Nick Rozanski, Eoin Woods, "Software Systems Architecture-Working with Stakeholders Using Viewpoints and Perspectives," 2nd edition, Addison-Wesley publications, NJ, 2012.
4. Michael R. Blaha, James R Rumbaugh, "Object-Oriented Modeling and Design with UML," 2nd edition, Prentice-Hall publications, NJ, 2004.
5. Grady Booch, "Object-oriented Analysis and Design with Applications," 2nd edition, Addison-Wesley publications, 1994.
6. R. J. A. Buhr, R. S. Casselman, "Use Case Maps for Object-Oriented Systems," Prentice Hall publications, 1999.
7. Paul Clements, Felix Bachmann, Len Bass, David Garlan, James Ivers, Reed Little, Paulo Merson, Robert Nord, Judith Stafford, "Documenting Software Architectures-Views and Beyond," 2nd edition, Addison-Wesley Professional, 2010.
8. Christine Hofmeister, Robert Nord, Dilip Soni, "Applied Software Architecture," 1st edition, Addison-Wesley Professional, 1999.
9. "IEEE1471-2000 Recommended Practice for Architecture Description of Software-Intensive Systems" Prepared by Software Engineering Standards Committee of the IEEE Computer Society, 2000.
10. Ivar Jacobson, "Object Oriented Software Engineering, A Use Case Driven Approach," 1st edition, Addison-Wesley Professional, 1992.
11. Ivar Jacobson, Grady Booch, James Rumbaugh, "The Unified Software Development Process", 1st edition, Addison-Wesley Professional, 1999.
12. Donald A. Norman, "The Design of Everyday Things," Basic Books, Reprint, 2002.
13. John Reekie, Rohan McAdam, "A Software Architecture Primer", Angophora Press, 2006.

Lossless Compression of Climate Data

Bharath Chandra Mummadisetty, Astha Puri, Ershad Sharifahmadian, and Shahram Latifi

1 Introduction

Data compression is the field where information is encoded using fewer bits than the original representation. It helps to improve the system performance and to reduce the redundancy in stored data. Data compression helps to transmit and store more information in less space, thereby, causing more storage space. By using data compression, significant reduction in climate data file size can be accomplished without loss of any important information.

Data compression is categorized as either lossy compression or lossless compression. Lossy compression techniques involve some loss of information and data compressed using this technique data cannot be recovered or reconstructed exactly. On the other hand, in case of lossless compression techniques, as the name implies, involves no loss of information. If the data is compressed using this technique, it is possible to recover the original data in exact form from the compressed data. [1] The compression technique i.e. lossless or lossy compression to be used depends on the application where it is being implemented. Some areas such as text compression do not have scope for lossycompression, they should be compressed using lossless compression only because if used with lossy compression a slight change or missing value can create a big difference with respect to the original data. On the other hand, in case where lack of exact reconstruction is not a problem, lossy compression is favorable. For example, when speech signal is stored or transmitted, the exact value of each sample of speech signal is not necessary. Depending on the quality required of the reconstructed speech, varying amounts of loss of information about the value of each sample can be tolerated.

B.C. Mummadisetty • A. Puri • E. Sharifahmadian (✉) • S. Latifi
Department of Electrical & Computer Engineering, University of Nevada, Las Vegas, 4505 S. Maryland Parkway, Las Vegas, Nevada 89154, USA

The most common way to measure the efficiency and accuracy of the compression technique is by finding the compression ratio which is defined as the ratio of number of bits required to represent the data before compression to the number of bits required to represent the data after compression. It is defined as the ratio of number of bits required to represent the data before compression to the number of bits required to represent the data after compression as in equation (1).

$$Compression\ Ratio = \frac{Uncompressed\ Size}{Compressed\ Size} \tag{1}$$

Another measure by which we can have a better compression of our data is by first analyzing about our data. The more we know about our data, the better we can compress it. With enough information about data's characteristics including the type, size, and statistical characteristics of data, it is possible to define or choose an appropriate method for data compression. [2]

We also worked on the prediction of climate data values using artificial neural networks, popularly known as ANN. The prediction of climate data is of concern to numerous fields like science, agriculture, environment etc. The prediction of climate data values can even help us in data compression especially when the predicted value is close to the original values. By the method of ANN prediction, desirable results and good accuracy are achieved for solar radiation.

Our experiments were performed on the data collected from the Nevada climate change portal which provides data collection of climate, hydrological, ecological, and hardware data and provides information pertaining to the state of Nevada. There are 12 sites in this data portal namely- sheep range mojave desert shrub, sheep range subalpine, sheep range blackbrush, sheep range pinyon-juniper, sheep range montane, snake range east salt desert shrub, snake range west pinyon-juniper, snake range east sagebrush, snake range west subalpine, snake range west

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_58,

sagebrush, snake range west montane, snake range east subalpine. Every site has data for 25 parameters, available in time gap of 1 minute and 10 minute. Also, average, maximum and minimum value for the same is available.

1.1 Related Work

We are basically focusing on lossless compression of climate data. A technique is presented in [3] to compress pressure, wind and precipitation data where uniform quantization is applied to the pressure data as a part of preprocessing the data. Optimal prediction techniques are used to predict the values based on the surrounding values and the differences are encoded. Wind data (wind velocity and wind direction) are transformed into polar co-ordinate form. Here entropy coding is performed after the preprocessing stage. Also, significant improvements in bandwidth can be realized through the use of common compression techniques such as wavelet/error grid or round/difference/BZIP2 compression [4]. The compression of temperature data in Kuala lampur from January 1948 until July 2010 by using Debauchies wavelet (D4) as the basis function is performed [5].

Many approaches for scientific data compression have been focused primarily on combining compression with data synthesis in order to increase throughput and conserve storage. Engel son [6] compressed sequences of double precision floating point values resulting from simulations based on ordinary differential equations. In theoretical approach the numbers are treated as integers and then compressed using predictive coding, with residuals being explicitly stored in case of lossless coding or truncated for lossy coding. Ratanaworabhan proposed a lossless prediction based compression method for double- precision floating point scientific data using a DFCM(differential finite context method) value method which is based on pattern matching using a hash table holding recent encoding context. The bitwise residual was then computed using XOR operator, with compressed representation consisting of finite number of leading zeros and remaining residual bits [7].

Weather forecasting is a vital application in meteorology and has been one of the most scientifically and technologically challenging problems around the world in the last decade. The use of data mining techniques in forecasting maximum temperature, rainfall, evaporation and wind speed works well [8]. This was carried out using artificial neural network and decision tree algorithms and meteorological data collected between 2000 and 2009 from the city of Ibadan, Nigeria. A data model for the meteorological data was developed and this was used to train the classifier algorithms. The performances of these algorithms were compared using standard performance metrics, and the algorithm which gave the best results used to generate classification rules for the mean weather variables. A predictive neural network model was also developed for the weather prediction program and the results compared with actual weather data for the predicted periods [9].

The investigation was done to develop artificial neural network for ambient air temperature prediction in Kerman city located in the south east of Iran. The mean, minimum and maximum ambient air temperature during the year 1961-2004 was used as the input parameter in feed forward network and Elman network. The values of R2, MSE and MAE variables in both networks showed that ANN approach is a desirable model in ambient air temperature prediction, while the results from Elman network are more precise than FNN network [10].

The organization of this paper is as follows. In Section 2, the proposed methods are described. Results and Conclusion is explained in Section 3 and 4 respectively.

2 Proposed Method

In this section, we present the overview and the details of the methods used for our study.

2.1 Overview

In this paper, a lossless data compression is proposed on climate data. The climate data from Nevada climate change portal is selected. Different types of sensors are placed in the desert of Nevada which measures climate parameters like temperature, pressure, solar radiation, photo synthetically active radiation and hardware related data like data logger power system voltage etc. and stored initially in files, which are then processed and stored in a database. Through Nevada climate change portal we get access to the database. Since it wastime series data we dealt with, a lot of information is stored minute by minute, so data compression is necessary. Our main focus is on solar radiation, voltage and photo synthetically active radiation data. The process includes three stages, as shown in the Fig 1. The first one is pre-processing of the data. Pre-processing of data improves data compression and data smoothing [9]. The Fig. 2 and Fig. 3 show solar radiation, photo synthetically active radiation data respectively for a week. Differential encoding is chosen to pre-process data. As shown in the Fig. 2 and Fig. 3 the data is similar to each other for the successive days. For these two climate parameters the data has leading and trailing zeros since these is no radiation in the early morning and night. This can be clearly observed from the graph.

The data logger voltage consists of non-integer data set. We converted it to integer format by multiplying it with

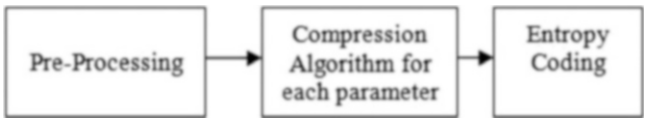


Fig.1. Schematic of proposed method.

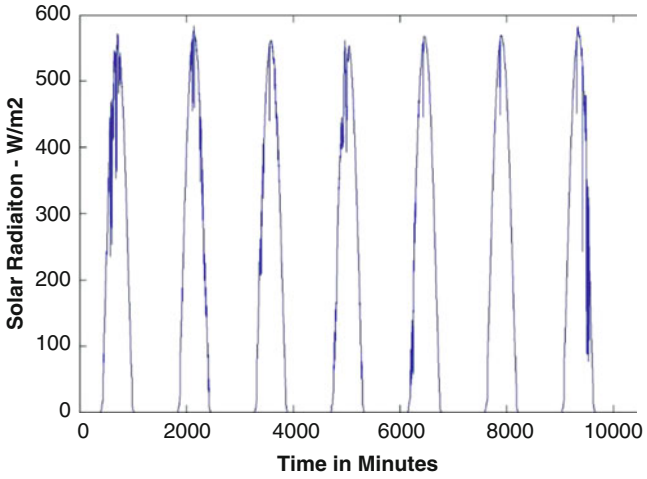


Fig. 2. Solar Radiation data (in W/m2) for 1-8 Jan 2012.

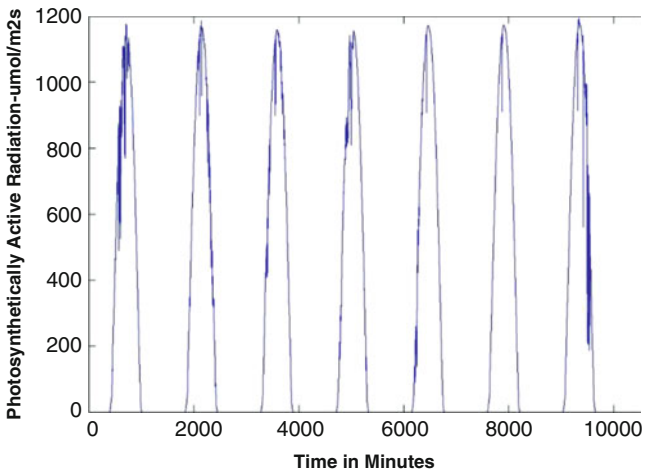


Fig. 3. Photo synthetically active radiation (in umol/m2s) for 1-8 Jan 2012.

hundred on which differential encoding was applied. Finally for all the above parameters Huffman coding is applied for data compression. All the computations and writing of algorithms is done in MATLAB.

Nevada climate change portal contains data for different parameters of climate for the years 2012 and 2013. For each site, parameter and date range, the corresponding data is available and can be downloaded in excel sheet. Solar

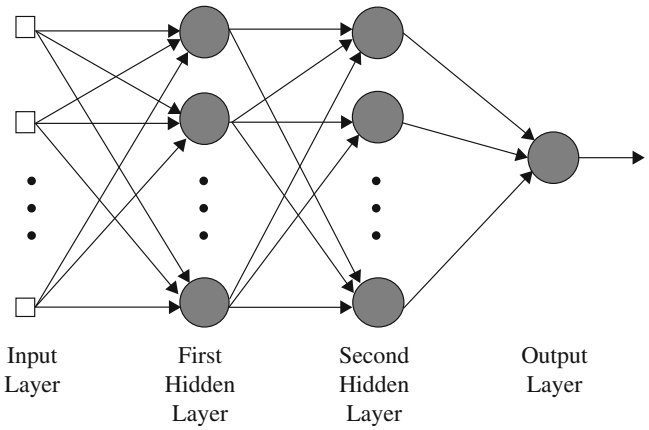


Fig. 4. Schematic of Neural networks

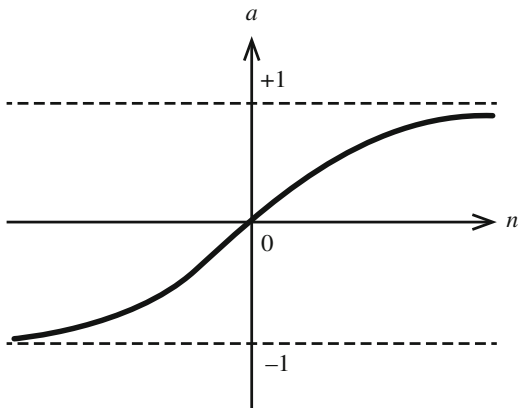


Fig. 5. Tansig transfer function.

radiation and photo synthetically active radiation are available in w/m2 and kw/m2. Data logger voltage has the units of V and KV. Temperature is available in degC, degK and degF.

A predictor model has been designed using artificial neural networks for the prediction of solar radiation data. The basic components of ANN are shown in the Fig 4. Our model uses a simple design using feed forward back propagation method. We use 6 inputs namely time, incoming longwave, incoming shortwave, outgoing longwave, outgoing shortwave, and photo synthetically active radiation data. The incoming (long wave and shortwave) and outgoing radiation data (longwave and shortwave) are measured in W/m2. The photo synthetically active radiation data is measured in umol/m2. The training data used for prediction is for 3 weeks in the month of January 2012. The corresponding values of solar radiation are used as output for the training data. We use Tan sigmoid as the transfer function which is represented mathematically as below and it is represented in Fig 5. We use 2 hidden layers each layer

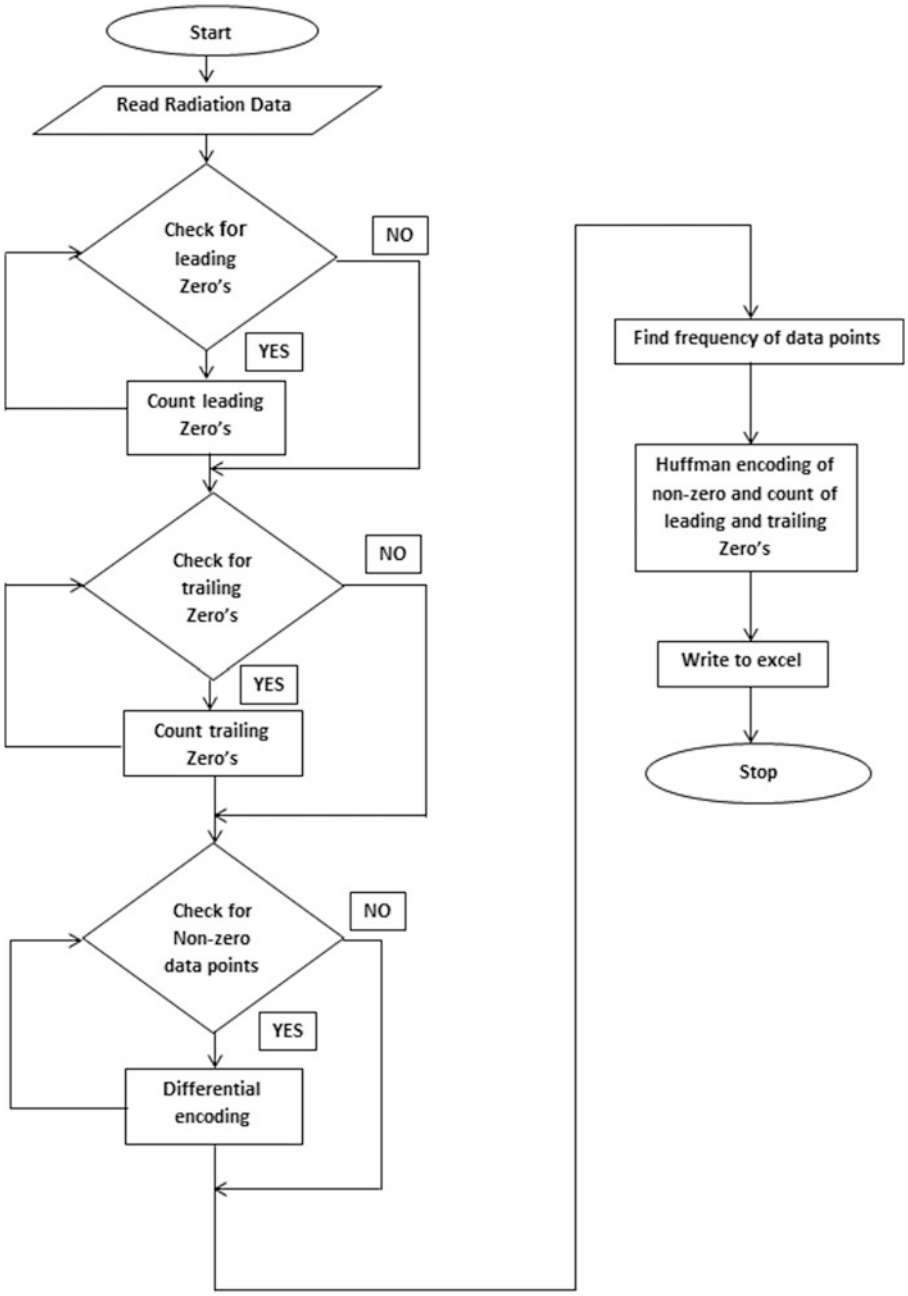


Fig. 6. Flowchart for Radiation data - Encoder.

having 10 nodes. By using this model we achieve a desirable compression ratio.

$$F_k(S_k) = \frac{2}{1 + \exp(-2S_i)} - 1 \qquad (2)$$

2.2 Methods

The methods utilized for the compression of three parameters-solar radiation, photo synthetically active data and data logger power system voltage are described in this section.

Solar Radiation and Photo synthetically active data.. The data is collected and analyzed from Nevada climate change portal for different days in the year 2012. The region selected is snake range black brush and data for 7th,14th and 21st day of different months are noted for implementing data compression for solar radiation and photo synthetically active data.The flow chart for encoder which shows the operation of the algorithm is shown in Fig 6. The algorithms functions as follows:

The solar radiation/photo synthetically active radiation data file is imported. The radiation data consists of series of leading zeros, non-zero data points, and trailing zeros.

The algorithm checks for the leading zeros and stores the count and the same process is done for the trailing zeros. Next, it checks for the non-zero data points and performs differential encoding, keeping the first non-zero digit as it is required during decoding operation. This is done as part of the pre-processing step. As the data points are close to each other to some extent, differential encoding gives smaller data points as compared to the actual data.Huffman coding is applied to the count of leading zeros, trailing zeros and the data obtained after applying differential encoding. The decoder reads the encoded data file for counts of zeros and differential encoded data. Both data sets are decoded, and the actual non-zero data points are obtained by the decoding algorithm.

$$X_{a1} = X_{d1} \tag{3}$$

$$X_{an} = X_{dn-1} + X_{dn} \tag{4}$$

Here X_{an} is the actual data. Differential encoding is not applied to the first data point as shown in equation (3), so the equation (4) is valid from the second data point. The flow chart for decoder is shown in Fig 7.

Data logger voltage data. The Data logger power system voltage data is taken for all months of year 2012. This data consists of floating point number up to a precision of 2 digits. As part of pre-processing, a hundred is multiplied to the data points to convert it into integers. Over this data, differential encoding is performed. The first data point is unchanged for the decoder operation. Huffman encoding is performed on these data points. The encoder algorithm is shown in Fig 8.

We then decode the data and with the help of equations (3) and (4) all the data points are obtained, and then by using the algorithm we divide the data points by hundred to obtain the actual data points. The flowchart for the algorithm is presented in Fig 9.

Prediction using Artificial neural networks for solar radiation data. For the Artificial neural network used in our design, the tansig function is used for the neurons in the hidden layer. The inputs are time, incoming longwave, incoming shortwave, outgoing longwave, outgoing shortwave and photo synthetically active radiation data.

The data sets are extracted from the input and target data for training, validation and test phases. The training set consists of 70 percent of data to build the model and determine the parameters such as weights and biases, validation data set includes 15 percent to measure the performance of network by holding constant parameters. And other 15 percent of data is used to increase the robustness of model in the test phase. The Fig 10 to Fig 13 show the plot of actual and predicted data for solar radiation for the dates January 23rd-26th 2012. The Fig 14 shows the plot of regression for training, validation and test data.

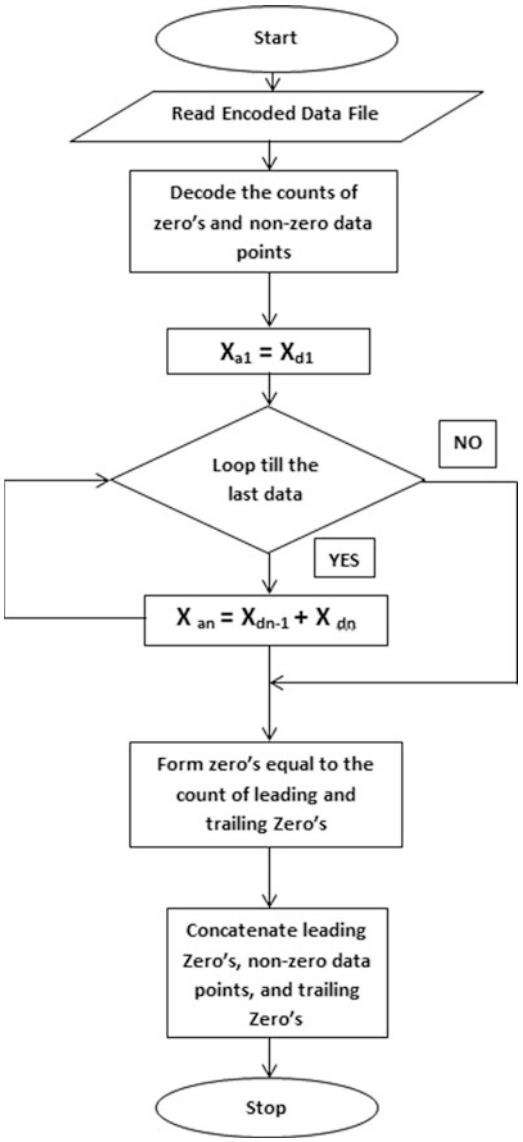


Fig. 7. Flowchart for Radiation data - Decoder

3 Results

3.1 Data Sets

The data sets used in this study consist of a few of the parameters from the Nevada climate change portal like solar radiation, photo synthetically active radiation, data logger power system voltage, incoming longwave, incoming shortwave, outgoing longwave and outgoing shortwave radiation data. The selected site is sheep range black brush. After careful analysis of different data sets we understood that solar radiation data and photo-synthetically active radiation data increase gradually during the day and by afternoon it reaches to maximum and then drops gradually.

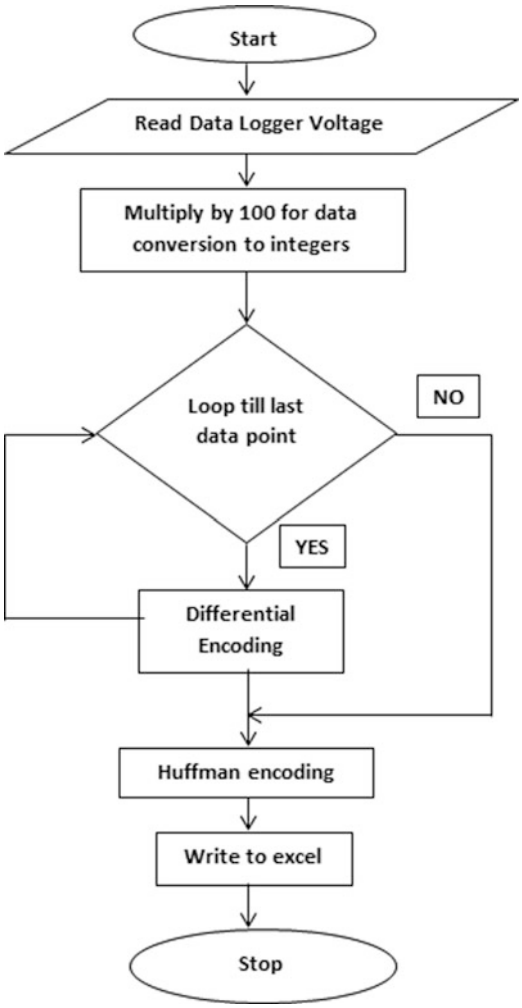


Fig. 8. Flowchart for Encoder – Data logger power system voltage.

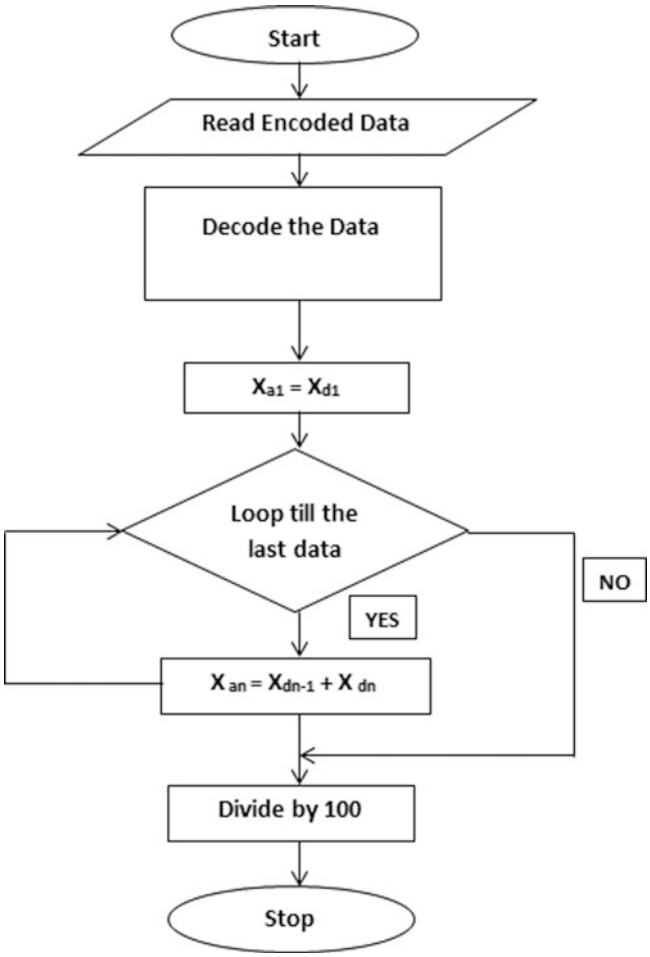


Fig. 9. Flowchart for decoder – Data Logger Power System voltage.

During mid-night and late evening this data is completely zeros. The transitions from one data point to the next is smooth during most of the days. For solar radiation and photo synthetically active data the daily values are considered as a sample whereas for data logger power system voltage monthly values are considered. Data logger power system is in the form of floating point data. We consider converting this to integer form to obtain good compression ratios. Photo synthetically active radiation, data logger power system voltage, incoming longwave, incoming shortwave, outgoing longwave and outgoing shortwave radiation data were used as inputs along with time for the computation of solar radiation for the predictor model using ANN. Using the techniques described above, we achieve compression ratios as high as 5.81 for solar radiation data. The compression ratios achieved were in the range of 2.79 to 5.81. For solar radiation data, the uncompressed bits are 13800.

4 Conclusion

We described a simple method for lossless compression of climate data based on differential and predictive coding. Our scheme provides good compression rates without sacrificing computational efficiently. We achieve good compression rates for solar radiation, photo synthetically active radiation data and data logger power system voltage and also prediction of solar radiation based on artificial neural networks. For future research we will perform compression of other parameters from nevada climate change portal based on predictor model. We will report these results in our forthcoming papers.

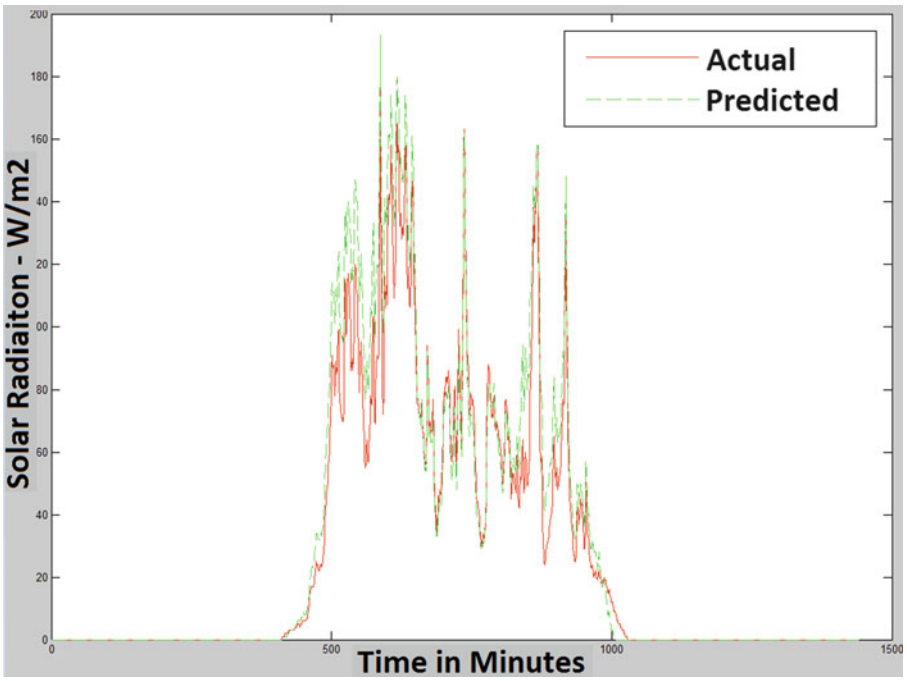


Fig. 10. Plot of actual and predicted data for Jan 23'2012.

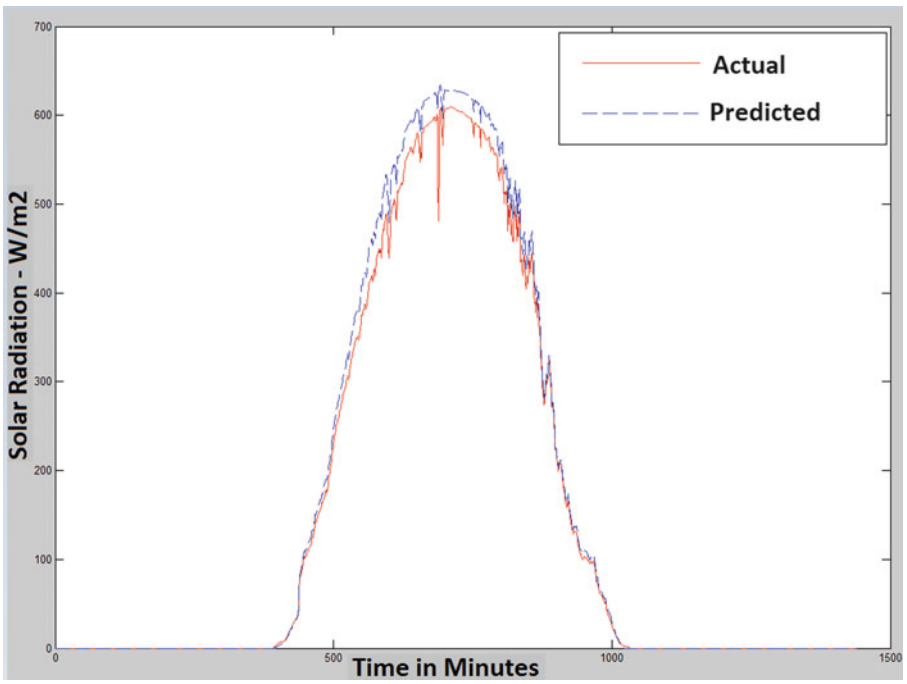


Fig. 11. Plot of actual and predicted data for Jan 24'2012.

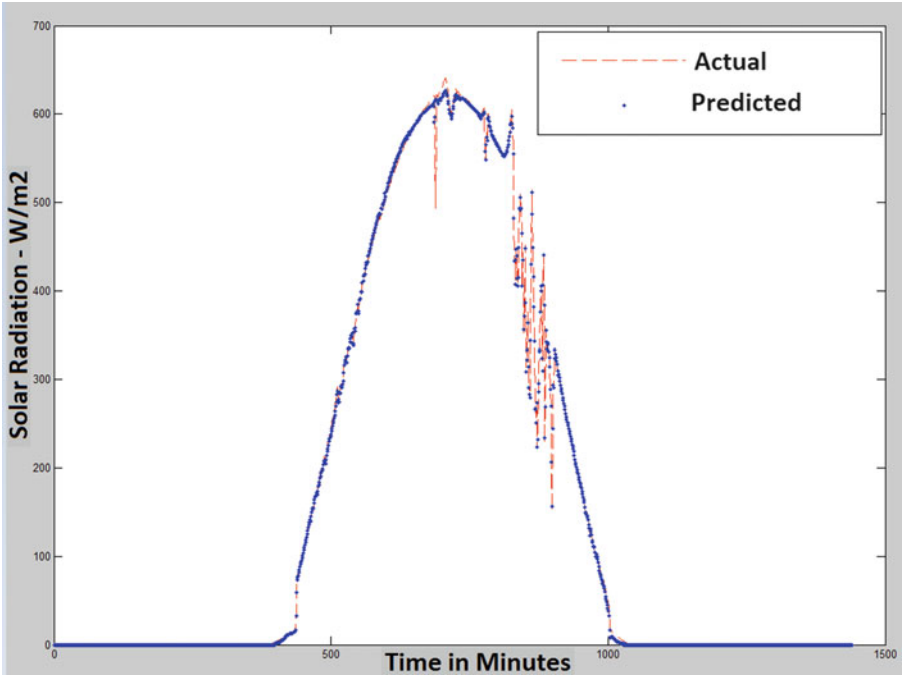


Fig. 12. Plot of actual and predicted data for Jan 25'2012.

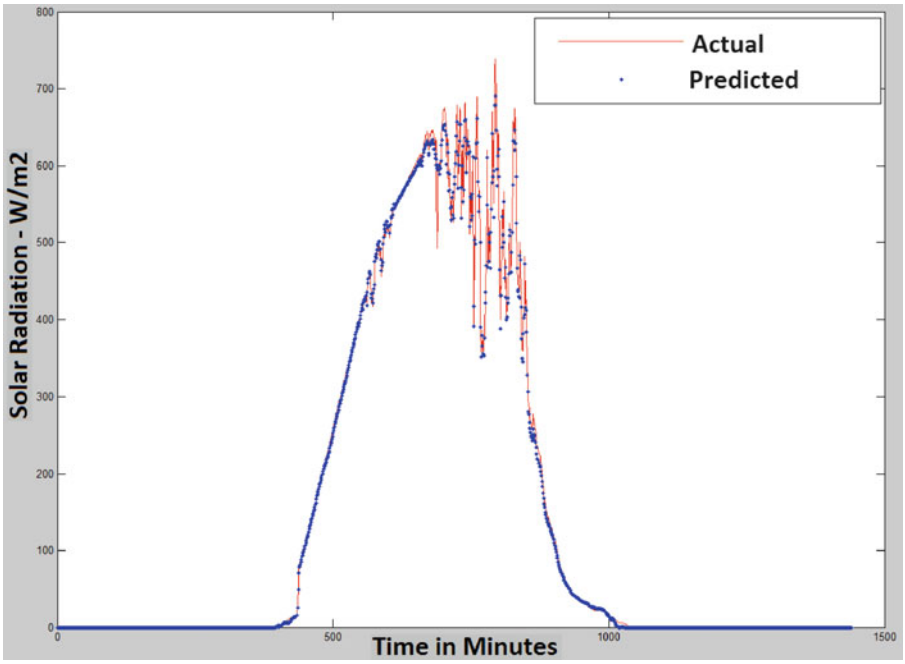


Fig. 13. Plot of actual and predicted data for Jan 26'2012.

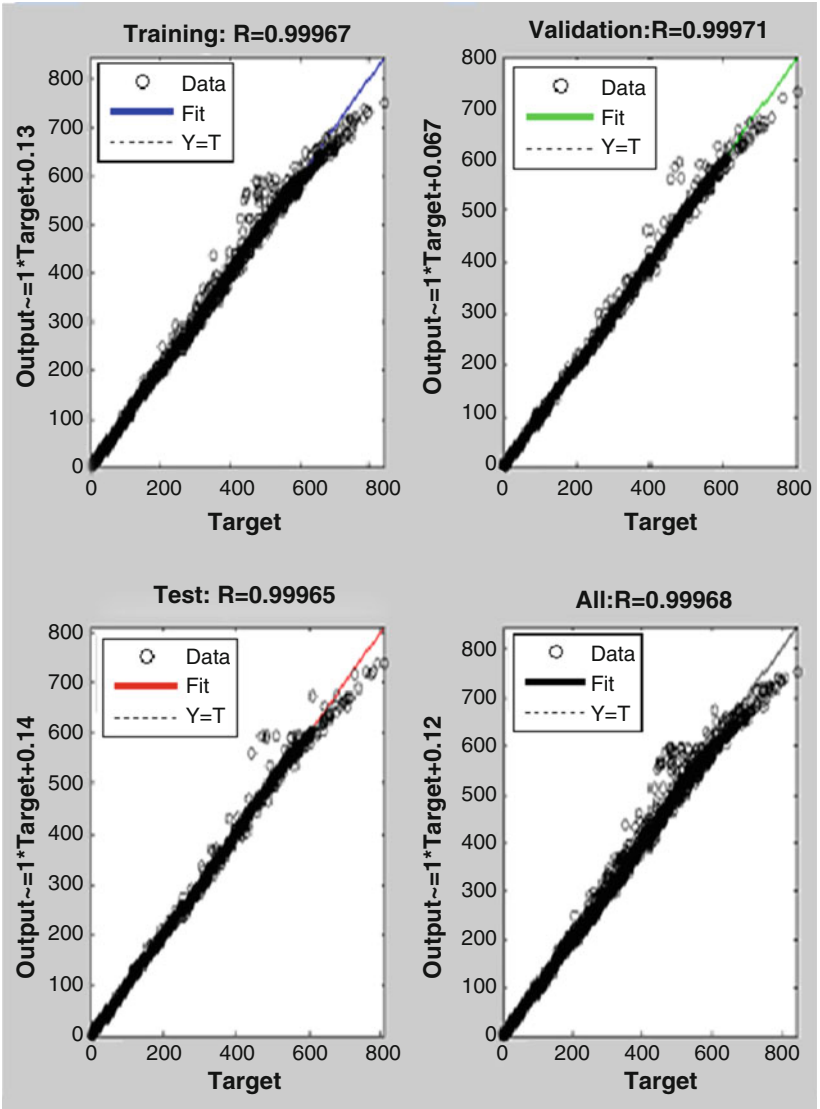


Fig. 14. Regression plot.

Table 1. Solar Radiation

	Day		Day		Day		Day	
Month	7th	CR	14th	CR	21st	CR	28th	CR
Jan	2443	5.65	2530	5.45	4058	3.41	2376	5.81
Feb	3096	4.46	4194	3.30	3434	4.02	4059	3.4
Mar	4415	3.13	4641	2.98	2508	5.51	4902	2.82
Apr	2668	5.18	4936	2.80	3530	3.91	2904	4.76
May	4267	3.24	4196	3.29	3721	3.71	3000	4.6
Jun	3014	4.58	2970	4.65	3219	4.29	3333	4.15
Jul	3269	4.23	5066	2.73	4433	3.12	4271	3.24
Aug	4962	2.79	3842	3.60	4690	2.95	2968	4.65
Sep	3999	3.45	2670	5.16	3350	4.11	4155	3.33
Oct	3233	4.27	2867	4.82	3145	4.39	3114	4.44
Nov	2578	5.35	3775	3.65	3321	4.15	3692	3.73
Dec	3195	4.32	4108	3.36	3612	3.83	2911	4.75

Table 2 Data Logger Power System Voltage

	Actual Data (in bits)	Compressed Data (in bits)	CR
Jan	534960	95894	5.58
Feb	500400	95792	5.23
Mar	534960	99148	5.40
Apr	501132	98782	5.07
May	518412	97072	5.35
Jun	517680	94725	5.47
Jul	534960	94642	5.66
Aug	534960	94690	5.65
Sep	517680	96089	5.39
Oct	534960	100084	5.35
Nov	517680	97984	5.29
Dec	517680	91230	5.68

Table 3. Photo Synthetically Active Radiation Data

	Day 7th			Day 14th		
Month	Compressed data	Actual data	CR	Compressed data	Actual data	CR
Jan	3171	15840	4.99	3069	15840	5.16
Feb	3640	17280	4.74	4692	17280	3.68
Mar	4908	17280	3.52	5168	15840	3.06
Apr	3380	17280	5.11	5485	17280	3.15
May	4861	17280	3.55	4821	17280	3.58
Jun	3713	17280	4.65	3737	17280	4.23
Jul	3982	17280	4.33	5749	15840	3
Aug	5386	17280	3.2	4627	15840	3.42
Sep	5166	17280	3.34	3895	15840	4.07
Oct	4059	15840	3.9	3756	15840	4.21
Nov	3786	15840	4.18	4956	15840	3.19
Dec	4089	15840	3.87	4947	15840	3.21

Table 4 Solar Radiation – Artificial Neural Networks

Date	Actual data (in Bits)	Compressed data (in Bits)	CR
Jan23	14400	4527	3.19
Jan24	14400	4780	3.02
Jan25	14400	3829	3.77
Jan26	14400	4285	3.37

Acknowledgment This work is supported (in part) by the Defense Threat Reduction Agency, Basic Research Award # HDTRA1-12-1-0033, and the National Science Foundation (NSF) award #EPS-IIA-1301726. Any findings, conclusions, or recommendations expressed in the material are those of the author(s) and do not necessarily reflect the views of NSF.

References

1. Sayood K, Introduction to Data Compression, Second Edition.
2. Sharifahmadian E, Choi Y, Latifi S.: "Multichannel Data Compression using Wavelet Subbands Arranging Technique", International Journal of Computer Applications (0975 – 8887) Volume 91 – No.4, April 2014.
3. Saupe D, Hartenstein H, and Wergen W. : "Compression of weather forecast data", institutional information processing systems conference, 2010
4. Steffen C, and Wang N. : "Weather data compression,"NOAA Research-Forecast Systems Laboratory.
5. Karim S,Karim B,Tahir M, Ismail M, Hasan M, and Sulaiman J.: "Compression of temperature data by using daubechies wavelets", International Conference on Mathematical Sciences,2010.
6. Engelson V, Fritzson D, and Fritzson P. : "Lossless Compression of High-volume Numerical", Data compression conference,2000.
7. Xie X, and Qin Q. : "Fast Lossless Compression of seismic Floating Point Data", Information Technology and Applications,2009.
8. Steinbach M, Kumar V. : Introduction to data mining–pang ning tan.
9. Olaiya F. : "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies", I.J. Information Engineering and Electronic Business, 2012, 1, 51-59.
10. Afzali1 M, Afzali A and Zahedi G. :"Ambient Air Temperature Forecasting Using Artificial Neural Network Approach", ICEC Conference, vol.19 2011.

Parameter Trade-off And Performance Analysis of Multi-core Architecture

Surendra Kumar Shukla, CNS Murthy, and P. K. Chande

1 Introduction

Today advancement in hardware and software are going ahead hand in hand. The need of speed, in heterogeneous integrated system, is growing at a very high rate. Which has lead to the use of multiple core of CPUs on single chip. Now mobile and other applications are depending upon these multi-core architectures. But flip-side is that, applications are not capable in taking full advantage of this. This is because parallel computing has many challenges like parallelism, scalability, inter-core communication delay etc. [1, 2]. However in a multi-core system, these are certain aspects, which could be utilized/acted upon the further honers the capabilities of multi-cores. These aspects could be identified/ recognized as parameters to model the system and study their effects towards performance enhancement. There are some parameters which can play a key role for the performance improvement of multi-core architecture. Parameters like parallelism(hardware/software), scalability, need of process communication (API') etc. are some known parameters. Many researchers have provided some techniques for improving the multi-core system performance [3] [4]. But their techniques focus on individual parameter like parallelism in isolation [5]. It may improve the performance of the system in terms of throughput based on a single parameter. In this research work we intent to identify and use multiple parameters in an optimal way. We would also intent to see if the parameters could be adjusted(effecting function on the device) to suit the application nature, and analyze the performance.

S.K. Shukla (✉) • C. Murthy
Department of CSE, CDGI, Indore, M.P., India
e-mail: surendr.shukla@gmail.com; cnsmurthy@gmail.com

P.K. Chande
School of Computer Science & IT DAVV Univercity, Indore, M.P., India
e-mail: pkchandein@yahoo.co.in

2 Literature Review

Bryan schauer [3] has mentioned his work entitled "Multi core Processors- A Necessity", that one important aspect in increasing the performance of the multi-core architecture is the speed up. Speed up can be achieved by increasing the clock speed of the processor. Also, another established fact is by increasing the number of cores, speedup can be further enhanced. However, again, if we increase the number of cores then there are other inevitable problems with memory and cache coherence. And communication between cores also would have to be considered. Researchers have also focused on interconnection networks. But the interconnection network tend to get overloaded if parallelism exist at instruction level. It will create the overhead of communication between instructions, which is a function of inter-dependency.

Raghavan Raman [4] in his PhD thesis has focused - how compiler can support to obtain the parallelism in the program. Hence it is specified that the advent of multi core processors has lead to the emergence of different kinds of programming models to utilise the parallel processing power available. Parallel programming models exist in three categories. These categories are based on their approach to exploit the parallelism; (a) Programming language approach (b) directive based approach and (c) library approach. In (a) execute the task asynchronously, and for scheduling the task, we need to create threads, and, threads are expensive.

Damian A. Mallon [5] in his research paper specified the importance of APIs for the inter process communication. This paper specifies that the current trend of multi-core architectures emphasise the need of parallelism too. This research work evaluated the performance on multi-core between unified parallel c (UPC) and openMP on multi-core architectures. Result showed that MPI is best choice for multi-core systems, with both hybrid shared and distributed memory, because it takes highest advantage of

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_59,

data locality. Here data locality is a key performance factor (parameter).

Dmitri Perelman [6], in his work "Exploiting Parallelism of Multi-Core Architectures", shows the importance of parallelism in multi core performance. He has specified that if threads(program parameter) can be synchronised efficiently, and data exchange between the threads is proper then we can exploit the parallelism. Thus the biggest issue of multi-core architecture in this scenario is synchronisation. He has introduced transitional memory to synchronise the multi-threaded program, but the drawback of this technique is that for the implementation of the scheme are requires to add a small auxiliary translational cache memory.

Kakoullie [7] in 2012 specified that hot spot is the main issue in the performance of multi-core architecture. Hot spot is a router which is responsible for the data exchange among the cores in multi-core processor. Researcher has focused on the issues, which are responsible for the creation of hot-spot in the multi-core systems. He has specified that if hot spot will be generated in inter-connection network, it will affect the performance of the multi-core architecture. Prediction of hot-spot in the network is a difficult task, because generation of hot-spot depends on the nature of the application. For the prevention of hot-spots, the technique described in his paper is based on the artificial neural network(ANN). ANN based technique predicts and avoids the hot spots in the network.

Tudor, B.M. and Young Meng Teo [8] in 2011 specified, that memory contention is a big issue in the performance of the multi-core architecture. They have specified that in a program, with large memory requirement, memory contention increased as 1000 %. Tests have conducted on 24 cores on an inter NUMA system. These tests have been done on SP.C program. Results shown in his paper are- if problem size is small then there is less cache miss, then when less main memory access, then less traffic and also busty traffic. If size of the problem is large, then there is more cache miss, more main memory access, and non-busty traffic. Model proposed in this paper has following limitations- accuracy is decreased for the programs, where less degree of contention is existing. Second limitation is accuracy is decreased for the programs, for low memory requirement. This model is best for the programs, where the memory requirement is higher.

Shahrivari, S & Sharifi, M, [9] specified that Task-Oriented Programming, can help in increasing the performance of the multi-core architecture. They have mentioned that programming languages available for the multi-core & distributed systems are not efficient. To increase the performance of the multi-core architectures, we should have to use new programming models like task-oriented programming. But in task-oriented programming for making the program we need to create threads, and creation of thread involves the system calls, which degrades the performance of the system. Drawback of task oriented programming is that in task oriented programming task and required data both are exists in remote computer and remote computer executes the task and returns it to the needy computer. If data is available far away from the task, it will put higher penalty on the performance of the multi-core architecture.

Bini et al., [10] in their research work specified the concept of virtualization for testing the performance of independent applications. In virtualization concept some part of the whole resource is assigned to a application. Application has the illusion that he is using whole resource. This concept is useful for the applications like mobile phone where resource is precious, so we need to provide the resources to the applications on the basis of their importance. For example there are some applications like display of mobile phone which has high demand of memory, so we must have to give the higher priority to this device when assigning resources like memory and CPU time. Virtualization (parameter) concept can be helpful on identification of those applications which have more requirement of resources.

Khan, et al., [11] in their research work "Improving Multi-Core Performance Using Mixed-Cell Cache Architecture" specified that cache memory can have two type of cells. Robust cell and non-robust cell. Robust cell required more energy because robust cells stores the data which are required write operation in cell. Non-robust cell stores data which are required reading operation in cell. Robust-cell required more energy(parameter) as compared to the non-robust cell.

Ubal, R et al., [12] have proposed a tool for the analysis of multi-core architecture performance. Tool has focused on the analysis of three major performance elements of multi-core architecture-processor cores, memory hierarchy, interconnection network. Tool has visualize graphically how cores executing threads, how many cores are ideal, how interconnection network is utilised by cores for inter-core communication. With the help of this tool we can identify the multi-core architecture performance parameters.

Durate, F, and Wong, S [13] specified the Accelerator scheme, with the help of this scheme data movement between main memory and cache memory can be increased, it can increase the performance of multi-core architecture. Limitation of this scheme is that, it can improve the performance of multi-core architecture only in the case when we are doing copy operation between the main memory and the cache memory. But this scheme is not applied when real time updation is done in main memory and the cache memory.

Magnus Broberg [14] has developed a tool, with the help of the tool a parallel program can be developed in sequential

machine. Tool provides the parallel processing environment in uni-processor. A programmer when writes their program in that virtual parallel environment, he feels that he is actually working in parallel processor. Tool has provided opportunity to make the program for multi-core architecture.

Julian Bui et al., [15] has specified in their research paper, that cache size, cache protocols, associate number etc. are all important parameters for performance, among which cache size is thought to be the most important parameter for performance. Cache size has the largest impact on cache performance and energy consumption.

The following parameters [16] throughput, execution time, energy, memory bandwidth, scalability, inter-core communication, CPU clock speed, number of cores and cache coherence are the important parameters for performance improvement of multi-core architecture.

N. Ramasubramanian et al., [17] has specified that cache memory plays a crucial role in deciding the performance of multi-core system. He has used M5sim tool for the cache memory parameters simulation with multi-core processor parameters like L1 and L2 cache size verses CPU frequency.

Ram prashad mohanty et al., [18] has specified that the evaluation of performance is dependent on internal network e.g. ring network and hybrid network. He has used the parameters, execution time and speed-up, to show the performance of ring network and hybrid network.

The above review had been summarised in the following: Some common identified parameters are listed below

Some common identified parameters are listed below [15–18].

In the literature review we have found following conclusions

[1]. All attempts address some isolated issues and some common issues
[2]. Little about programs and programming style.
[3]. No investigation for hybrid and integrated applications-like mobile, Internet and security.
[4]. No research on instruction types, other characteristics of process and their performance.
[5]. No research on issues of fault tolerance and real-time systems.
[6]. Nothing on load balancing.
[7]. Less research on inter-connection network.
[8]. Less on special support processors.
[9]. No special hardware.

With the above literature survey, we have come with following research objectives, given in below.

3 Scope & Objectives

We Visualise the multi-core processor architecture with above perspectives then

I. Identifying more effective and comprehensive hardware software system parameters.
II. Understanding correlation between parameters.
III. Exploring the scope of balancing the parameters and trade-off optimising the throughput.
IV. Exploring the possibility of on-line parameter optimisation through signalling mechanism.
V. Simulation and analysis which can visualise all parameter tuning operation.
VI. Obtaining the interface standard(protocol) between Intelligent parameter tuning device(IPTD) and multi-core architecture for the interrupts.
VII. Obtaining the performance issues when many parameters and many applications will affect/influence to each other, dependencies.

4 Description of Research Work

Based on the objectives, a research plan has been prepared. Performance of multi-core architecture in totality can be visualised as:

Performance of the multi-core architecture = relation(System parameters + program parameters + data parameters + environment parameters + instruction pattern)

Parameter: A parameter, in its common meaning, is a characteristic, feature, or a measurable factor that can help in defining effect on a particular system. measurable factor which helps to define a particular system.

(1) **System Parameters:** Parameters which exist in the components of the multi-core architecture, e.g. cache memory, main memory, cores in the processor, interconnection network.
(2) **Program Parameters:** parameters which exist in the program by its virtue. e.g. Execution time, parallelism.
(3) **Data Parameters:** parameters related to the principal of locality.
(4) **Environmental Parameters:** parameters which the surroundings of a physical system that may interact with the system. e.g. Energy, temperature, radiations.
(5) **Instruction Pattern:** it means for the execution of instruction, how much data is required. With the help of

instruction pattern information, we can mix the instructions, CPU bound instructions, I/O bound instructions. We can perform data fetching operation for the instructions which has required more data.

4.1 Proposed Scope of Performance Improvement

Temperature balancing in Multi-core Architecture scope for performance improvement: we are proposing a novel approach here for the performance improvement of multi-core architecture, and extra core for monitoring the temperature on cores of multi-core processor. If in any core it was found that the temperature in the core is high because large number of instructions are executing in the core. Load should be automatically balanced, with the core which has less temperature. So some instructions should be transferred to the core which has less temperature. In the figure 4.1 given below it is shown that, because temperature in core 1 is high so some instructions must have to be transferred to the core 4. For monitoring the temperature a special device (extra core) should be used which is responsible to watch continuously the temperature of all the cores existing in the multi-core processor.

Hybrid cores and applications scope for performance improvement/ balancing. We are proposing here RISC + CISC instruction set architecture both should be existing in the core. CISC and RISC both should be used in the dynamic form. Core must have capability to change their mode from CISC instruction set architecture to RISC instruction set architecture as per the characteristics of the application. If application contains instructions which if could be executed in RISC architecture instead of CISC, then core should have to change their mode from CISC to RISC. In this approach we are suggesting that core functionality can be adjusted depending upon load characteristic. If we will implement this technique, performance of the multi-core architecture will be improved. In the diagram 4.2 it is shown that integrated application can be scheduled from CISC mode to RISC mode and from RISC mode to CISC mode.

Common shareable hardware and impact on performance. We are proposing shareable hardware which stores mixed instructions. we are using slack time, (a cycle that can be used for some other operation by the CPU, as instruction is not required that cycle for their execution may be that instruction is doing I/O activity that time) can be used to bring data in advanced for the other instruction stored in the instruction cache. The multi-core processor can utilize this concept for improving the performance. Here slack-time may be a parameter.

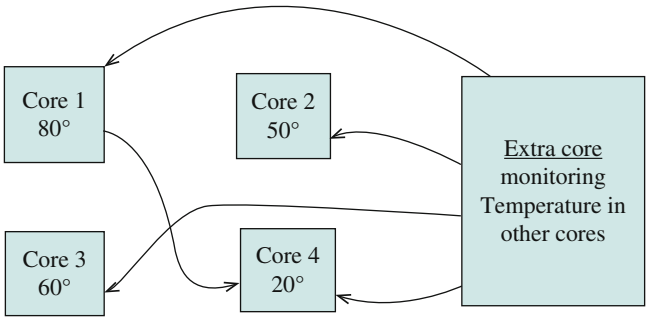


Fig. 1 Proposed approach for temperature control in Multi-core architecture

Intelligent Parameter Tuning device for performance improvement of Multi-core architecture. As our objective of research is to improve the performance of the multi-core architecture, for that we have proposed an intelligent device, IPTD(Intelligent Parameter tuning device). IPTD will act as supporting interface for the multi-core architecture for the performance improvement. As depicted in figure 4.1 any application will be given as a input to the multi-core architecture, machine will execute the application and simultaneously IPTD will watch the whole process of execution. If IPTD finds that any parameter is trying to degrade the performance, it will tune it with the help of the parametric interrupt to the multi-core architecture.We can test the performance of one application, two applications or n applications. We can analyse the performance of multi-core architecture one application to one parameter, one application to many parameter, many application to many parameters and many applications to one parameter. Intelligent parameter Tuning device is a firmware, we need to embedded the parameters in the chip. IPTD device has also a capability to generate the interrupt to the multi-core machine on the basis of parameters. Here parameters are

Functioning of IPTD and multi-core architecture. As shown in the figure 4.3 an application is used as a input to the multi architecture machine. Multi-core architecture machine executes the application. IPTD starts keen observation of the execution of application, with all their parameters and stores the throughput of the machine. As IPTD founds that performance is degrading, it tries that if he can change the some parameter value and performance will again come in the same label. Below we have given two cases of parameters tuning.

Different cases for the parameters tuning.

CASE 1

If IPTD founds that hot spots have been created in the interconnection network, IPTD will change their parallelism parameter and it will stop other parallel parts of the program to be given to the multi-core machine. And when he see that

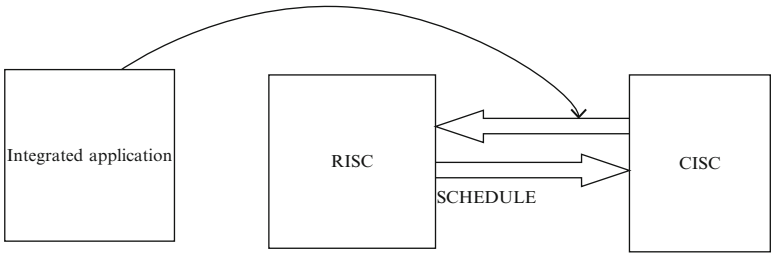


Fig. 2 Proposed Hybrid cores RISC & CISC

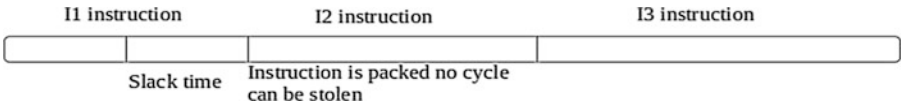


Fig. 3 Proposed Common shareable hardware

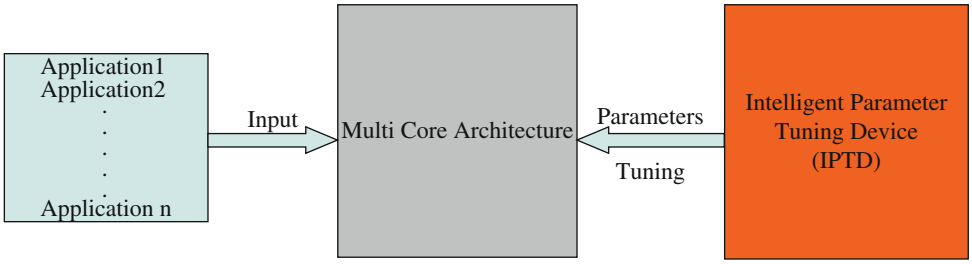


Fig. 4 Proposed Intelligent parameter tuning device

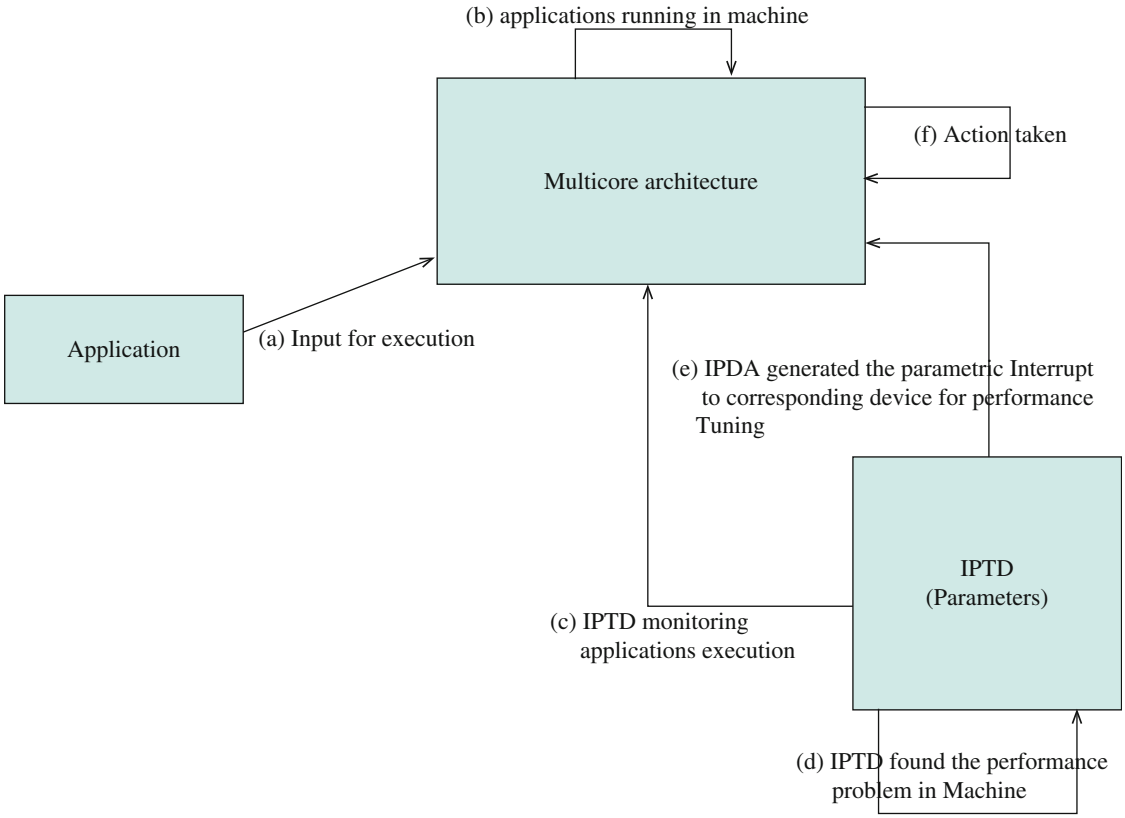


Fig. 5 proposed working of Intelligent parameter tuning device(IPTD).

Table 1 Parameters Description & Correlation.

S. No	Parameter	Parameter description	Correlation with other parameters
1	Speed-up	In parallel computing, speedup refers to how much a parallel algorithm is faster than a corresponding sequential algorithm.	Parallelism
2	Parallelism	Scope of parallel instruction in program.	Speed-up
3	Data locality	Data locality is a, typical memory reference feature of regular programs.	Execution time
4	No of threads	A thread is a, single sequential flow of control within a program.	Inter-connection network
5	Percentage of memory contention	A situation in which, two different programs, or two parts of a program, try to read items in the same block of memory at the same time.	Memory band-width
6	Virtualization	Virtualization, means to create a virtual version of a device or resource, such as a server, storage device, network.	No. of threads
7	Percentage of robust-cell required	Robust-cell, is a cell in cache memory, which is requires more energy.	Energy
8	No of processor cores	Total small CPUs, exist in single chip.	Speed-up
9	Inter-connection Network.	Network used to connect processor cores.	Percentage of memory contention
10	Memory Accelerator	An approach by which, data movement between main memory and cache memory can be increased.	Cache access time
11.	Memory Hierarchy	A ranking of computer memory devices, with devices having the fastest access time at the top of the hierarchy, and devices with slower access times but larger capacity and lower cost at lower levels.	Execution time.

now performance is up to the mark, it will again increase the parallelism parameter.

CASE 2

Suppose that IPTD have observed that, because of memory contention performance is decreasing, he may instruct to the multi-core machine that lot of memory operations are going on so, stop for some time this. In this case again IPTD will have to tune the parallelism parameter.

CASE 3

Suppose that IPTD have observed that the interconnection network have higher traffic because of finer level of granularity, (instructions level parallelism), all the cores are trying to communicate to each other, because of that the performance is degrading. IPTD will tune the compiler and instruct to the compiler for finding the course grain parallelism, so that traffic in the interconnection network can be reduced. After some time when traffic is normal in the interconnection network IPTD will again tune the compiler for the finer grain parallelism (instruction level parallelism).

5 Conclusion

It is clear therefore, that performance of multi-core architecture can be improved with the help of parameters tuning (adjustment). In this research, we will identify the performance parameters, and analyse them for the performance improvement of multi-core architecture. Analysis of performance parametes can provide us various opturnaties on performance improvement in future. If the proposed IPTD will be implemented in future it will be a seprate harware device for performance improvement.

Acknowledgements The authors thanks to management of Chameli Devi Group of Institutions, Indore, M.P. India for providing excellent research enviourment in the college. A Special thanks to the Chairman **Shri Vinod Kumar Agrawal ji**, CDGI Indore, India for motivating to the faculties of the institutioon on research activities.

References

1. http://software.intel.com/en-us/blogs/2008/12/31/top-10-challenges-in-parallel-computing
2. Advanced Computer Architecture, parallelism, scalability, programmability, Kai hwang 2nd ed. McGraw-Hill, 01-Feb-2003
3. Bryan Schauer "Multi core Processors – A Necessity" Pro Quest Discovery Guides, September 2008
4. Raghavan Raman, "Compiler Support for Work-Stealing Parallel Runtime Systems", Ph.D. dissertation, Dept. Computer Science, Rice University, Houston, Texas, May 2009
5. Damian A. Mallon, et al., "Performance Evaluation of MPI, UPC and OpenMP on Multi-core Architectures" in under Project TIN2007-67537-C03-02, Spain.
6. Dmitri Perelman, "Exploiting Parallelism of Multi-Core Architectures", Ph.D dissertation, Department of Electrical Engineering, Israel Institute of technology, Haifa Israel September 2012.

7. Kakoullie, E. et al., "Intelligent Hot-spot Prediction for Network-on-Chip-Based Multi-core Systems" in computer aided Design of Integrated Circuits and Systems, IEEE Transactions on, vol.61 no.3, pp. 418 – 431, 2012.
8. Tudor, B.M. And Young Meng Teo, "Understanding Off-Chip Memory Contention of Parallel Programs in Multi-core Systems", IEEE Conference in Parallel Processing (ICCP), Taipei, Singapore, 2011, pp. 602 - 611
9. Shahrivari, S & Sharifi, M, "Task-Oriented Programming: A Suitable Programming Model for Multi-core and Distributed Systems", IEEE Conference in Parallel and Distributed Computing (ISPDC), Cluj Napoca, Iran, 2011, pp. 139 - 144.
10. Bini, E et al., Resource Management on Multi-core Systems: The ACTORS Approach, published in IEEE Conference in Micro, 2011, pp. 72-81.
11. Khan, et al. "Improving Multi-Core Performance Using Mixed-Cell Cache Architecture", in High Performance Computer Architecture (HPCA2013), IEEE 19th International Symposium, Shenzhen, China, 2013, pp. 119 - 130
12. Ubal, R et al. "Multi2Sim: A Simulation Framework to Evaluate Multi-core-Multithreaded Processors", in Computer Architecture and High Performance Computing,19th International Symposium, Rio Grande do Sul, 2007 pp. 62-68
13. Durate, F and Wong, S "Cache-Based Memory Copy Hardware Accelerator for Multi-core Systems", Published in Computers, IEEE Transactions, (Volume:59, issue:11), 2010, pp. 1494-1507.
14. Magnus Broberg, "Performance prediction and improvement techniques for parallel programs in multi processors", Ph.D dissertation, Department of software Eng. & computer SC., Blekinge institute of technology Sweden, 2002
15. Julian Bui et al.,"Understanding Performance Issues on both Single Core and Multi-core Architecture", ACM conference in Computer Organizatione'07, Charlottesvill, 2007
16. http://www.cse.wustl.edu/~jain/cse567-13/ftp/multicore
17. N. Ramasubramanian et al. "Performance of cache memory subsystems for multi-core architectures", in International Journal of Computer Science Engineering and Applications, (IJCSEA) on vol.1, no.5, October 2011
18. Mohanty et al. "Performance Evaluation of Multi-core Processors with Varied Interconnect Networks", in Advanced Computing Networking and Security (ADCONS), 2nd International Conference, Mangalore, India, 2013, pp. 7-11

Approximation algorithms for utility-maximizing network design problem

Maciej Drwal

1 Introduction

Network utility maximization (NUM) problem is stated as a problem of maximizing concave objective function over a set of linear constraints [9]. The objective function aggregates the "utility" of a collection of flows, which is a nondecreasing function of the amount of flow transmitted between a source and sink nodes. Flows share common links along their paths, and total flow allocation in each link is limited by a fixed capacity.

The formulation of NUM problem is as follows. Let F be the set of flows (or commodities), $|F| = m$. For each flow $k \in F$ let $x_k \geq 0$ represent the transmission rate assigned to it. The network consists of L links, each of capacity $c_l > 0$. Each flow passes through a subset of links, which is expressed by a binary $m \times L$ matrix $\mathbf{A} = [A_{kl}]$, defined as: $a_{kl} = 1$ iff. kth flow passes through lth link. For each flow we have utility function $u_k : \mathbb{R} \to \mathbb{R}$, which assigns utility (payoff) to rate x_k of a flow. The goal is to maximize the sum of all utilities:

$$\max_{\mathbf{x}} \{ \sum_{i \in F} u_i(x_i) : \mathbf{Ax} \leq \mathbf{c}, \mathbf{x} \geq 0 \}. \tag{1}$$

An instance of NUM consists of a set of flows F, set of functions u_i for each $i \in F$, matrix $\mathbf{A}$ and vector $\mathbf{c}$.

While NUM problem, formulated in this way, can be used to model performance of transmission control mechanisms in computer networks, there are also several related issues that can be captured within this framework. In this paper we focus on the *network design* problem from the perspective of utility maximization. The basic model (1) assumes fixed flow assignment, expressed in terms of matrix $\mathbf{A}$. This is reasonable, since in typical IP networks routing problem is separated from transmission rate control problem. One may ask however, whether it is possible to solve these problems jointly (in an efficient way)? Matrix $\mathbf{A}$ can be seen as a result of transformation of input set consisting of source-destination pairs into a flow-to-link assignment. We consider the problem of selecting a matrix that represents such paths for each flow, that the corresponding NUM problem has the greatest optimal solution over all possible sets of paths.

1.1 Problem formulation

We state a mixed-integer programming formulation of the considered problem as follows. Consider a complete graph G on n vertices, $V = \{1, \ldots, n\}$, with capacity function $c : V \times V \to \mathbb{R}_+ \cup \{0\}$. Let $I = \{(s_1, t_1), (s_2, t_2), \ldots, (s_m, t_m)\}$ be a subset of $V \times V$. Given is a set of functions $u_k : \mathbb{R} \to \mathbb{R}$, $k \in F = \{1, \ldots, m\}$.

Let $x_k \geq 0$ be a real variable denoting the rate of kth flow, let y_{ijk} be binary variable, assuming the value 1 if and only if kth flow uses edge (i, j) on its path from source s_k to destination t_k, and value 0 otherwise. We wish to maximize the following objective function:

$$U(\mathbf{x}, \mathbf{y}) = \sum_{k \in F} u_k(x_k) \tag{2}$$

subject to constraints:

$$\forall_{k \in F} \quad \forall_{i \in V} \quad \sum_{j=1}^{n} y_{ijk} - \sum_{j=1}^{n} y_{jik} = \begin{cases} 1, & \text{if } i = s_k, \\ -1, & \text{if } i = t_k, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

$$\forall_{(i,j) \in V \times V} \quad \sum_{k \in F} x_k y_{ijk} \leq c(i,j), \tag{4}$$

$$\forall_{k \in F} \quad x_k \geq 0. \tag{5}$$

M. Drwal (✉)
Institute of Computer Science, Wrocław University of Technology, Wrocław, Poland
e-mail: maciej.drwal@pwr.wroc.pl

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_60,

$$\forall_{i,j,k \in V \times V \times F} \quad y_{ijk} \in \{0,1\}, \tag{6}$$

We call this problem Utility-Maximizing Network Design problem (UMND). The problem can be seen as maximizing over all routing matrices $\mathbf{A}$ in (1), constrained to represent m simple paths between given set of source-destination node pairs I, in addition to selecting the best transmission rates $\mathbf{x}$. The underlying connection structure can be, in general, an arbitrary graph, since capacity function c may assign values of zero to some edges.

2 Related works

Problems of maximizing the aggregate utility of flows in network are called *network utility maximization* (NUM) problems [2], [3], [6], [9]. NUM is related to the well-known *maximum multicommodity flow* problems [10]. In the latter we are given a set of k source-destination pairs (each representing a single commodity) with a demand D_k, and the goal is to find flow assignment function for each pair, such that link capacities are not exceeded and the flow conservation law is preserved on each node, and the common fraction of all routed commodity demands is maximized. In NUM demands are not fixed, but expressed using utility functions.

Combinatorial problems in which we are concerned with selecting subgraphs from a family of graphs are usually called *network design* problems. One commonly encountered problem in this area is called *buy-at-bulk network design* [1]. In this problem we are given a set of source-destination pairs in an undirected graph with given edge lengths, and we need to connect these pairs of nodes by purchasing the capacity c of any edge at some cost $f(c)$ per unit of length. The goal is to purchase sufficient amount of capacity for each edge so that it is possible to route total demand d_i between ith source-destination pair, while minimizing the cost paid. This problem is similar to UMND in the aspect that the cost is usually modeled using concave function f for purchasing capacity, which reflects the fact that increasing allocated resource for larger values yields reduction in cost.

In [7] a network design problem is formulated in which we look for set of shortest paths between all vertex pairs with total weights no greater than given threshold. It is established that such problem is NP-complete. In [12] a similar problem of joint routing and rate control to the one considered in this paper was formulated, and for general utility functions it was shown to be NP-complete. Here we consider only the class of iso-elastic utility functions, and show NP-completeness for such restricted class of problems. Moreover, we indicate that for such class of utility functions the randomized rounding [4], [11], technique may lead to a constant-factor approximation algorithm.

3 Main Results

It is easy to see that for an optimal solution we should assign to each flow the *widest simple path* between source and destination (this is such a path on which the minimal capacity of edges is maximized, and there are no repeated nodes). In case $m = 1$ this is easy to find just by computing maximum spanning tree in given graph, and restricting it to the path between s_1 and t_1. However, for $m > 1$ the paths resulting in highest bandwidth (and highest utility) for different flows may not be disjoint, which would violate capacity constraints.

For general utility functions u_i the problem appears to be difficult to solve. However, in practical applications, such as in computer networking, we are usually interested in utility functions which allow for *fair* allocations [8]. Henceforth, we restrict the choice of utility function to the family of *iso-elastic* functions:

$$u_i(x_i) = \begin{cases} w_i \dfrac{1}{1-\gamma} x_i^{1-\gamma} & \gamma > 0, \gamma \neq 1, \\ w_i \log x_i & \gamma = 1, \end{cases} \tag{7}$$

where $\gamma \in (0, 1]$ is a parameter. Such form of utility has the property that each flow $i \in F$ receives a share of capacity proportional to its weight w_i. If a link of capacity c is the bottleneck link for a collection of flows F, then maximal value of $\sum_{i \in F} u_i(x_i)$ for any $i \in F$ is:

$$x_i^* = c \frac{w_i^{1/\gamma}}{\sum_{k \in F} w_k^{1/\gamma}}.$$

If in addition to the network size n, also the number of flows m is given as a part of the input, then the utility-maximizing network design problem is very likely to be intractable.

Theorem 1. *Problem UMND is NP-complete.*

Proof. We reduce the BIN-PACKING problem [5] to UMND with iso-elastic utility functions with parameter $\gamma = 1$. In the decision version of BIN-PACKING problem we are given a set of items $U = \{a_1, a_2, \ldots, a_m\}$, each with integer size $w(a_k) = w_k > 0$, and two integers $B, K > 0$, and we ask whether it is possible to pack all the items into K bins of size B each.

Given an instance of BIN-PACKING, we construct an instance of UMND as follows. Each $a_i \in U$ corresponds to

one source node s_i. Let $C = \sum_{i=1}^{m} w_i$. There is a central layer of K edges of capacity B, such that each of m source nodes is connected with each of these K edges via an edge of capacity w_i, (thus there are K outgoing edges from each source node s_i). Each of K edges from the central layer is connected to a single edge of capacity C. The other endpoint of this edge is the terminal node for all m flows. This is illustrated in Figure 1.

Each route from a source node to the terminal node must pass through exactly one of K edges of capacity B in the central layer and through the single terminal edge of capacity C. If the latter edge were a bottleneck, its capacity must have been completely filled, and since the utility functions are iso-elastic, each flow would be assigned a fraction w_i of its capacity. Consequently, the optimal solution of UMND problem would have the value:

$$v^* = \sum_{i=1}^{m} w_i \log w_i$$

We claim that such flow assignment is achieved if and only if the given instance of BIN-PACKING problem has a solution. To see this, suppose that all items U can be packed into K bins of size B. Then each flow of rate $x_i = w_i$ would pass through one of the central layer edges, and then through the terminal edge of capacity C, contributing to filling it completely. On the other hand, if not all items from U fit into K bins, the only way to route all flows through central layer would be to reduce the rate of at least one flow below the corresponding value w_i. But that would result in a solution of UMND strictly less than v^*.

Consequently, UMND is NP-hard. Moreover, its decision version obviously belongs to NP, since given a routing matrix and transmission rates, we can easily verify feasibility and evaluate the value of objective function in polynomial time, using formulation (2)–(6). Then it remains to compare the value of objective with a given threshold value. □

3.1 Linear-factor approximation algorithm

Consequently, we are interested in designing polynomial time approximation algorithms for UMND. Perhaps one of the simplest heuristics is based on finding the widest path for each flow separately, followed by squeezing the flows appropriately in order to fit into all links whose capacities were violated. This is summarized as Algorithm 1. As shown in Theorem 2 such algorithm never returns a solution that is less than a factor $O(\frac{1}{m})$ of optimal value.

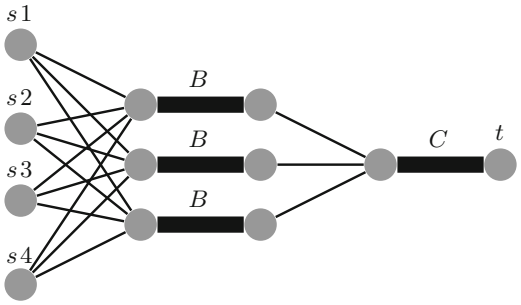


Fig. 1. Example network resulting from reducing instance of BIN-PACKING for $m = 4$ items and $K = 3$ bins of size B.

Theorem 2. *Algorithm 1 is $O(\frac{1}{m})$-approximate for UMND.*

Proof. Let $(x_i^*)_{i \in F}$ be rates in optimal solution of UMND, and let *OPT* be the value of optimal solution. Let $(x_i)_{i \in F}$ be rates in a solution produced by Algorithm 1, and let v be the value of this solution. Let P_j denote the path computed by Algorithm 1 for realizing flow j (i.e., set of edges that make up this path).

For each flow $i \in F$ in solution given by Algorithm 1, let $e(i)$ denote the bottleneck edge of i. Observe, that if $e(i)$ is a bottleneck edge for some flow i in the solution produced by Algorithm 1, it must be that $x_i^* \leq c(e(i))$, as the Algorithm 1 in its first phase finds the widest edge for flow i, unconstrained by the presence of other flows, so no higher allocation for i can be in optimal solution. But since $e(i)$ is a bottleneck edge, some subset of flows contribute to filling its capacity: $\sum_{j:e(i)\in P_j} x_j = c(e(i))$ (this sum always includes flow i). In the special case this sum may include all flows, if $e(i)$ happens to be the bottleneck for all flows. Thus we have:

$$\sum_{j\in F} x_j \geq \sum_{j:e(i)\in P_j} x_j = c(e(i)) \geq x_i^*,$$

which, after summing over all $i \in F$, gives:

$$\sum_{i\in F}\sum_{j\in F} x_j = m \sum_{j\in F} x_j \geq \sum_{i\in F} x_i^*. \tag{8}$$

Let $M = \max_{(i,\, j) \in V \times V} c(i, j)$. Since each $u_j(x_j)$ for $x_j \in [0, M]$, can never assume value greater than $w_j \frac{1}{1-\gamma} M^{1-\gamma}$, we can write:

$$\sum_{j\in F} u_j(x_j) \geq \alpha \sum_{j\in F} x_j,$$

where $\alpha = \min_{i\in F} \{w_i \frac{1}{1-\gamma} M^{-\gamma}\}$, and, consequently from (8):

$$\tilde{v} = \sum_{j\in F} u_j(x_j) \geq \frac{\alpha}{m} \sum_{j\in F} x_j^*.$$

Algorithm 1

Require: Graph $G = (V, c)$ with $|V| = n$ vertices and edge capacity function c. Set of s-d pairs $I \subset V \times V$ of size $|I| = m$.
Ensure: Set of paths $P_k \subset V \times V$ and rates of flows $x_k \geq 0$, for $k = 1, \ldots, m$.

Find maximum spanning tree in G. Denote the set of its edges by T.
for $k = 1, \ldots, m$ **do**
  Let $P_k \subset T$ be a path between $(s_k, t_k) \in I$ in the maximum spanning tree.
  Let $x'_k = \min_{(i,j)\in P_k} c(i, j)$.
end for
for $(i, j) \in V \times V$ **do**
  Let S be the set of flows passing through edge (i, j), i.e., $S = \{k : (i, j) \in P_k\}$.
  if $\sum_{k\in S} x'_k > c(i, j)$ **then**
    For all $k \in S$ let $x_k = c(i, j)\frac{x'_k}{\sum_{l\in S} x'_l}$.
  end if
end for

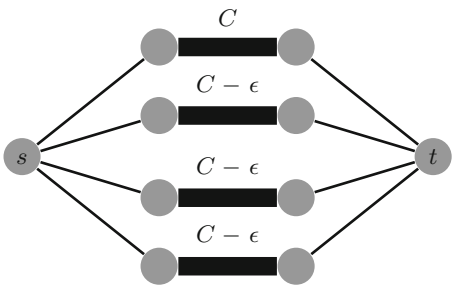


Fig. 2. Example network illustrating the worst case instance for Algorithm 1.

Since for all $j \in F$, $\alpha x_j^* + \alpha \geq u_j(x_j^*)$, we obtain:

$$\tilde{v} \geq \frac{\alpha}{m}\sum_{j\in F} u_j(x_j^*) - \alpha = \alpha\frac{1}{m}OPT - \alpha = O(\frac{1}{m})OPT. \qquad \square$$

To see that the bound of $\frac{1}{m}$ is strict, consider the network of m parallel links, of which the first one has capacity C, and the remaining $m - 1$ of them have capacities $C - \varepsilon$ each, for some small $\varepsilon > 0$ (see Figure 2). There are m identically weighted flows originating from common node s and terminating at common node t (other edges have sufficiently high capacities to never be bottlenecks). Algorithm 1 would assign all flows to the first link, as it greedily looks for the widest paths for each of them separately. The value of solution would be $\sum_{i\in F} u_i(C/m) = m^{\gamma-1}\sum_{i\in F} u_i(C)$. Optimal solution would consist of each flow assigned to a separate link, thus giving solution $\sum_{i\in F} u_i(C) - \varepsilon'$, where $\varepsilon' > 0$ can be arbitrarily small.

3.2 Constant-factor randomized approximation algorithm

It is possible to obtain a better algorithm by allowing randomization. The idea is to use the the mixed-integer program capturing UMND problem, and to relax integrality constraints. This allows to obtain an upper bound on the optimal solution in polynomial time, by solving the relaxation. Although we allow only for a single path for each flow, after the relaxation, the fractional solution might split a flow among several paths. However, such solution can be treated as a probability distribution on paths for a given flow. Subsequently, we can apply randomized rounding of such solution.

Since constraints (4) are nonlinear, we first need to reformulate it to an equivalent linear program. Let us introduce variables $\mathbf{z} = [z_{ijk}]$, and substitute (4) by the following set of constraints, to make sure that $z_{ijk} = x_r y_{ijk}$:

$$\forall_{i,j,k\in V\times V\times F} \quad z_{ijk} \leq C y_{ijk}, \tag{9}$$

$$\forall_{i,j,k\in V\times V\times F} \quad z_{ijk} \leq x_k, \tag{10}$$

$$\forall_{i,j,k\in V\times V\times F} \quad z_{ijk} \geq x_k - C(1 - y_{ijk}), \tag{11}$$

$$\forall_{i,j,k\in V\times V\times F} \quad z_{ijk} \geq 0. \tag{12}$$

Here $C = \max_{i,j} c(i, j)$. A path chosen for a flow is represented by a set of selected edges, that is a set of indices of variables, such that $y_{s_k u_1} = y_{u_1 u_2} = \ldots = y_{u_q t_k} = 1$. Observe, that constraints (3) guarantee that for each flow k there would be exactly one path selected.

Algorithm 2

Require: Instance of UMND problem in form of mixed-integer program (2)–(3), (5)–(12).
Ensure: Set of paths $\mathbf{y}$ and rates of flows $\mathbf{x}$.

1: Solve the LP relaxation of (2)–(3), (5)–(12), denote fractional solution by $(\tilde{x}_k, \tilde{y}_{ijk})$.
2: **for** $k = 1, \ldots, m$ **do**
3: $u \leftarrow s_k$
4: **repeat**
5: Select randomly $v \in V$, with probability of selecting v being:

$$Pr[v \text{ is selected}] = \frac{\tilde{y}_{uvk}}{\sum_{w:\in V} \tilde{y}_{wuk}}.$$

6: $y_{uvk} \leftarrow 1$ (and $y_{uwk} = 0$ for all other $w \in V$)
7: $u \leftarrow v$
8: **until** $v = t_k$
9: $x_k \leftarrow \frac{1}{e}\tilde{x}_k$
10: **end for**
11: Return solution $(\mathbf{x}, \mathbf{y}) = ([x_k], [y_{ijk}])$.

By relaxing constraints (6) to $y_{ijk} \geq 0$ we obtain a linear program. Let $(\tilde{x}_k, \tilde{y}_{ijk})$, for $i, j, k \in V \times V \times F$, denote an optimal fractional solution. Due to the constraint (3), for each flow k the sum of values of edge selection decision variables $\tilde{y}_{ijk}$ must be equal to 1. Since $\tilde{y}_{ijk}$ may now assume fractional values in the range [0, 1] that sum up to 1, we may use these values as a probability distribution of selecting a path among a subset of paths between s_k and t_k, that fractional solution would contain.

However, there is a concern that with nonzero probability the rounded solution may be infeasible, due to the violation of constraint (4). To overcome this, we may shrink the values of rates obtained from solving linear relaxation by multiplying them by a common factor. This would allow us to bound the probability of constraint violation. This idea leads to the Algorithm 2.

Proposition 1. *For any $\gamma \neq 1$, Algorithm 2 returns a feasible solution for UMND with nonzero probability. Such a solution is no worse than $e^{\gamma-1}$ times the optimal solution.*

Proof. The Algorithm 2 selects for each flow a path from the source s_k to destination t_k, by randomly forking on each edge, with probability given by fractional solution $\tilde{y}_{uvk}$. For each flow k, a particular vertex v is added to the currently constructed path with conditional probability:

$$Pr[edge(u, v) \text{ is selected} | \text{current path include vertex } u],$$

computed in step 5. In order to show that a solution constructed this way can be feasible, we calculate the probability of violating constraint (4).

Let $Y^k_{(i,j)} \in \{0, \tilde{x}_k\}$ be a random variable with $Pr[Y^k_{(i,j)} = \tilde{x}_k] = \tilde{y}_{ijk}$. Consider a random variable $Y = \sum_{k \in F} Y^k_{(i,j)}$. Its expected value is clearly $\mu = \sum_{k \in F} \tilde{x}_k \tilde{y}_{ijk} \leq c_{ij}$. We apply Chernoff bound to (4) including rounded solution $y_{ijk} \in \{0, 1\}$. For any $(i, j) \in V \times V$ and any $\delta > 0$ it holds that:

$$Pr[\sum_{k \in F} \tilde{x}_k y_{ijk} \geq (1+\delta)c_{ij}] = Pr[\sum_{k \in F} x_k y_{ijk} \geq c_{ij}] \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{c_{ij}},$$

where $x_r = \frac{1}{1+\delta}\tilde{x}_k$ is rate allocated by Algorithm 2.

Let us denote $C = \min_{(i,j) \in V \times V} c_{ij}$. Then:

$$Pr[\sum_{k \in F} x_k y_{ijk} \geq c_{ij}] < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{C} = \eta. \quad (13)$$

The rounded solution generated by Algorithm 2 is feasible if all constraints in set (4) are satisfied. From (13), any constraint concerning edge (i, j) is satisfied with probability at

Table 1 Performance evaluation of Algorithms 1 and 2.

n	d	m	A1 sol.	A1 time	A2 sol.	A2 time	OPT	time
20	50	10	128.63	1	137.71	1	144.39	2
20	100	10	122.33	1	167.8	3	175.68	13
50	100	10	151.13	1	153.26	3	153.26	3
50	100	25	334.69	1	379.49	4	384.88	5
70	400	10	121.35	1	162.71	3	174.00	40
70	250	12	142.30	1	178.61	5	179.37	1115
70	400	20	212.49	2	314.92	5	> 380.0	> 2h

least $(1 - \eta)$. Since there can be up to n^2 constraints, the probability that all of them are satisfied simultaneously cannot be less than $\varepsilon = (1-\eta)^{n^2}$. We show that $\varepsilon > 0$. By taking logarithm of (13) we obtain:

$$\log \frac{e^{\delta}}{(1+\delta)^{(1+\delta)}} = \log \eta^{1/C},$$

and for $\delta \geq e - 1$:

$$\begin{aligned} \log \eta^{1/C} &= \delta - (1+\delta)\log(1+\delta) \leq \log(1+\delta)(\delta - 1 - \delta) \\ &= \log \frac{1}{1+\delta}. \end{aligned}$$

Thus fixing $\delta = e - 1$ (see step 9 of Algorithm 2) we get $\log \eta^{1/C} = \log \frac{1}{e}$, and consequently $\varepsilon \geq (1 - e^{-C})^{n^2}$, which is greater than zero.

Finally, we evaluate the objective function, given solution $(x_k)_{k \in F}$ returned by Algorithm 2, and under the assumption that u_k are of the form (7). Since $\sum_{k \in F} u_k(\tilde{x}_k) \geq \sum_{k \in F} u_k(x_k^*) = OPT$, we have:

$$\sum_{k \in F} u_k(x_k) = \sum_{k \in F} u_k(\frac{1}{e}\tilde{x}_k) = e^{\gamma - 1} \sum_{k \in F} u_k(\tilde{x}_k) \geq e^{\gamma - 1} OPT. \square$$

Table 1 contains results of computational experiments for several randomly generated input data sets, comparing performance of two presented approximation algorithms: values of solutions and corresponding running times (in secs.). Last two columns contain value of optimal solution (OPT) and running time of branch and cut algorithm from CPLEX solver; n is the number of nodes, d is the number of edges, and m is the number of flows.

4 Conclusions

The utility-maximizing network design problem with iso-elastic utility functions has been shown to be NP-complete. Approximate solutions can be obtained in polynomial time deterministically, but presented study suggests that randomization may yield better results. As Algorithm 2 may fail to give a feasible solution, further investigations are needed to determine the required number of runs.

References

1. B. Awerbuch and Y. Azar. Buy-at-bulk network design. In: *38th Annual Symposium on Foundations of Computer Science, 1997*, pages 542–547. IEEE, 1997.
2. M. Chiang, et al. Layering as optimization decomposition: A mathematical theory of network architectures. *Proc. of the IEEE*, 95 (1):255–312, 2007.
3. M. Drwal and D. Gasior. Utility-based rate control and capacity allocation in virtual networks. In *Proc. of the 1st European Teletraffic Seminar*, 2011.
4. M. Drwal and J. Jozefczyk. Decomposition algorithms for data placement problem based on Lagrangian relaxation and randomized rounding. *Annals of Operations Research, DOI: 10.1007/s10479-013-1330-7*, 2013.
5. M. Garey and D. Johnson. *Computers and intractability*, Freeman New York, 1979.
6. D. Gasior and M. Drwal. Pareto-optimal Nash equilibrium in capacity allocation game for self-managed networks. *Computer Networks*, 57(14):2675–2868, 2013.
7. D. Johnson, J.K. Lenstra, and A. Kan. The complexity of the network design problem. *Networks*, 8(4):279–285, 1978.
8. F. Kelly. Fairness and stability of end-to-end congestion control. *European Journal of Control*, 9(2-3):159–176, 2003.
9. F. Kelly, A. Maulloo, and D. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49(3):237–252, 1998.
10. T. Leighton and S. Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM*, 46(6), 1999.
11. P. Raghavan and C. Tompson. Randomized rounding: a technique for provably good algorithms and algorithmic proofs. *Combinatorica*, 7(4):365–374, 1987.
12. J. Wang, L. Li, S. Low, and J. Doyle. Cross-layer optimization in TCP/IP networks. *IEEE/ACM Transactions on Networking*, 13(3): 582–595, 2005.

Network Energy Reduction via an Adaptive Shutdown Algorithm

Mohamed K. Watfa, L'emir Bachir Chehab, and Zayed Sulaiman Balbahaith

1 Introduction

The world is only beginning to grasp the complex consequences of the exploding growth of Internet. The rapid growth of the Internet makes it difficult to understand its current impact, let alone its future impact. With the tremendous growth of products and applications over the past decade, the internet sector is often a topic of discussion concerning energy utilization. The relationship between the internet and energy use would have profound implications for long-term economic and energy forecasting in this country. It would affect key national issues such as how costly it will be for the us to reduce the energy consumption by the internet and therefore reduce its emissions of heat trapping gases that scientists believe contribute to global warming and climate change. The number of internet users continue to increase year by year, which would lead to more energy usage where routers consume at least 9 % of the total electricity consumption. Electricity generation is the leading cause of industrial air pollution in the UAE. Most of our electricity comes from coal, nuclear, and other fossil fuels power plants. Producing energy from these resources takes a severe toll on our environment, polluting our air, land and water. The information and communications technology are responsible of 2 % of the global carbon dioxide emission and energy consumption is not taken into consideration when configuring the network components. Unfortunately, energy usage is not proportional to its usage. Networks devices and equipments stay fully on even if they are not used. Networking devices are estimated to increase by a factor of 20 by the year 2020. Also, as the router throughput increases the power consumption of each router increases exponentially [15]. Previously proposed internet protocols did not take fully consideration of the need to conserve energy and thus cannot achieve high energy efficiency. It costs about 1 kilogram of coal to create, package, store and move 2 megabytes of data which implies that the Internet is burning up an awful lot of fossil fuel in the process of delivering what we are looking for. Getting the bits from a web server to your PC requires a lot of electricity. Cisco's routers, for example, need about 3 kilowatts of power on average and the wireless Web draws even more power, because its signals are broadcast in all directions, rather than being tunneled down a wire or fiber. Redesigning the network infrastructure is required whenever there is an increase in the number of users or whenever new services are added. Given that the existing network designs are causing a huge consumption of power related to the network interfaces being used all the time and the energy consumption of network links is not proportional to the actual end user usage, this paper proposes a new technique in network interface status aggregation.

The rest of this paper is organized as follows. In Section 2, the related research work is summarized. Section 3 presents detailed analysis and comparison of existing algorithms for energy savings followed by our proposed adaptive algorithm. We conclude this paper in Section 4 followed by a list of references.

2 Related Work

The rapid growth of the Internet makes it difficult to understand its current impact, let alone its future impact. In part this is due to Metcalfe's Law of Networks, which states that the usefulness of the network equals the square of the number of users. In other words, while the number of users grows arithmetically, the value and impact of the

M.K. Watfa (✉) • L.B. Chehab • Z.S. Balbahaith
Faculty of Engineering, University of Wollongong in Dubai (UOWD), Knowledge Village, Dubai, UAE
e-mail: MohamedWatfa@Uowdubai.ac.ae; Lebc810@uowmail.edu.au; zsmb271@uowmail.edu.au

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_61,

Internet grows exponentially. There have been several works in the literature that discuss green network strategies. These strategies include the process of using energy efficient networking technologies like virtualization, server consolidation, system management, upgrading & renewal, and substitutions to minimizing resource use whenever possible [1]. The benefits of green networking from the environmental side include the reduction of gas emissions and prevention of global warming. The benefits of green network strategies from an economical side include the reduction of electricity consumption to increase competitiveness, and minimize power wastage. The benefits from the management side include saving time, reducing cost, and easy control & management to maximize profits. Obviously green technologies face challenges on how to economically build the network, how to deal with the techniques in order to find ways to reduce the power for the network while operation [3].

In [2], green networking was analyzed in both wired & wireless networks and its consequent environmental and economic impacts. Redesigning the network architecture itself was suggested to minimize the greenhouse gases emission. It was also suggested to minimize the geographical delocalization of services to reduce the operations related to energy supply. From an engineering direction, it was also suggested to carry out the given task while having the same level of performance. One of the prominent green networking strategies is the energy-aware infrastructure approach which defines the way network devices are installed in the network and how they communicate with each other in a low power consumption mode. The existing design of the network devices interfaces use the full speed of the network link which leads to consume more power even when the network device is being used by a process or it is idle. This will cause the network to drain the enterprise energy resources and will lead to unwanted network traffic which can increase network uplinks utilization and slow down the productivity of the end users accessing network services like ERP, Media server among others.

Building energy-aware infrastructures utilize traditional ways which are still used today including: the incremental Process, and the clean-state process [4][5].

- *Incremental Process*: The process of building over an existing infrastructure to arrange all the processes and packets destined to the same router. This process has a major disadvantage which is increased end to end delay. An example of this process is the periods of activity and sleep inside the network [4].
- *Clean-state process*: The process that completely rebuilds the new architecture using optical network extensively which reflects on saving the consumption of the power. A number of papers study how to combine optical transport & packet processing [5].

In this paper, we will focus on the Adaptive Link Rate (ALR) technique [6–8]. ALR reduces the consumption of energy by reducing the utilization of the network link and it is considered one of the most widely developed and used techniques to date. The results in [6–8] present that the consumption on the links independent of its utilization. Practically, network peripherals like work stations, laptops, printers and servers transmit synchronization traffic all the time even though no one on the network is communicating with them. An example of the synchronization traffic is the traffic sent for network discovery services, nearby devices, broadcast packets and others. So in that case all the network links will stay on all the time communicating in full mode. The network infrastructure interface offers a 1Gbps link or 100Mbps link where network devices will utilize the highest speed possible and therefore lead to exponential increase in power consumption. A number of published work [10] have attempted to solve these issues where it was suggested to shut down the links when there are idle periods which is referred to as the "sleeping mode". Other works [11–13] suggest to reduce the rate of the link based on the utilization needed on the link which is referred to as "rate switch".

2.1 Sleeping mode – Shut down

The normal scenario for the link is a fully working mode but with the sleeping mode, it will turn off the link during idle periods. The authors of [9] [10] allow the nodes to decide on their interfaces status by measuring packet sequential arrival times. If this interval is long enough, they will lower the link speed for energy savings between two consecutive frames. The efficiency of such a strategy is directly tied to the inter-arrival distributions. Different sleeping modes were investigated including:

- Stay in a sleeping mode and drop the packets coming during sleeping mode.
- Only awake when receiving packets.
- Store the packets coming during the sleeping mode then process them when the device is awake.

2.2 Rate Switch Mode

In a normal scenario of network operation, the link operates with a fixed rate limited by the network's core switches and the capabilities of the network interface of the computer peripheral from the start until the end of operation. Increasing the rate of the link would increase the energy consumption. If there are no pending tasks on the link then the rate switch technique would reduce the rate of the link until a load becomes available where the rate would be

consequently increased. Based on the utilization of the link, selecting the proper rate will reflect on the consumption of the power. Ethernet protocols communicate on the network via several transmission rates, from 10 Mb/s, 100 Mb/s, 1Gb/s and 10 Gb/s. The authors of [11] showed that there are minor changes in power consumption on various transmission rates. An increase of the data rate of a Network Interface Card from 10 Mb/s to 1 Gb/s results in an increased energy consumption of about 5 W which represents about 8 % of the overall system energy consumption. The authors of [11] proposed successive improvement of the rate control policies, based only on the current system state [12] or on historical analysis [13].

3 Algorithm and Experimental Analysis

In this section, we will compare and analyze some of the exiting techniques presented in Section 2 and propose a new energy efficient technique. In all our experiments, a network device is connected to the power outlet via a power meter. Several workstations running different processes are connected to the network device and generate predetermined traffic at several time intervals. The device configuration program modifies the various configuration states running the different energy efficient algorithms at the network interface with varying link rates. The collected information is then processed by an analyzer program to generate various energy proportionality indices and other link metrics. In Figure 1(a), when all the devices are always on, the network devices interfaces use the full speed of the network link which would lead to more power consumption in both idle and active states. Figure 1(b) simulates the operation of the sleeping mode and rate switch mode respectively in terms of energy usage and link rates.

As noted in Figure 1, in the sleeping mode, the energy usage alternates between the different idle states for several peak usages while in the rate switch mode, the peak mode is less than that of the sleep mode and energy curve has a less overall slope.

3.1 *Rate Switch* mode vs. *Sleep* mode

To further analyze both modes, the *Sleeping Mode* and *Rate Switch* mode were applied at a network infrastructure level. From the energy saving point of view, the comparison highlights a network utilization threshold below which the sleeping mode performs better than rate switch and vice-versa. Moreover, we compare two types of rate sets. In the first case, the rates are distributed exponentially (e.g., 100 Mb/s, 1000 Mb/s, 10 Gb/s) while in the second case they are distributed uniformly (e.g., 290 Mb/s, 540 Mb/s, 1100 Mb/s). Results show that equally distributed rates perform better than the dynamic distributed ones, in terms of added delay and average rate reduction. Other works such as [9], [14] provide a relative comparison of sleeping mode versus rate switch strategies, when applied to processors and servers, respectively. In this case, both works are in favor of sleeping mode strategies because of a lower management complexity for a comparable performance level. The lower complexity comes with simpler optimization goals, which are minimizing idle energy and transition time.

Figures 2 and 3 show that both techniques will still cause network utilization and power consumption even when the network is idle. The red shaded area is what we will attempt to improve in our proposed *Adaptive Shutdown Rate* mode.

3.2 Adaptive Shutdown Rate (ASR)

We consider a combination of the sleeping mode and adaptive switch rate and present Adaptive Shutdown Rate (ASR). The major characteristics of ASR will be adding the sleeping mechanism to the adaptive switch rate to reduce the consumption of power on network links for communication. ASR is to be implemented on all network interfaces in the network infrastructure to reduce the power consumption of the processing which would lead to the reduction in energy all over the network as described in Figure 4.

Key Features:

- Network link speed will be proportional to the link utilization of process.
- Network link power consumption will be proportional to the utilization of process.
- Network link will go to idle mode while unutilized by any process.
- Network link will offer link speed based on process speed requirements.

The pseudo code of the *Adaptive Shutdown Rate* algorithm is as follows:

```
– Executes while network device is on
    If (Link rate is zero)
      If (frame length equal zero)
      Handshake for lowest link rate
- Executes when receiving a frame
    If (Link rate is greater than zero)
      If (frame length greater than zero)
      Handshake for frame length
- Executes on shifting from higher or
  lower process using higher or lower
  link rate to another process
    If [(Link rate is greater than frame
    length) or (Link rate is smaller
    than frame length)]
      Handshake for new frame length
```

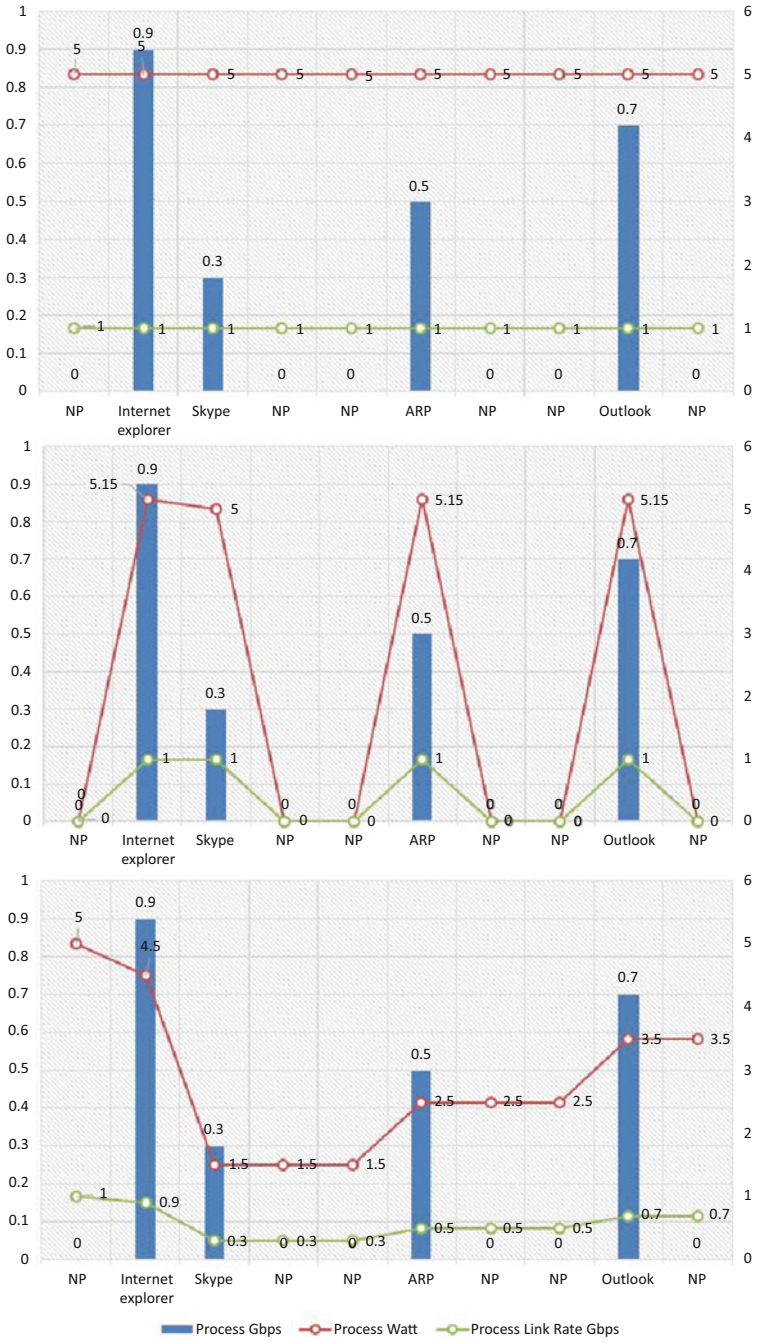


Fig. 1 An analysis of the link rates and energy usage using (a) Always on mode, (b) Sleeping mode, and (c) Rate Switch mode respectively.

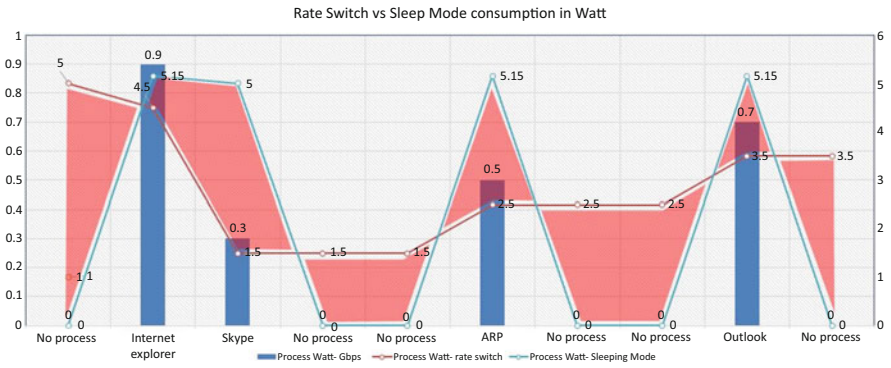


Fig. 2 *Rate Switch* compared to *Sleep Mode* measured in watts

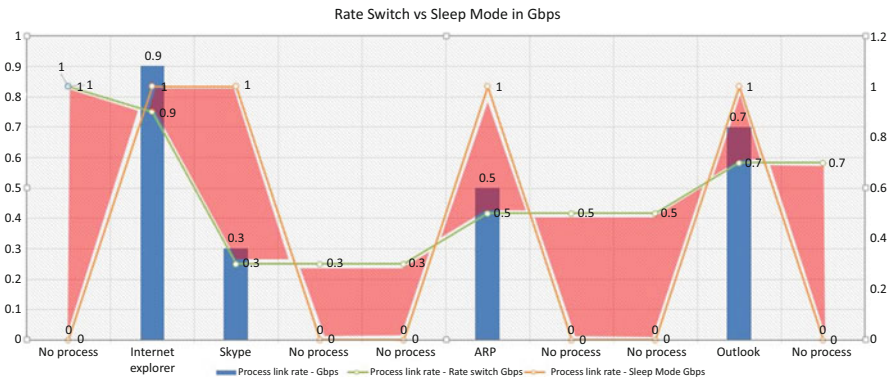


Fig. 3 *Rate Switch* compared to *Sleep Mode* measured in Gbps

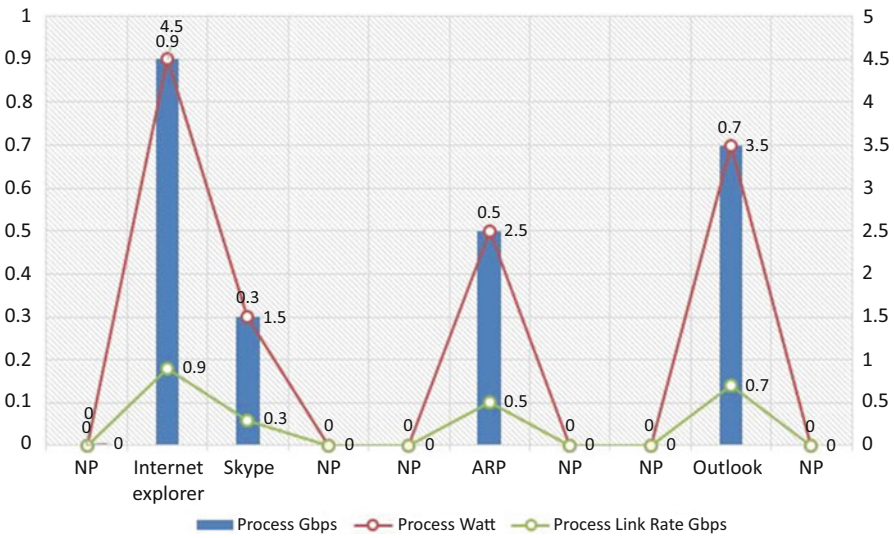


Fig. 4 An analysis of the link rates and energy usage using ASR.

The Adaptive Shutdown Rate will reduce the power consumption from 50 watts / 10 Gbps to 12 watts / 2.4 Gbps for the same running process during the same time frames. In another analysis based on existing technologies implemented, workstations and network components network interface links will stay ON with 1 Gbps links which would consume 5 watts during the life span of the computer or the process. ASR will lead to a lower consumption of power and transmission link speeds based on the actual utilization and requirement of the processes, as described in Table 1. ASR shows an improvement in energy savings of 76 %, 57 % and 41 % over a normal, Rate Switch mode and Sleeping modes respectively.

4 Conclusions

Networking equipments such as modems and routers have a significant power draw. Data shows that networking components in an office setting account for about 1/4 the electricity demand of computers and monitors (that is, for every 100 kWh of demand from desktop PCs, another 25 kWh is needed for networking components). Electricity generation is the leading cause of industrial air pollution. Most of our electricity comes from coal, nuclear, and other fossil fuels power plants. Producing energy from these resources takes a severe toll on our environment, polluting our air, land and

Table 1 *Always On*, *Rate Switch*, *Sleeping Mode* and *ASR Comparison*

		Always on		Rate Switch		Sleeping mode		Adaptive Shutdown rate	
Process	Gbps	Watt	Link Rate Gbps	Watt	Link Rate Gbps	Watt	Link Rate Gbps	Watt	Link Rate Gbps
No process	0	5	1	5	1	0	0	0	0
Internet explorer	0.9	5	1	4.5	0.9	5.15	1	4.5	0.9
Skype	0.3	5	1	1.5	0.3	5	1	1.5	0.3
No process	0	5	1	1.5	0.3	0	0	0	0
No process	0	5	1	1.5	0.3	0	0	0	0
ARP	0.5	5	1	2.5	0.5	5.15	1	2.5	0.5
No process	0	5	1	2.5	0.5	0	0	0	0
No process	0	5	1	2.5	0.5	0	0	0	0
Outlook	0.7	5	1	3.5	0.7	5.15	1	3.5	0.7
No process	0	5	1	3.5	0.7	0	0	0	0
	2.4	50	10	28.5	5.7	20.45	4	12	2.4

water. In this paper, we analyzed existing solutions used to minimize energy consumed in computer networks and proposed a new adaptive technique that overcomes some evident disadvantages in current techniques. Our results are promising leading to a 76 % decrease in energy consumption over a normal network operation.

References

1. Aruna Prem Bianzino, Claude Chaudet, Dario Rossi, Jean-Louis Rougier: Survey of green Networking Research. Telecom parishtech (2010)
2. D. Pamlin and K. Szomol´anyi: Saving the Climate @ the Speed of Light–First Roadmap for Reduced CO^2 Emissions in the EU and Beyond. World Wildlife Fund and European Telecommunications Network Operators' Association (2007)
3. S. Nanda and T.-C. Chiueh: A Survey on Virtualization Technologies.Tech. Rep. TR179, Department of Computer Science, SUNY at Stony Brook (2005)
4. S. Nedevschi, L. Popa, G. Iannaccone, S. Ratnasamy, and D. Wetherall: Reducing Network Energy Consumption via Sleeping and Rate-Adaptation. In Proceedings of the 5th USENIX Symposium on Net worked Systems Design and Implementation (NDSI2008), (San Francisco,California, USA), (2008)
5. G. Da Costa, J.-P. Gelas, Y. Georgiou, L. Lefevre, A.-C. Orgerie, J.-M. Pierson, O. Richard, and K. Sharma: The GREEN-NET Framework: Energy Efficiency in Large Scale Distributed Systems. In Proceedings of the High Performance Power Aware Computing Workshop (HPPAC2009) in conjunction with IPDPS 2009, Rome, Italy (2009) 1–8
6. J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, and S. Wright: Power Awareness in Network Design and Routing. In Proceedings of the 27th IEEE Annual Conference on Computer Communications, INFOCOM 2008) (2008) 457–465
7. H. Hlavacs, G. Da Costa, and J.-M. Pierson: Energy Consumption of Residential and Professional Switches. In Proceedings of the IEEE International Conference on Computational Science and Engineering (CSE 2009), vol. 1, Vancouver, Canada (2009) 240–246
8. P. Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan: A Power Benchmarking Framework for Network Devices. In Proceedings of IFIP Networking 2009, Aachen, Germany (2009)
9. S. Albers: Energy-efficient algorithms. Communications of the ACM, 53 (5), 86–96 (2010)
10. M. Gupta and S. Singh: Greening of the Internet. In Proceedings of the ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM 2003), Karlsrhue, Germany (2003) 19 – 26
11. C. Gunaratne, K. Christensen, and B. Nordman: Managing energy consumption costs in desktop PCs and LAN switches with proxying, split TCP connections and scaling of link speed. International Journal of Network Management, (2005) 15(1) 297–310
12. C. Gunaratne, K. Christensen, and S. W. Suen, "Ethernet Adaptive Link Rate (ALR): Analysis of a buffer threshold policy", in Proceedings of the IEEE Global Communications Conference, GLOBECOM, California, USA, (2006)
13. C. Gunaratne, K. Christensen, B. Nordman, and S. Suen, "Reducing the Energy Consumption of Ethernet with Adaptive Link Rate (ALR)," IEEE Transactions on Computers, (2008) 57 (1), 448–461
14. A. Wierman, L. L. H. Andrew, and A. Tang: Power-Aware Speed Scaling in Processor Sharing Systems. In Proceedings of the 28th Annual IEEE Conference on Computer Communications, INFOCOM 2009, Rio de Janeiro (2009)
15. Tucker et al: Energy Consumption in IP Networks, 34th European conference on optical communication (2008) 11- 16

Improving TCP Performance in Mix Networks

Mohamed K. Watfa, Mohamed Diab, and Nikhil Stephen

1 Introduction

Advances in communication technology over the years have broken geographical barriers, making communication networks ubiquitous and seamlessly accessible. Widespread use of these networks by the public has led to the development of network-based personalized services involving sensitive and private information. Security is what we can understand, but privacy is what we need to control. With the rapid growth of internet and the wide use of wireless communications many concerns have been raised regarding user privacy and security. People, cars, and even homes are now communicating through wireless devices which provide rich and unified network connectivity. In the next generation internet, the use of these wireless computing devices will dramatically increase which will worsen the issue of users' privacy. Furthermore, the popularity of ubiquitous devices created new threats such as user tracking which is a common privacy threat in a large growing wireless mobile world. These risks have been already identified by others and have even been highlighted in some particular technologies such as RFID and 802.11 tracking and profiling.

These privacy breaching risks have a unified treatment because regardless of the associated technology, they share the same problem as tracking tasks rely on the user's transmission of unique identifiers. The current security and encryptions techniques could ensure the confidentiality of the message contents, but cannot ensure user profiling and tracking. Consequently, anonymity has become essential in modern network communications.

In 1981, David Chaum [14] pioneered the basic idea of the anonymous communication systems and he proposed Mix Network (mixnet) or mixes technique. A mixnet is a multistage system that accepts an input batch of quantities and produces an output batch containing the cryptographically transformed, permuted input batch. The change of appearance and the random reordering of the batch by the mixnet prevents trace back from output to input, hence achieving untraceability between the input and output batches. The design of a mixnet is based on providing anonymity for a batch of inputs, by changing their appearance and removing the order of arrival information. As shown in Fig. 1, the main component of a mixnet is the stage, also known as the mix, that performs mixing on a batch of inputs. Note that the inputs may arrive at the stage at different times. The mixing operation involves a cryptographic transformation using either decryption or encryption, that changes the appearance of inputs, followed by a permutation on the batch of transformed inputs. The mixed batch is then forwarded in parallel by the stage at time Tout to the next destination. The batching and the permutation together hide the order of arrival information of the inputs.

The work in [14] spurred significant interest in the development of anonymous channels for communication networks and for network applications requiring privacy. Mix networks depend on data packets batching and reordering to improve anonymity. A successful anonymous communication system requires two basic elements: First, the level of user's anonymity the system can accomplish. Second, the quality of service (QoS) a network service can deliver. Balancing between these two elements is considered as the real challenge in mix networks.

Batching and reordering false fast retransmit is very likely to occur in mix networks which limits both the maximum congestion window and, thus, the TCP throughput. In this paper, we analyze different studies on mix networks TCP performance in flow based applications and introduce a hybrid approach for improving the TCP performance using an intelligently adapting TCP's duplicate threshold `dupthresh` approach based on monitoring real packet loss.

M.K. Watfa (✉) • M. Diab • N. Stephen
Faculty of Engineering, University of Wollongong in Dubai (UOWD), Knowledge Village, Dubai, UAE
e-mail: MohamedWatfa@Uowdubai.ac.ae; diabeg@gmail.com; nikhil.stephen@gmail.com

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_62,

The rest of this paper is organized as follows. In Section 2, related research work is summarized. Section 3 presents detailed analysis and comparison of existing algorithms followed by simulation results in Section 4. We conclude this paper in Section 5 followed by a list of references.

2 Related Work

User privacy protection in a network with ubiquitous devices is a very real concern nowadays. General consensus is that these concerns are not well addressed by the current technology. There are known design challenges like the need to conceal identities and not to leak them, yet reveal enough information to trusted parties [1]. The following key properties of ubiquitous devices create privacy risks: 1) Mobility; 2) Wireless connectivity; 3) Embedded use; 4) Diversity and 5) Scale [2]. Based on anonymous communication as proposed by [3], there have been numerous proposals including the Tor system [4] which have looked to counter the risk factors enlisted above. Mixed networks conceals endpoints in communication using concepts like batching, reordering and multiple layers of encryption. The main disadvantages of these solutions are their overhead and complexity. Strong anonymity can be counterproductive by reducing accountability which makes this technology less used broadly.

A mix network consisting of multiple mix servers can provide enhanced anonymity as researched in [15, 16, and 17]. In a mix network, senders route their messages through a series of mixes. Therefore, even if an adversary compromises one mix and discovers the correlation between its input and output messages, other mixes along the path can still provide the necessary anonymity. A sender can choose different routes for each message or use one route for all his or her messages through the mix network. There are several research works that encourage the use of mix networks by improving performance and hence reducing the overhead. Mix networks can be used in flow based or message based applications. Cypherpunk remailer by Hughes and Finney [5] and Mixmaster [6] are two message based email anonymity applications. Popular examples of low latency flow based communication which have used a core mix network are Tor [4] and Freedom [7].

The major two concepts which we choose to concentrate in this paper are TCP's duplicate threshold `dupthresh` and actual packet loss monitoring. RR-TCP [8] is one algorithm which can dynamically change its `dupthresh`. RR-TCP uses a loop control with selective acknowledgments (SACKS) and DSACKS, and fast transmit events and time-out events to do so. But experiments in [9] show that RR-TCP is not feasible for mix networks. Since mix networks involve batching and re-ordering, the impact of out-of-order packet delivery is very important. There have been a lot of research on this topic as can be seen in the work of Kats and Ludwig [10] which used a scheme of timestamps. S. Bohacek et al [11] proposed an idea which used timers to check how long ago a packet was transmitted. Differentiating between actual packet loss and packet loss due to out-of-order packets by using additional sequence numbers stored in the TCP header was also suggested in [12].

3 Algorithm and Experimental Analysis

A TCP connection transmits packets in bursts. The number of packets sent in one burst is the instantaneous congestion window size `cwnd` in the case of no packets dropped and a large enough receiver advertised window. When a mix node receives a burst of packets from a sender, it may change the order of packets before forwarding them to the next mix or receiver. For example, the sender transmits packets 1, 2, 3, 4, and 5 in order, while the receiver (after one or more mixes reorder the packets) may receive packets in the order 2, 3, 4, 5, and 1. A TCP receiver sends an immediate duplicate ACK whenever an out-of-order segment arrives. By the time packet 5 has been received, the receiver has already generated three duplicate ACKs, because packet 1 has not yet been received. The three duplicate ACKs cause a false fast retransmit at the sender, which assumes that the three duplicate ACKs signal a packet loss. The sender exercises the fast recovery and the TCP congestion avoidance process and cuts the TCP instantaneous `cwnd` in half. Intuitively, such unnecessary retransmits will have a significant impact on TCP throughput in a mix network since the size of instantaneous `cwnd` limits the maximum number of packets the TCP sender can send at one time. False fast retransmit is very likely to occur in a mix network with batching and reordering applied, which limits both the maximum congestion window and, thus, the TCP throughput.

Xinwen Fu et al[9] implemented an algorithm for improving TCP performance in flow-based mix networks by increasing `dupthresh` value. Their intuitive approach was to suppress false fast retransmits by increasing the number of duplicate ACKs that the sender must receive before it realizes that the network has dropped a packet. The regular `dupthresh` is fixed at three duplicate ACKs in the fast retransmit specification. They proved through theoretical analysis and simulations that there is a positive linear relation between TCP Throughput and `dupthresh`. They presented two theorems as a theoretical evidence that TCP performance is improving by manipulating the duplicate threshold in a mix network. Their simulation illustrated that there is a changing trend of the maximum congestion window and TCP throughput whenever the `dupthresh` increased.

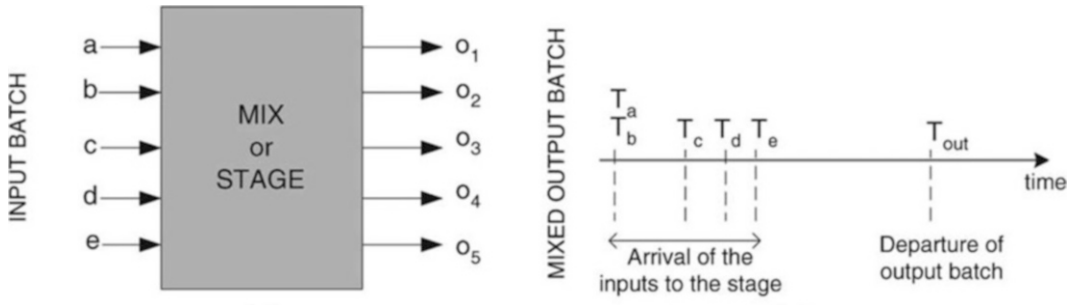


Fig. 1 Mixing by stage changes appearance of inputs and also removes order of arrival information. Output batch is a permutation of the transformed input batch.

Let D represent the `dupthresh` and $E(w)$ is the maximum congestion window, as can be noted in Eq(1), when D increases, $E(w)$ increases, which would lead to the increment of the TCP throughput.

$$E(w) = \sum_{n=D+1}^{\infty} nP(W = n) \leq \sum_{n=D+1}^{\infty} \frac{D^{n-D} \times (D-1)}{n-1} \quad (1)$$

The limitation of this solution is because of larger `dupthresh` values, TCP sender responds more slowly after real packet drops. When a real packet loss happened, the TCP sender waits for duplicate ACKs to start fast retransmit and fix the connection. If `dupthresh` is too large, duplicate ACKs may not be enough to trigger fast retransmit, and this will generate time-outs and which will lead to a longer packet delay. Even if enough duplicate ACKs return when a real packet loss occurs, the fast retransmit will still be delayed until the sender receives all of the required duplicate ACKs. This will dramatically increase the end-to-end delay for dropped packets.

We have seen the drawback of increasing *dupthresh* in our analysis. In the related work Section, we mentioned RR-TCP which uses the idea of *adaptive dupthresh* which was not practical to use in mix networks. We would like to build on the idea of *adaptive dupthresh* by combining with the idea of a packet loss monitor. The challenge we identified in changing `dupthresh` is how to differentiate between actual packet loss and packet loss due to batching and re-ordering. Explicit Congestion Notification (ECN) [13] is a good way of identifying actual packet loss. Routers configured with the function of ECN can notify initial congestion where the notification can sometimes be through marking packets rather than dropping them. Dropping packets is used by RED (*Random Early Detection)* as a method of congestion notification. Bits 6 and 7 in the IPv4 TOS octet are designated as the ECN field. The 6th bit is responsible for the ECN-Capable transport (ECT) flag that is set by the data sender to indicate that the senders and receivers of the transport protocol are ECN-capable. The 7th bit is the Congestion Experienced (CE) bit that is modified by the router to inform the receiver that the IP packet has experienced congestion. When a CE packet has been received, the receiver can set ECN-Echo (ECE) flag in the TCP header (Bit 9) so as to inform the data sender. A Congestion Window Reduced (CWR) flag in the TCP header (Bit 8) is set so that the sender can inform the receiver that the congestion window has been reduced. The characteristics of the ECN can help the sender to distinguish between packet losses due to high bit errors and packet losses due to network congestion.

Our proposed approach uses ECN to identify actual packet data loss and uses this information while adapting `dupthresh`. The intuitive result would be that when there is actual packet data loss, `dupthresh` would be reduced so that fast retransmits will be enabled faster and hence flow of packets will be better. On the other hand, if the packet loss monitor does not identify actual packet data loss, the `dupthresh` can be maintained high to enable high performance.

4 Results and Analysis

We have covered in our analysis how throughput increases with increasing `dupthresh`. In this section, we present simulation results to validate our analysis in Section 3 and to evaluate the adpative dupthresh-based approach combined with ECN for improving TCP QoS in a mix network. These results are obtained by using the popular network simulation software ns-2. We have implemented different mix boxes with different batching and reordering strategies in ns-2. A mix box is an ns node that should be placed between sender and receiver nodes. The following simulation results are obtained in a one-mix network where there is only one mix between the sender and receiver. Our simulation setup is the classical dumbbell topology, which is used for various TCP performance studies. We intentionally set the same bandwidth for all links, so there will be no bottleneck link and congestion. We set the queue size for all links to infinity. Therefore, if there

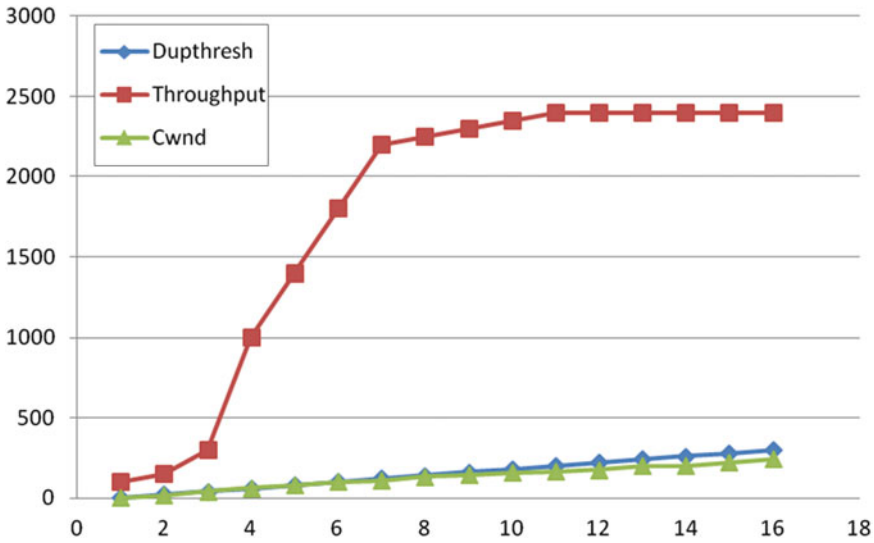


Fig.2 Throughput and `cwnd` increases with the increasing `dupthresh`

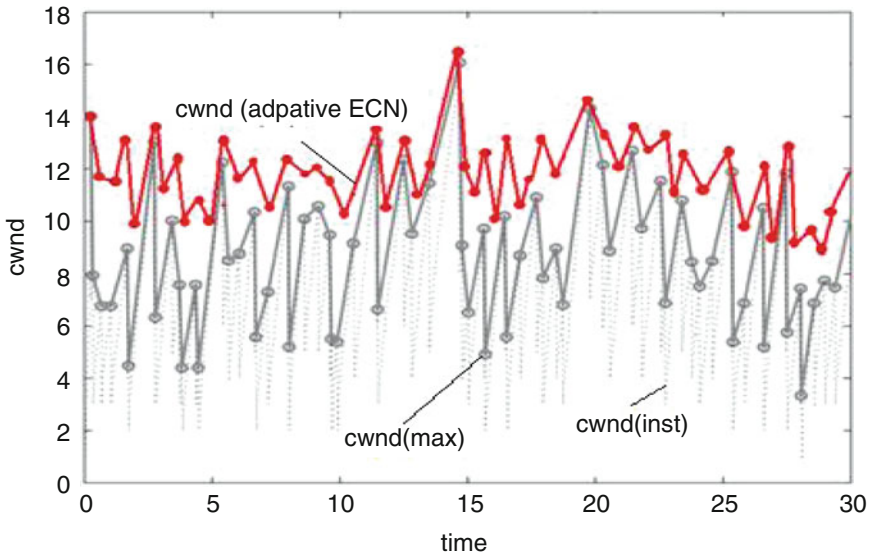


Fig. 3 The congestion window as time progresses using our modified adaptive algorithm.

is any fast retransmit, it is caused by batching and reordering strategies. We set the packet size to 512. We set the advertised window for TCP flows to infinity. An FTP session is created between the sender and receiver.

We discussed several attempts in improving TCP performance by increasing the duplicate ACK threshold `dupthresh`. Fig. 2 shows the changing trend of the mean of the maximum congestion window and throughput in terms of the increasing `dupthresh` respectively. By increasing `dupthresh`, we can increase TCP's maximum congestion window. Increasing the maximum congestion window necessarily implies a corresponding increase in TCP throughput. The limitation of this solution is because of larger `dupthresh` values, TCP sender responds more slowly after real packet drops. When a real packet loss happened, the TCP sender waits for duplicate ACKs to start fast retransmit.

In the second experiment, our modification of wireless TCP in sender is based on Reno version. The intermediate router is configured with the function of ECN in order to determine the cause of packet drops in a wireless network running the TCP protocol. This is done by observing the Congestion Experienced (CE) bit in the duplicate ACKs received if the TCP is ECN-capable. When the BER packet loss occurred, the sender estimates the packet loss rate.

The change in `cwnd` as we predict is shown in the Figure 3. It shows the changing trend of the congestion window with time. The dotted line is the instantaneous `cwnd` changing with time, the bold solid line is the maximum `cwnd` changing with time and the red line is our modified adaptive TCP algorithm using ECN. As expected, the `cwnd` is our modified version with ECN is less volatile and hence the congestion control will be called into action much less. This in turn will improve performance in mix networks.

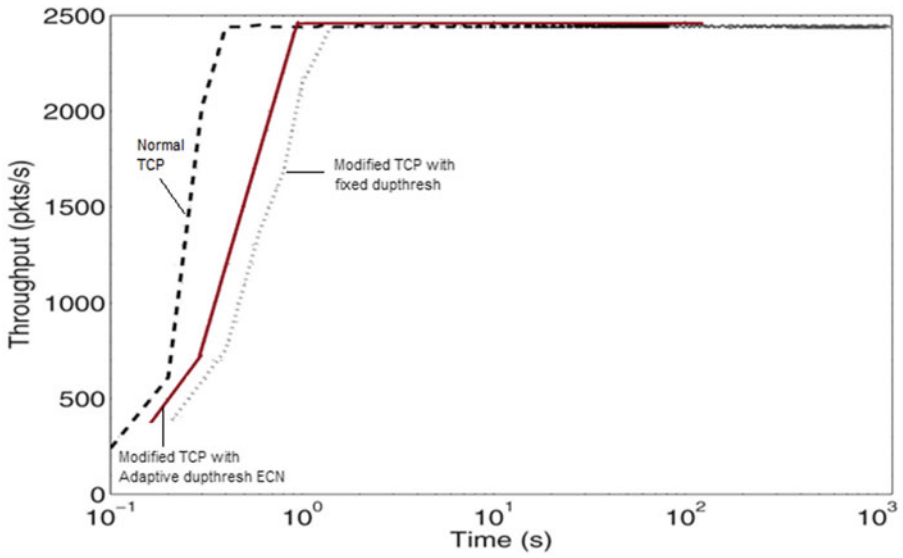


Fig. 4 Performance of adaptive `dupthresh` with packet loss monitor compared to normal TCP and modified TCP with fixed `dupthresh`.

Figure 4 shows the performance when we combine adaptive `dupthresh` and a modified TCP with ECN which will give information about the actual packet loss. Our modified TCP can dynamically change the duplicate threshold. Because of the increasing `cwnd`, TCP throughput increases to the maximum. However, since it keeps track of the actual packet loss, the transit time for our modified version is faster than for TCP Reno with a static `dupthresh` in a mix network to reach the maximum throughput.

5 Conclusions

Anonymity can be provided by cryptographic constructions and network architectures. But, for anonymous communication in terms of untraceability from the receiver to the sender, the mixnet is the most attractive solution when both security properties and scalability need to be satisfied. The peer-to-peer-based anonymity solutions can offer unconditional sender and receiver anonymity and are attractive in terms of protection from the consequences of the use/misuse of anonymity. However, in their current form, these solutions fail to meet performance requirements. In this paper, we examined the degradation of TCP performance in flow-based mix networks incorporating batching and reordering techniques. Our analysis and simulation results demonstrate that TCP performance dramatically degrades in such a mix network. The reason is that TCP throughput has an approximately linear relationship with the mean of the maximum congestion window. To improve TCP performance in such a flow-based mix network, we examined increasing TCP's duplicate threshold parameter `dupthresh` adaptively while keeping track of real packet loss using ECN. Our simulations show that we can intuitively be confident that the combination of ECN as a packet loss monitor and adaptive `dupthresh` in Mix network should make it a popular choice in flow based applications.

References

1. Paul, S et. al: Architectures for the future networks and the next generation Internet: A survey. Computer Communications, (2010), 34(1), 2-42
2. Seshan, S. : Collaborative Research NeTS-FIND: Protecting User. (2010), available at: www.nets-find.net/Funded/pdf_files/Protecting.pdf
3. Chaum, D.: Untraceable electronic mail, return addresses, and digital pseudonyms. Communications, (1981), 4(2), 84-90.
4. Dingledine, R. et al.: Tor: The second-generation onion router. Proceedings of the 13th conference on USENIX Security Symposium - Volume 13 (2004).
5. Parekh, S: Prospects for Remailers-Where Is Anonymity Heading, (1996). available at: http://freehaven.net/anonbib/cache/remailer-history.html.
6. Lier, M., Cottrell, L.: Mixmaster Protocol-Version, (2000). Available at: http://www.eskimo.com/~rowdenw/crypt/Mix/draft-moeller-mixmaster2-protocol-00.txt
7. Reed, M.G., Syverson, P.F., Goldschlag, D.M.: Anonymous connections and onion routing. Selected Areas in Communications, IEEE Journal on, (1998) 16(4), 482,494.
8. Ming Zhang; Karp, B.; Floyd, Sally; Peterson, L.: RR-TCP: a reordering-robust TCP with DSACK. Proceedings of the 11th IEEE International Conference on Network Protocols, (2003), 95–106.
9. Xinwen, Fu, et al.: TCP Performance in Flow-Based Mix Networks: Modeling and Analysis. IEEE Transactions On Parallel And Distributed Systems, 20(5), (2009).
10. Katz, R.: The Eifel Algorithm: Making TCP Robust against Spurious Retransmission. ACM Computer Comm. Rev, 30(1), 30-36, (2000).
11. Bohacek, S.; Hespanha, J.P.; Junsoo Lee; Chansook Lim; Obraczka, K.,: TCP-PR: TCP for persistent packet reordering. Proceedings of the 23rd International Conference on Distributed Computing Systems (2003) 222-231.

12. Zhang, F., Zhang, Y.: Improving TCP Performance over Mobile Ad-Hoc Networks with Out-of-Order Detection and Respo. Proceedings of the 3rd ACM international symposium on Mobile ad hoc networking & computing, (2002) 217 - 225.
13. Han, H: Performance improvement of TCP_Reno based on monitoring the wireless packet loss rate. Proceedings of the 3rd IEEE International Conference on Communication Software and Networks (ICCSN), (2011) 469-472.
14. D. Chaum, B: Untraceable electronic mail, return addresses, and digital pseudonyms, Commun. ACM (1981), 24(2), 84–88,
15. G. Danezis, R. Dingledine, and N. Mathewson: Mixminion: Design of a Type III Anonymous Remailer Protocol. Proc. IEEE Symp. Security and Privacy (S&P '03), (2003).
16. M. Wright, M. Adler, B.N. Levine, and C. Shields: An Analysis of the Degradation of Anonymous Protocols. Proc. Network and Distributed Security Symp. (NDSS '02), (2002).
17. M. Wright, M. Adler, B.N. Levine, and C. Shields, "Defending Anonymous Communication against Passive Logging Attacks," Proc. IEEE Symp. Security and Privacy (S&P '03), May 2003.

An Epidemic Routing with Low Message Exchange Overhead for Delay Tolerant Networks

Teerapong Choksatid and Sumet Prabhavat

1 Introduction

Communications plays a major part in our lives. We always communicate with other people on many ways such as sending messages to friends who live far away, check hot news via social networks, etc. To make the convenient conversation for our lives, researchers have continuously improved the communications from the basically communicated systems (i.e., LAN) to wireless communication system (i.e., WLAN). At present, many researchers have developed the communication system that allows devices to communicate with each other without relying on infrastructure. This kind of communication system is known as Ad Hoc network.

In this paper we focus on Delay Tolerant Network (DTNs), which is a form of Ad Hoc networks. DTN nodes are continuously moved and have limited energy that make it difficult for establish the communication path from source nodes to destination nodes. Therefore, Store and Forward Messages Switching has been used as technique for forwarding the messages to other node time after time and storing the messages until the nodes are able to forward a message to destination node successfully. In addition, probability of a message reaching destination depends on the number of nodes which store destination's message.[1]

Although communication on DTNs network is very flexible, there are many problems that can reduce the performance of distributing messages. In this case, the authors pay attention to one biggest overhead: Message Exchanges Overhead. This overhead make traffic congestion of the communication path and limited energy in communicating nodes. Both problems are often found in DTNs, which always use broadcast to distribute messages continuously. Therefore, in this article, we propose development of Epidemic Routing to reduce Message Exchanges Overhead in DTNs without impact the performance of spreading messages.

Section 2 reviews of related literatures which attempted to improve communication strategies in DTNs. Then, in section 3, introduces our proposed routing strategy and performance evaluation along with comparison with existing schemes are presented in section 4. Finally, the section 5 concludes the development of this novel strategy.

2 Related Work

Delay-Tolerant Networks (DTN) is an interesting alternative of communication in Ad Hoc Network. It has ability to transmit messages between devices in situations where traditional communications technology cannot accomplish, such as wireless sensor networks, vehicular networks and mobile Ad-Hoc networks, etc. The natures of these networks show characteristics of network structure, which change node position over time. Then, the ability to distribute a message depends on the probability of encountering other nodes during message transmission. Moreover, because it sensitive to nodes mobility, the communication path could be difficult to predict and maintain. This is the reason why in next paragraph we describe the challenges in DTNs before develop Epidemic Routing to reduce Message Exchanges Overhead. [2]

The challenges that the authors are taking into account include: First, Kapadia S., Krishnamachari B. and Zhang L. mentioned clearly that "Contact schedules are information about pre-calculate and determine the best path to forward its information". It makes nodes blindness because it don't know who is the next contract that can forward messages to destination.[3, 4] Second, network capacity or link bandwidth is amount of data that can be exchanged between a pair of nodes. If the number of messages exchange increases or the messages' size gets bigger, the network have more data

T. Choksatid (✉) • S. Prabhavat
Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
e-mail: teerapongc@outlook.com; sumet@it.kmitl.ac.th

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_63,

transmission and will become congested finally other than unknowing messages that opposite nodes don't have. Third, most nodes in DTN have limited amount of energy, which will be consumed continuously when distributing messages until the nodes can no longer function because of insufficient energy. Therefore, it is important to conserve node's energy.

Direct contact is a simplest routing strategy to send a message. When the source nodes have to send the messages, they will not spread the messages to other nodes or calculate the best path to forward the messages, but it will wait until it comes in contact with the destination node and transmit the messages directly. [5] The flexibility of DTNs gives birth to a lot of routing strategies, which intend to resolve communication problems on different network environments. Jones E. and Ward P. have surveyed and classified routing strategy into two categories: Flooding Strategies which rely on distributing messages to different nodes, even when the nodes continuously move unlike the Forwarding Strategies that collect information from network to analyze and select the best communication path to forward the messages to next nodes. [6, 7] In addition, Jones E., Li L., Schmidtke J. and Ward P. has proposed Minimum Estimated Expected Delay (MEED) path metric, which uses the observed contact historyto predict future contact schedule for estimate waiting times on each contact.[8]

The previous studies demonstrate the importance of selecting routing strategy for appropriate network situation, makes the authors focus on Epidemic Routing that has ability to provide message transmission on a network which nodes move randomly. In such situation, it is impossible to predict when two nodes will come in range of one another. Anti Entropy is a pattern of Epidemic Routing of Flooding Strategies. It works by letting the nodes exchange a list of messages that used to check message which have and don't have. After that, the node can send only other request instead of sending all messages as shown in Figure 1.

From Figure 1, when node B gets to know node A via Hello. Node B will send Message List to determine message that stored in it. Then, node A uses Message List Request to request copies of messages that it has been unseen. This message causes node B can send some messages instead of sending all messages. Show that Anti Entropy protocol can resolve network congestion, which usually occurs on a network that nodes have to exchange a lot of messages such as nodes operating under Flooding Strategies. Additionally, Vahdat A., Becker D. and others have further developed Anti Entropy by allowing nodes to only request the messages that they need. [9]

In recent years, Lu X. and Hui P. have presented a routing strategy to reduce the loss of nodes' power by reduce time for sending a message. They use the advantage of communication on wireless device that transmits data by broadcasting. The wireless device (node) in this strategy will send

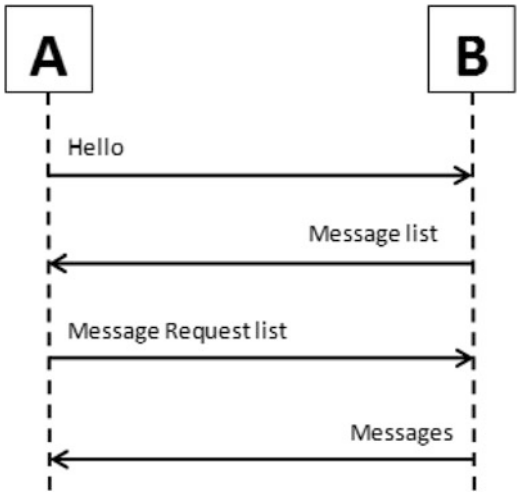


Fig. 1 Anti Entropy

broadcast messages when neighbor nodes reach the specific number. Then, each broadcasting can increase the number of nodes that receive messages ($\geq$ *the specific number*) while broadcast for one time. [10] Moreover, De Rango F. and Amelio S. also developed n-Epidemic Routing that provides a mechanism to select forward messages (Dynamic Forwarding) by consider the density of nodes in each group.[11] Example of n-Epidemic shows in Figure 2. Wireless device (Phone) define number of neighbor nodes equal 3. They will broadcast Messages after receiving Hello packet from neighbor nodes reach 3.

3 Routing Strategies Description

The reason that Epidemic Routing can send messages in network that has unpredictable movement of nodes is shown in Figure 3.

Epidemic Routing is one of Flooding Strategies, which allows a pair of nodes to exchange messages to each other until they find the destination nodes. Figure 3 shows a group of nodes with node S that cannot directly send message to node D due to node D being out of transmission range of node S. At time t1, node S sends a message to node C1 and C2 to increase the number of nodes that received the message in the network. After that, node C1 sends the message to node C3. At time $t2 > t1$, since the mobility of nodes are random, some nodes, such as node C3 may come within range of node D. Therefore, node D can receive a message originally transmitted from node S from node C3. Additionally, the messages will not be lost when source nodes disappear, because messages have already been transmitted to other nodes. However, when consider the number of messages in another view, we can find a lot of Message Exchanges Overhead that

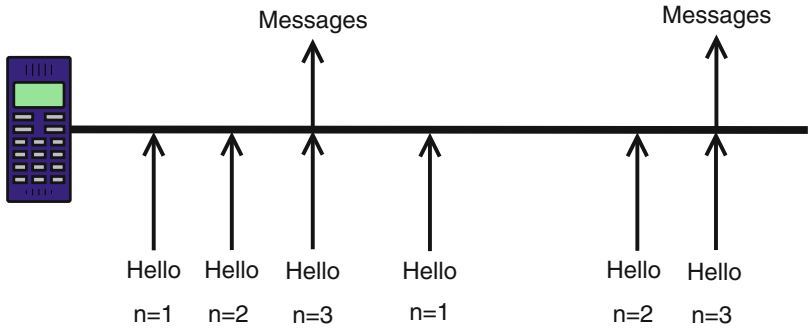


Fig. 2 n-Epidemic

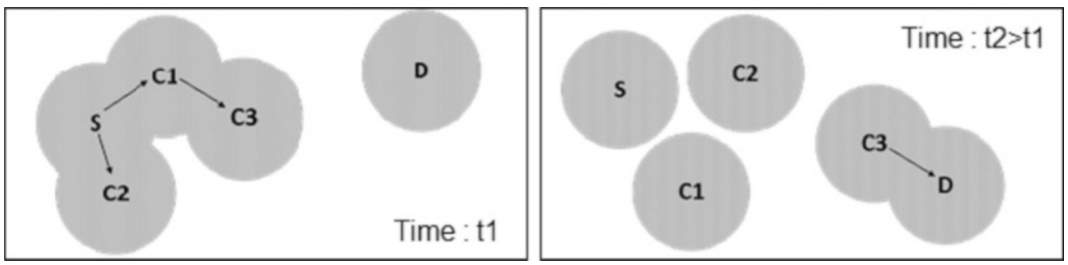


Fig. 3 Epidemic Routing

increase the congestion in the network and increase energy consumption during data transmission.

Anti-Entropy and n-Epidemic approach have inspired the authors to develop Epidemic routing that can potentially reduce the number of messages exchanging on network without sacrifice delivery time. We will be referingreferring to Epidemic Message with Message List Advertisement as EM-MA for the rest of the article.

3.1 EM-MA Description

EM-MA works by spreading messages to surrounding nodes. The process occur when source node announce its messages to neighbor nodes in period of time and then neighbor nodes will check its requirements and send request back. Finally source node collects messages that have been requested and forward those nodes in next period time.

Figure 4 describes the process of EM-MA. Upon the expiration of each period time (T), Announcer will broadcast Message List Advertisement (Message List) to the surrounding Receivers. When these Receivers have received Message Lists, they will be created Message Request List to request messages which it doesn't have from Announcer's messages. Then, Announcer will respond by forwarding those messages (Messages[n]) in next interval time. In practice, we will be add a slight random delay before broadcast Message List. This delay is used to prevent synchronization similar to DIFS in IEEE 802.11 [12]. For simplicity, we assign each node starts announcement not same time in simulations. Then, a slight random delay is omited.

We expect that, One: using Message List instead of Hello, Two: the message aggregation of several Message Request List, Three: assigned nodes only send messages in their period time which resemblance to Slotted Aloha, can reduce retransmission due to Collision as proven in the article [13] and the number of Message Exchanges Overhead. EM-MA features detailed below.

Advantage of Message List Advertisement. The different between Anti Entropy and EM-MA is Anti Entropy's node send Message List after recive Hello message. It's like they wait for some greeting from a friend. However, EM-MA reduces each step by creating "All in one Message". Each node will be announced Message List based on the time interval continuously and sometimes it is also attached messages that have been requested by another nodes. Sending message once can communicate with surrounding nodes: "I live here.", "I have had these things.", "She wants something from me?" and "And the last time she asked for these things from me."

Advantage of time period. In Figure 2, n-Epidemic proposed the mechanism for saving energy by the broadcasting of wireless transmission. In the mechanism, each node needs to wait until the number of neighber nodes reach the defined number for broadcasting messages. [10] This process is similar to the process of EM-MA shown in Figure 4. When

announcer broadcasts Message List, it collects Request Message List from surrounding nodes over a time period before attach the messages that have been requested to the next Message List.

3.2 A variety of Interval Time

Because of the performance of EM-MA depends on the period time. In this section, the authors have created simulations for messages exchange between nodes in network environment that is defined the detail in Section 4.1.

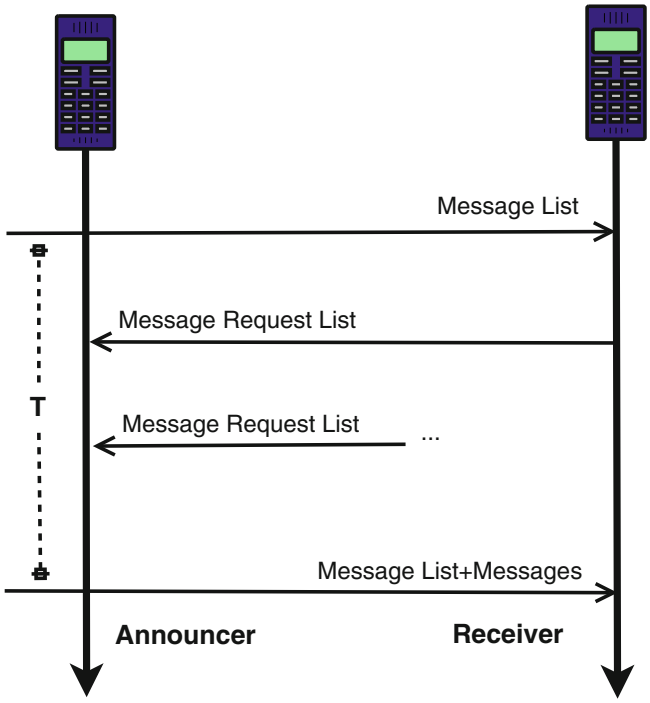


Fig. 4 EM-MA protocol

Message Transmission Rate. Message Transmission Rate (MTR) is total number of messages that were sent in the network per time units. The reduction of it means that the amount of data transferred each time units and number of data transmission rate decreased. Network Performance reduces when MTR increases due to the network being less congested. At the same time the network's nodes have highly energy consumption due to exchange big or a lot of messages.

Figure 5 shows average MTR with varying number of nodes in the network. The MTRs of different intervals of Broadcast Message are presented. The figure shows that the result of the scenario with 0.5 seconds Broadcast Message interval attains maximum MTR, followed by result of 1 second interval. Finally, the scenario with 2 seconds interval has minimal MTR. This means the network has a fewer time period to send fewer messages that makes nodes increase frequency of broadcast messages. MTR will increase with the inevitable.

Delivery Time. Delivery Time is the length of time it will take for all messages to be delivered from source nodes to their destinations. Simulation results are presented in Figure 6. From the graph, it is seen that the result of scenario with 0.5 second, scenario with 1 second and scenario with 2 second are slightly different. Although the increase of the period to send messages will cause the reduction of Message Transmission Rate, but it did not make Delivery Timetime to communicate differently.

4 Performance Evaluation

To demonstrate that EM-MA can resolve network congestion and increases energy efficiency when comparing with existing methods without having an impact on the performance of spreading messages. We present Message Transmission Rate Graph to indicate the reduction of

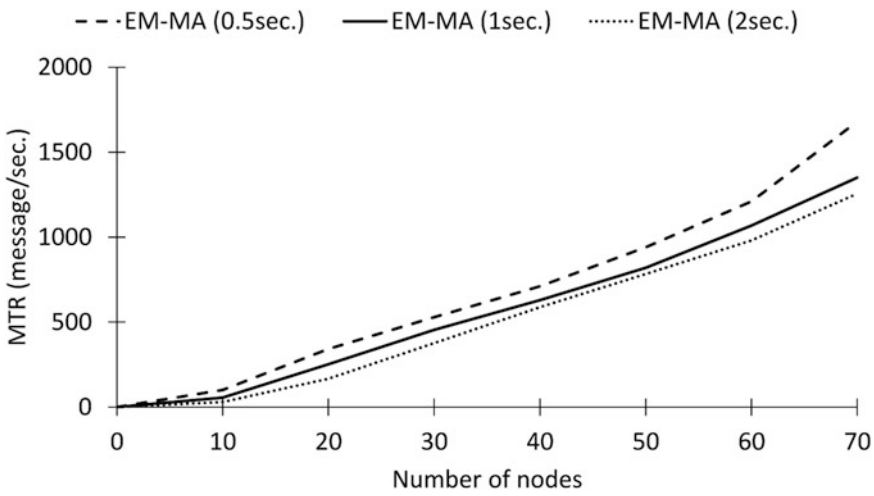


Fig. 5 Message Transmission Rate of EM-MA on the width of the period time

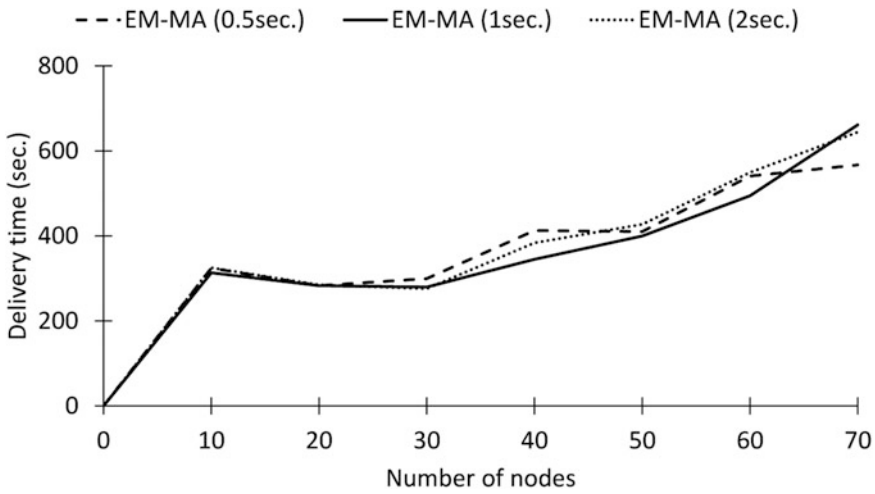


Fig. 6 Delivery Time of EM-MA on the width of the period time

Message Exchanges Overhead. However, the problem of duplicating and distributing messages will occur when this overhead decrease. Then, we show message delivery time to confirm that EM-MA does not impact the performance of duplicating and distributing messages. This evaluation study is the same way with the study article of Samo G. and Anders L., where they study and develop models for test DTNs routing strategies as to communicate in the real world. Lead to modeling 4 strategies: Epidemic (Direct Mail), Anti Entropy, n-Epidemic and EM-MA. The experimental results show through following graphs. [3]

Table 1 Simulation Parameters

Parameters	Value
Number of nodes	10-70
Simulation area	1500 m x 300 m
Transmission range	100 m
Message length	1 KB
Period of message list advertisement	1 sec
Max. transmission rate	54 Mbps
Mobility model	Random waypoint
Mobility rate	0-20 m/sec

4.1 Network Environment

Simulation parameters are listed in Table 1. In the simulation scenario, 90 percent of the total nodes generate messages and the other 10 percent only act as relay and do not generate any message. Additionally, a node that generate messages will try to send message to all other nodes that can generate messages. For example, if the network has 50 nodes, there are 45 nodes that can generate messages and one node will create a total of 44 messages to be sent to other 44 nodes However, the other 5 nodes will not generate any messages. Therefore, to test the communication network with 50 nodes, 1,980 messages have to be created. Furthermore, In order to make communication in the network more realistic, we assume that the node's processing delay range from 0 to 0.01, with uniform distribution.

We use a simple mechanism called Direct Mail to represent a basic mechanism of Epidemic routing. Because of Direct Mail's node will send all of its messages when receive Hello packet from any node in the network. [14] Moreover, Because n-Epidemic can be configured within $n \geq 1$, so in this experiment has set $n = 3$ which is the middle of the test ($n = 1$-5) occurred in [10]. Likewise, announcement interval in EM-MA can be set to many value. In section 3, we see that if the value is increasing, MPR is also increasing. But the change values does not affect the delivery time. Therefore, in this simulation, the interval is equal to 1, which is similar to the intevalinterval of sending Hello message in other strategies that were used in the comparison.

4.2 Number of Message Exchanges

Message Transmission Rate. In figure 7, show Message Transmission Rate that change by density of node. X-axis is the number of nodes in network and Y-axis is the number of messages per second that have been sent out to the network. Each line represents the routing strategy types. In line with high level of messages transmission would have a lot of Message Exchanges Overhead which cause network congestion and lost energy more than the line that have lower. This graph shows that when density of nodes in the network increased, the MTR also increased. Direct Mail has the highest MTR, followed by 3-Epidemic, Anti Entropy and EM-MA (1sec.) respectively. It evidences that MTR of EM-MA can reduce the number of message exchanges when compare to other strategies in this

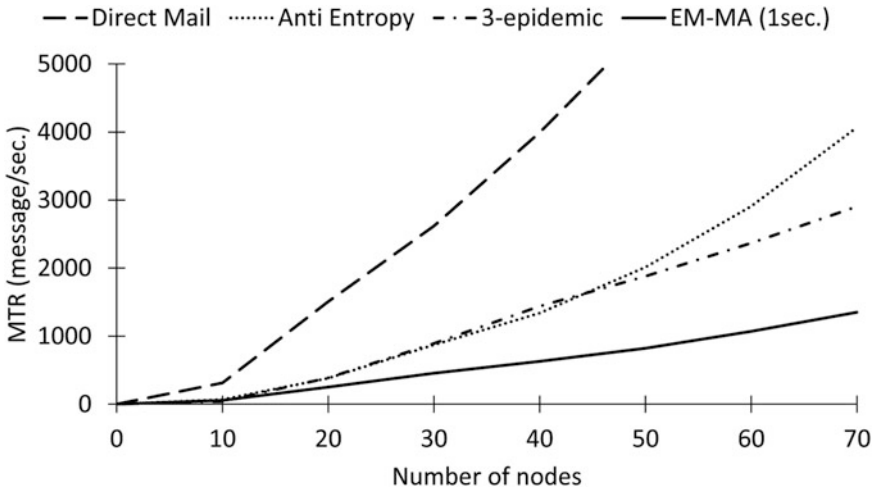


Fig. 7 Message Transmission Rate of Direct Mail, Anti Entropy, 3-Epidemic and EM-MA

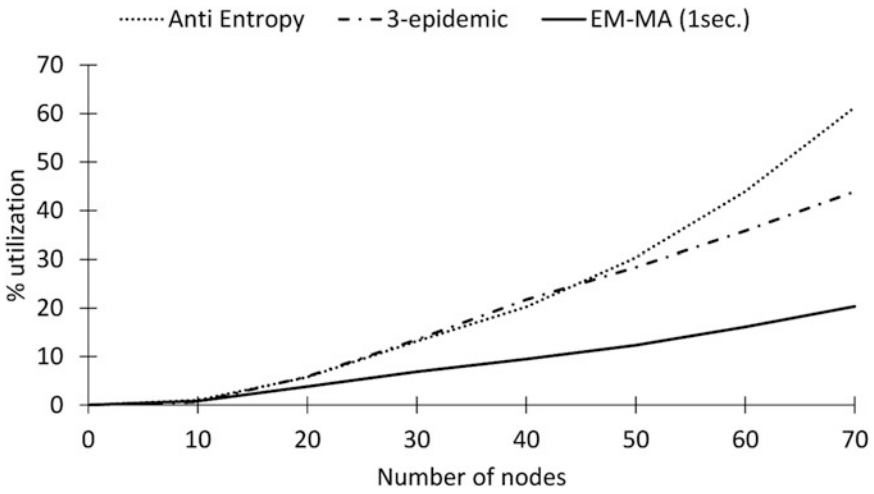


Fig. 8 Utilization Percentage of Direct Mail, Anti Entropy, 3-Epidemic and EM-MA

graph. MTR of Direct Mail have been clearly higher than other strategies because the slope of Direct Mail line is higher than other lines. Therefore, in Utilization Percentage graphs, we don't considerate Direct Mail.

Utilization Percentage. Network utilization is the amount of traffic on the network comparing to the peak amount that the network can support. It compares to traffic on a road while cars passing each other all time. If it has a few vehicles on the road, traveling would be comfortable but if there are many vehicles, the opportunity to meet the traffic jam and accidents on the road are highly. Therefore, Utilization Percentage can show up at network congestion and network collision. In this simulation, maximum data rate is the maximum number of bits that are transferred per unit of time. IEEE 802.11G assign Top Data Rate equal to 54 Mbps. [15]

Figure 8 is Utilization Percentage graph. It shows average Utilization percentage compared to density of nodes in the network. X-axis is the number of nodes in network and Y-axis is average Utilization percentage. Each graph line represents the routing strategies. Level of each line represents amount of message on network that network can support. Furthermore, average Utilization percentage also represents the chance of network congestion so the chance of network congestion of 3-Epidemic and Anti Entropy more than Utilization Percentage of EM-MA. In this graph, the value of each line shows the amount of messages that sent by nodes to the network. Therefore, it can be said that Message Transmission Rate and Utilization Percentage are sequel to each other. Therefore, Figure 8 and Figure 9 show that the simulations and the experimental result in accuracy significantly.

4.3 Delivery Time

Figure 9 shows that reducing number of exchanged messages does not affect the efficiency of EM-MA while duplicating and distributing the messages. Delivery Time is the time until all nodes receive their intended messages.

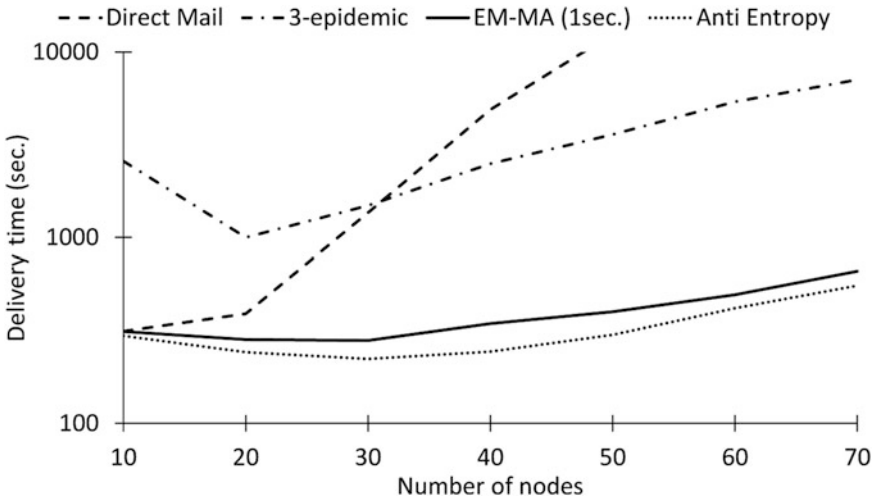


Fig. 9 Delivery Time of Direct Mail, Anti Entropy, 3-Epidemic and EM-MA

The graph shows the Delivery Time in Figure 9. X-axis is the number of nodes in network and Y-axis is average Utilization percentage. In this graph, Direct Mail spent the most time when compare to other strategies. 3-Epidemic is likely to increase the time period when number of nodes increases. Although EM-MA would take a little time less than Anti Entropy, but it comes with a number of lower Messages Exchanges Overhead as shown in Figure 7. Anti-Entropy spent minimal time because it do not have the delay process before sending messages like n-Epidermic and EM-MA (Figure 2 & 4) In addition, it will quickly send back the messages to other which contact itself. Therefore, it can spread the messages faster.

5 Conclusion

If the network has a small number of nodes that need to send messages to the destination node, routing protocol with high delivery ratio should be strongly considered. However, in a real situation, each node creates a message to communicate for various nodes in the network continuously. Two questions that important than Delivery Time is: How to increases life time of node? and How to transfer data in the network continuously? Those answers will prevent communication failures due to network congestion and losable powe. We proposed EM-MA and demonstrated it can reduce energy consumption of the nodes and decrease congestion in DTNs. Our simulation shows that the total number of messages sent by network operating under EM-MA is less than other existing strategies without affecting the message delivery time. Additionally, our simulation results also show that EM-MA is better at reducing network congestion than the other existing strategies. Therefore, EM-MA can potentially improve message transmission in DTNs.

References

1. S. Burleigh, A. Hooke, L. Torgerson, K. Fall, V. Cerf, B. Durst, K. Scott, and H. Weiss, "Delay-tolerant networking: an approach to interplanetary internet," *Communications Magazine, IEEE* **41**(6), pp. 128–136, 2003.
2. M. J. Khabbaz, C. M. Assi, and W. F. Fawaz, "Disruption-tolerant networking: A comprehensive survey on recent developments and persisting challenges," *Communications Surveys & Tutorials, IEEE***14**(2), pp. 607–640, 2012.
3. S. Grasic and A. Lindgren, "An analysis of evaluation practices for dtn routing protocols," in *Proceedings of the seventh ACM international workshop on Challenged networks*, pp. 57–64, ACM, 2012.
4. S. Jain, K. Fall, and R. Patra, *Routing in a delay tolerant network*, vol. 34, ACM, 2004.
5. S. Kapadia, B. Krishnamachari, and L. Zhang, "Data delivery in delay tolerant networks: A survey," *Mobile Ad-hoc Networks: Protocol Design*, pp. 565–578, 2011.
6. E. P. Jones and P. A. Ward, "Routing strategies for delay-tolerant networks," *Submitted to ACM Computer Communication Review (CCR)*, 2006.
7. Y. Cao and Z. Sun, "Routing in delay/disruption tolerant networks: A taxonomy, survey and challenges," *Communications Surveys & Tutorials, IEEE* **15**(2), pp. 654–677, 2013.
8. E. P. Jones, L. Li, J. K. Schmidtke, and P. A. Ward, "Practical routing in delay-tolerant networks," *Mobile Computing, IEEE Transactions on* **6**(8), pp. 943–959, 2007.
9. A. Vahdat, D. Becker, *et al*., "Epidemic routing for partially connected ad hoc networks," tech. rep., Technical Report CS-200006, Duke University, 2000.
10. X. Lu and P. Hui, "An energy-efficient n-epidemic routing protocol for delay tolerant networks," in *Networking, Architecture and Storage (NAS), 2010 IEEE Fifth International Conference on*, pp. 341–347, IEEE, 2010.

11. F. De Rango and S. Amelio, "Performance evaluation of scalable and energy efficient dynamic n-epidemic routing in delay tolerant networks," in *Performance Evaluation of Computer and Telecommunication Systems (SPECTS), 2013 International Symposium on*, pp. 167–173, IEEE, 2013.
12. IEEE-SA, "Ieee std 802.11-2012," 2012.
13. L. G. Roberts, "Aloha packet system with and without slots and capture," *ACM SIGCOMM Computer Communication Review* **5**(2), pp. 28–42, 1975.
14. A. Demers, D. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. Sturgis, D. Swinehart, and D. Terry, "Epidemic algorithms for replicated database maintenance," in *Proceedings of the sixth annual ACM Symposium on Principles of distributed computing*, pp. 1–12, ACM, 1987.
15. i. Cisco Systems, "Capacity, coverage, and deployment considerations for ieee 802.11g," 2005.

EEIS: an Energy Efficient at Idle Slots MAC layer Protocol for Wireless Sensor Networks

Usha Jhadane, Pramod Kumar Singh, and Abhishek Patel

1 Introduction

A WSN consists of spatially distributed autonomous sensors, which are used to monitor environmental conditions, such as pressure, temperature etc. and then their data is passed to the main location. In WSN, there are four basic components viz: (a)Sensor Nodes; equipped with microprocessor, radios and limited battery power; (b) An interconnection network (c) base station(BS) (for receiving all the data) (d) Some computing resources at the BS to handle data querying and data mining etc. Communication in a WSN can be divided into several layers. One of those is the MAC layer [3]. This layer can be described by MAC protocols, which try to ensure that no nodes are interfering each other at the time of transmissions and deal with the situation when they do. The proposed protocol is also a MAC layer protocol. Wireless sensor network has an additional aspect as: sensor nodes are generally battery-operated, which mean they have limited source of power. Therefore it is important to focus on the energy consumption, in the present era for WSN. Before designing an energy efficient MAC layer protocol, we should know the reasons of energy losses into MAC layer [4, 7, 2].

First and the main reason of energy loss, at which we have focused in the proposed protocol is idle listening. In variable traffic load, node does not know when it is the receiver of a message from one of its neighboring node, therefore it must keep its radios on or in the receiving mode. So energy loss occurs because of listening, of idle channel. And this kind of energy loss is known idle listening problem. As we know that receiving energy equals to the idle energy. Therefore at low duty cycles, most of the energy loss occurs due to the idle listening.

U. Jhadane (✉) • P.K. Singh • A. Patel
ABV- Indian Institute of Information Technology and Management, Gwalior, [M.P.], Madhya Pradesh
e-mail: mona.jhadane@gmail.com; pksingh7@gmail.com; mrabhi.patel@gmail.com

Other sources of energy losses are collision, overhearing and control Signal overhead. Therefore for an energy efficient MAC protocol, it is important that the protocol must have less idle listening, zero collision, minimum overhearing of signals and minimum control signal overhead. We have used cluster based architecture that tackles all the problems associated with idle listening and certain other issues.

Cluster based architecture uses Time Division Multiple Access (TDMA) and Carrier Sense Multiple Access (CSMA). Schedule based and contention based protocols represent themselves respectively. Schedule based protocols provide maximum energy efficiency, but they do not give flexibility to change with respect to traffic load in WSN. Contention based protocols provide flexibility and robustness by providing RTS and CTS control signals. These signals provide collision avoidance, when two nodes are sending data in the same channel at the same time. But these kinds of protocols provide more energy loss, because of control signal's overhead. As we described above, CSMA and TDMA both have their own advantages and disadvantages. Therefore we have proposed a hybrid MAC layer protocol, which has capabilities of both TDMA and CSMA based protocols, known as hybrid protocols [1]. There are many hybrid MAC layer protocols have been proposed for WSN [4, 5, 8].

In this paper an energy efficient MAC protocol has been proposed, using a hybrid access protocol, based on cluster based architecture for variable traffic load, which provides energy efficiency and deals with idle listening problem in low duty cycles.

2 Related Work

In the proposed protocol, cluster based architecture has been used. A cluster based protocol divides all the sensor nodes into clusters. So every sensor node has a unique CH. In this kind of architecture only cluster member nodes can

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_64,

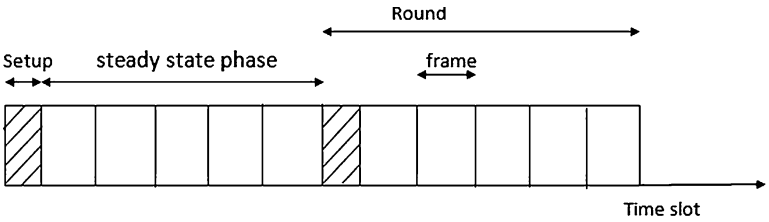


Fig. 1 Phases and Operations in leach protocol

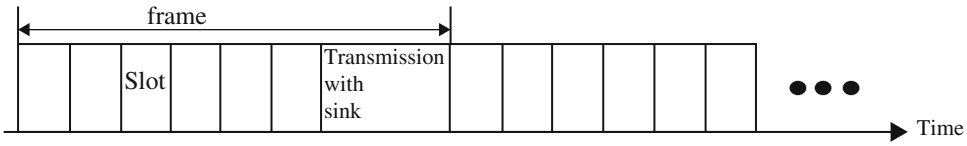


Fig. 2 Time frame format for the CH

communicate with their respective CH. But Member nodes cannot communicate with each other. CH receives data from member nodes then aggregates the data and sends it to the BS.

A number of cluster based protocols have been proposed for WSN [3, 2, 5, 8]. However, LEACH [4, 5] is considered as the most popular routing protocol, it uses cluster based architecture in order to minimize the energy consumption. But at variable traffic load or low duty cycle it faces a problem of idle listening, because of idle time slots. Therefore we have compared energy efficiency in variable traffic load, between LEACH and EEIS. Basic structure of LEACH protocol is given in Fig 1.

LEACH [4] is a hybrid access MAC layer protocol for WSN. It is based upon cluster based architecture. Cluster based architecture has both the property of CSMA and TDMA. First of all, all nodes are partitioned into clusters. Each cluster has a CH and only CH is responsible for creating TDMA slots for their respective cluster members. After accepting all the data from cluster members, the aggregated data are sent to the BS. In LEACH protocol, BS has infinite number of resources. Operations of LEACH protocol is divided into many rounds as shown in Fig 1.

There are 2 kind of phases defined into LEACH protocol

2.1 Setup Phase

2.2 Steady-State Phase

2.1 Setup phase:

Each round in LEACH begins with cluster setup phase. In this phase, CHs are created. Then they broadcast the schedule to the member nodes. So every member nodes has its own time slots for sending data to the CH.

2.2 Steady-state phase:

In this phase, Data transmission begins; member nodes send their own data within their allocated time slots to the CH, as shown in Fig 2. When one node sends data to the CH, then other member nodes go to the sleeping mode. After receiving all the data from the member nodes, CH aggregates that data and sends it to the BS.

LEACH is able to perform local aggregation of data in each cluster to reduce the amount of data are transmitted to the BS. Although LEACH protocol acts in a good manner, it suffers from many drawbacks.

One of them is, when a CH does not have data to accept into the low duty cycle, and then CH remains into the idle listening mode until the time slot of their respective member node finishes. As we know listening energy is equal to the reception energy, therefore energy loss will be more.

One more reason of idle listening is when a node becomes dead in a round, then the node time slot wastes and CH remains into sensing mode until the round finishes. There is more loss of energy due to idle listening at the CH, but in our protocol CH switches to the sleeping mode after waiting T_w time, means it remains into sleeping mode for T_{all}-T_w time slot. And then again it switches to the wake up mode. Therefore it provides high energy efficiency.

Since LEACH-C [6] provides energy efficiency by selecting CHs on the basis of available energy. In variable traffic load, there are still many other idle slots available. Therefore, the energy is wasted at the CH, especially under low traffic level. Now, we have introduced our new hybrid access protocol to cope with all these drawbacks.

We have used scenarios of CA-MAC [8] protocol and TEEN [11] protocol. These protocols work on the threshold value of data, if current value of data in a node is less than

threshold value, then cluster member sends data to the CH; else it does not send data to the CH. Main drawback of TEEN protocol was, if data is less than threshold, then data will not be sent to the CH. EEIS overcomes this overhead by sending whole data at the last frame's time slot in every round, from cluster members to the CH.

3 Proposed Methodology

EEIS protocol has improved energy efficiency by using it's scenario in variable traffic load and tackled with other problems like control signal overhead, over hearing of signals and collision also. It has saved CH's energy in idle slots.

Following assumptions are made in the architecture of the networks to design EEIS protocol:

- All nodes into the network, are homogeneous, i.e. all nodes have equal capacity with respect to communication capability and power computation.
- The sensor nodes in the network are aware of information about themselves, like the node IDs, locations and energy levels.
- The Sink node has unlimited resource.
- All sensor nodes are stationary.
- The number of transmitting nodes is varying, which mean, the system is event-driven and traffic load is dynamic.

There are two kinds of phases defined into EEIS algorithm

3.1 Set-up phase
  a) Advertisement phase
  b) Cluster setup phase
3.2 Steady-state phase

3.1 Set-up phase

Each node decides independently whether it wants to become a CH or not. They take decision, based upon the last round they have became a CH. The node that hasn't been a CH for long time is more likely to elect itself than nodes that have been a CH recently.

a) In the advertisement phase, CHs inform their neighbors with an "advertisement packet" that they are CHs into the current round. Then Non-CH nodes pick up the advertisement packets.

b) In the cluster setup phase, the member nodes inform their respective CH, that they have become a member to that cluster by sending "join packet" containing their IDs using CSMA. After that, CH receives member nodes IDs. Then it creates a TDMA schedule for every node which belongs to their respective cluster and broadcast it with last frame time slot (LF_SLOT), as shown in Fig 5. After following this procedure steady-state phase begins.

3.2 Steady-state phase

In this phase, data transmission begins; member nodes send their own data within their allocated TDMA slot to the CH. In EEIS each round consists of frames and each frame consists of time slots, as shown in Fig 3 and frame format of CH shown in Fig 4, where T_1, T_2, T_3 etc represent allocated time slot of $node_1$, $node_2$ and $node_3$ respectively. For sending data at their own time slots, member nodes check following conditions.

EEIS Protocol algorithm. ***In steady state phase***

At transmission side (nodes; belonging to CH)

if (no_of_packet > buffer_threshold) (1)

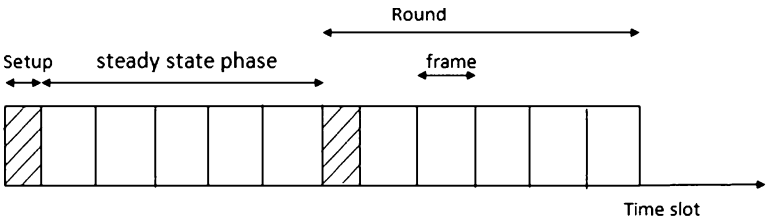


Fig. 3 Operations of EEIS protocol

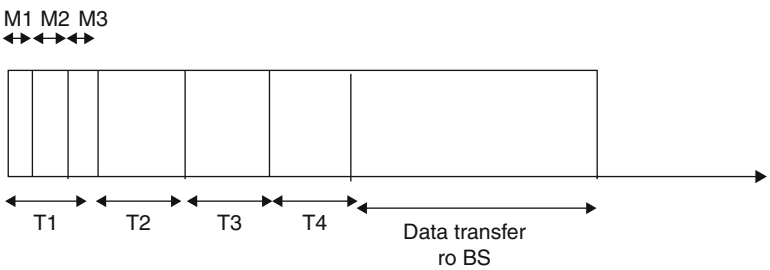


Fig. 4 Frame format of CH

{ On radio;
Send data to the CH;}
else if (current_frame == last_frame_slot) (2)
{Send data;}
else if (current_node_energy < node_threshold) (3)
{Send data;}
else
{Switch off radio;}
At receiving side (CH side)

CH waits for T_w time, if it does not find any data to accept then CH switches to the sleep mode for T_{all}-T_w time
else
{Switch on radio;}

In our methodology, if CH does not has data to accept variable load then it switches to the sleeping mode, but when next time slot begins, CH faces more energy loss as compared to the previous leach protocol, because of startup energy. And startup energy is always higher than the listening or idle energy. Therefore we have increased length of time slots. As shown in EEIS protocol algorithm.

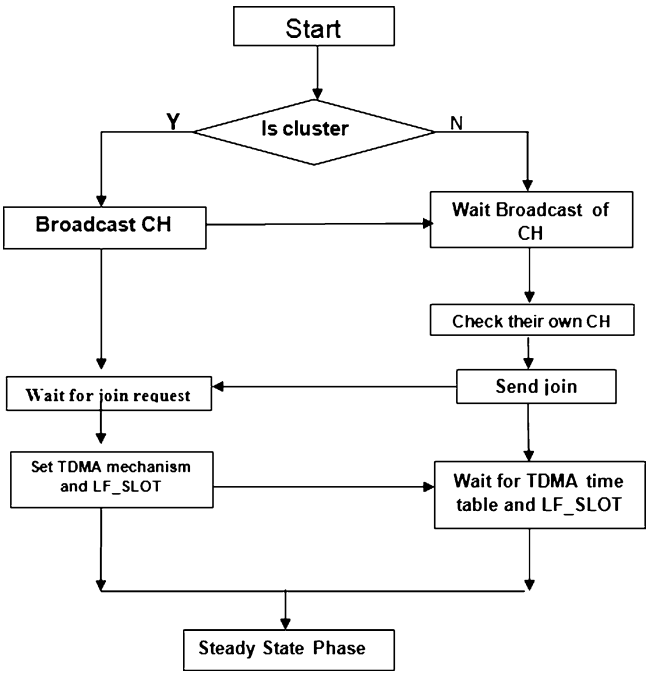


Fig. 5 Flow chart of Setup phase of EEIS protocol

At (1) if the buffer threshold is greater than or equals to number of packets in a member node, then it sends data to the CH, therefore at low duty cycle, it proceeds to save energy.

At (2) after following procedure (1), the node that has data, but number of packets available into the node is less than packet threshold. Then that node will not be able to send data to the CH. To overcome this drawback CH broadcasts last frame slot to the member nodes.

At (3) if a node has very less energy remaining, then it switches to the dead mode earlier, without sending data to the CH. Therefore every member nodes check their own current node energy. If current energy is smaller than node energy threshold, then node sends data to the CH.

As shown in Fig 6, every node waits for T_w time, if it senses data at T_w. Then CH receives data from the member nodes. If there is no data to accept in T_w time, then CH switches off its radio till next time slot. After T_{all}-T_w time CH switches to the startup mode and if CH has data to sense in T_w time slot. Then it receives data in their respective time slot. M1, M2, M3 represents messages in time slot T1. After accepting the data, again it waits for the data for T_w time slot. It follows same procedure. CH accepts data at owner's node time slot. If it satisfies the conditions given in steady state phase, after that all data sends to the BS using Code Division Multiple Access (CDMA). After finishing a round again CHs are created and it follows the same procedure.

4 Simulation Environment

We have simulated results on 100 sensor nodes, where BS placed at (58,145) location and compared energy of LEACH and EEIS protocol at variable traffic load. This section examines the performance of the proposed protocol using NS2.34 under Linux Operating System. The main purpose of this work is to analyze and develop an energy efficient clustering protocol and the factors influencing it include energy efficiency of nodes.

We are assuming the network is in a 100 x100 square area, all sensor nodes are uniformly randomly deployed in it.

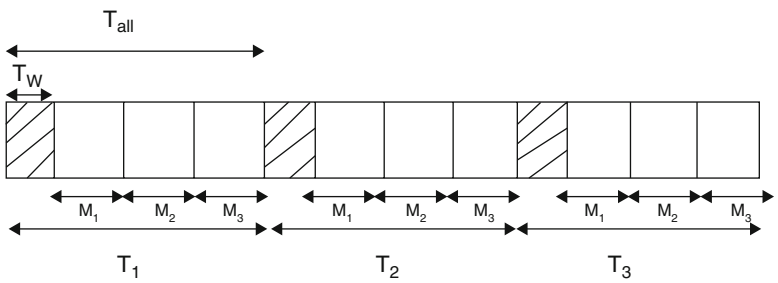


Fig. 6 Format of Time slots in a frame of CH

Table 1 Summary of the parameters used in the simulation experiments.

S. No.	Parameters For Our Simulation	
1	Header Size	25
2	Data Payload Size	500 Bytes
3	Beam forming energy	5e-9 J/bit
4	Radio Circuitry	50e-9 J/bit
5	Initial Energy	2 J
6	Channel Bandwidth	1Mbps
7	Energy Threshold	0.3 J
8	Receiving Energy	13.5 mW
9	Sleeping Energy	15 Micro W
10	Startup Energy	16 mW
11	Packet Threshold	3
12	Energy Threshold	0.3 J
13	Total number of Sensor nodes	100
14	Max packets in if queue	30

5 Results and Discussion

In this simulation, all the nodes begin with the same energy, 2 J. We have calculated total energy dissipation with respect to time slots.

The following result shows the improvement of energy with respect to average energy dissipated and number of nodes alive. We have compared energy efficiency of LEACH and proposed EEIS protocol in variable traffic load.

5.1 Average energy dissipated

This figure shows dissipation of energy per node over time in the network, as it various functions such as transmitting, receiving, sensing, aggregation of data etc.

As shown in Fig 7, the results are calculated in variable traffic load. In starting EEIS has more energy dissipation as compare to LEACH protocol. Because in starting, every node has data to send and all nodes store data and then send that data to their respective CH, if they satisfy the conditions as described in steady state phase of EEIS protocol, therefore there is more energy loss, because of additional startup energy. But when the idle slot increases average energy dissipation of EEIS protocol decreases and as a number of dead nodes are increases into a round then energy dissipation of all nodes are also decreases into EEIS protocol.

5.2 Total number of nodes alive

This metric indicates the life time of the network. More importantly, it provides an idea of the coverage area of the network over time.

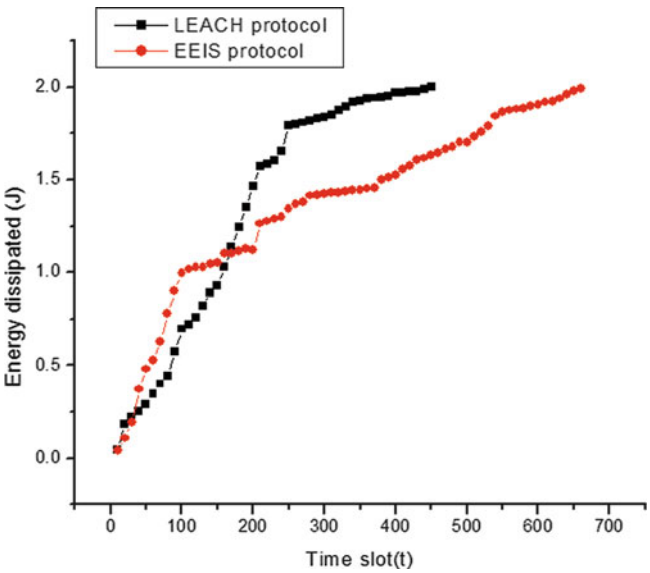


Fig. 7 Comparison of average energy dissipation

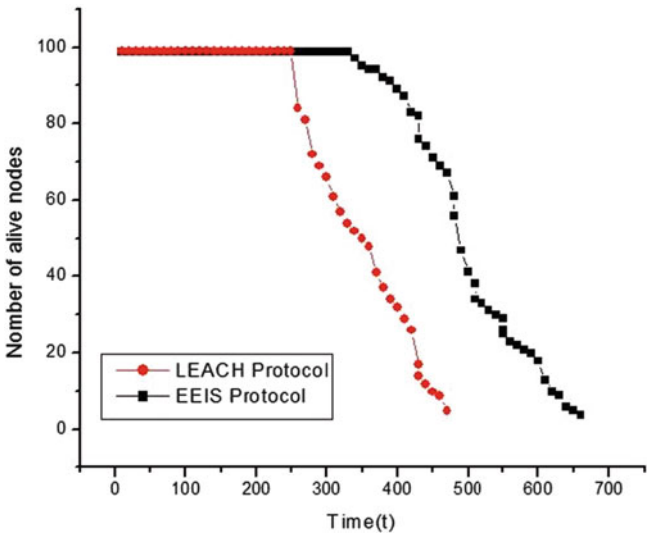


Fig. 8 Comparison of number of nodes alive

As shown in Fig 8, network life time of LEACH has decreased, because of low duty cycles. In LEACH, CH senses data continuously, whether a node is sending data or not. Therefore it proceeds to more energy loss at CH into the form of idle energy. But in EEIS protocol, CH switches to the sleeping mode, if it does not sense anything at T_w time. Since member node does not have anything to send or a node is dead in between a round then, CH switched to the sleeping mode till the next time slot, we have increased length of time slots on the basis of threshold value. It proceeds to the minimum energy loss into the form of startup energy. Therefore EEIS protocol has performed much energy efficiently than LEACH protocol.

6 Conclusion and Future Work

In this paper, we have introduced a new MAC layer protocol for WSN. EEIS protocol is well suited for that time when network is into the less contention. Simulation has shown that EEIS is more energy efficient then LEACH.

Work done in this paper is basically for low idle cycles, here we have provided packet threshold for the cluster members to the CH, Moreover, work can be extended to analyze relationship between number of packets and the distance from cluster members to the CH. In this protocol, at the last frame slot all the data are sent to the CH whether it is less then threshold or not. Therefore it is preceded to the less bandwidth utilization of data. We can use context aware data or provide priorities for sending important data to increase latency of data.

References

1. I. Rhee, A. Warrier, J. Min, Lisong, Xu,: DRAND: Distributed Randomized TDMA Scheduling for Wireless Ad Hoc Networks. In: IEEE Transactions on Mobile Computing, vol. 8, no. 10, pp. 1384-1396. (2009).
2. W. Ye, J. Heidemann, D. Estrin.: An Energy-Efficient MAC Protocol for Wireless Sensor Networks. In: Proc. IEEE INFOCOM, vol. 3, pp. 1567-1576. (2002).
3. I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci.: Wireless sensor networks: a survey. In: Computer Networks, vol. 38, no., pp. 393-422. (2002).
4. W.B. Heinzelman, A.P. Chandrakasan and H. Balakrishnan.: An Application-Specific Protocol Architecture for Wireless Microsensor Networks. In: IEEE Transactions on Wireless Communications, Vol. 1, No. 4, pp. 660-670. (2002).
5. M. Zhang, A. Babaei and P. Agrawal.: Networks Layer SCL: A Cross-Layer Protocol for Wireless Sensor Networks. In: 44th IEEE Southeastern Symposium on System Theory. Jacksonville, Florida (2012)
6. F. Xiangning and S. Yulin.: Improvement on LEACH Protocol of Wireless Sensor Network. In: International Conference on Sensor Technologies and Applications, pp. 260-264, (2007).
7. W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan.: Energy- Efficient Communication Protocol for Wireless Microsensor Networks. In: proceedings of the 33rd hawaii international conference on system sciences. Hawaii (2000).
8. A. A. Abbasi, M. Younis.: A survey on clustering algorithms for wireless sensor networks. In: Computer Communications. vol. 30, no.14-15, pp. 2826-2841, (2007).
9. Kyung Tae Kim, W. Choi, H. Youn.: CA-MAC: Context Adaptive MAC Protocol for Wireless Sensor Networks. In: Wireless Communications and Networking Conference, pp.1–6, (2009)
10. A. Manjeshwar, D. Agrawal.: TEEN: a protocol for enhanced efficiency in wireless sensor networks. In: Proceedings of the 15th International Parallel & Distributed Processing Symposium. pp. 189. (2001).

Identification of Redundant Node-Clusters for Improved Face Routing

Laxmi Gewali and Umang Amatya

1 Introduction

Sensor networks have been extensively used in several areas of science and engineering that include emergency response, remote sensing and monitoring of environmental data, law enforcement, manufacturing, robotics, object recognition, intelligent transportation, and unmanned vehicle systems. There is no centralized control in sensor network and most maintenance and computations are performed in a distributed manner. A sensor network is formed by wirelessly connecting sensor nodes which are distributed on a two dimensional surface or embedded in some equipment or gadgets. A sensor node is essentially a small electrical device containing (i) a small amount of memory, (ii) a low capacity processing unit, (iii) a radio communication component with range up to 300 meters, and (iv) a few sensing components for measuring physical quantities such as temperature, pressure, humidity, etc. Nodes within the transmission range can exchange information wirelessly. Nodes outside the transmission range can communicate by establishing a sequence of in-range intermediate nodes between them.

In this paper, we present an approach for identifying and/or removing unnecessary node clusters for improved face routing in sensor network by introducing the concept of 'floating chains' and 'external clusters'. The proposed approach is very appropriate if the locations of source node and destination node are confined within known regions.

2 Review of Planar Networks and Routing

We start with a brief review of planar networks and face routing algorithms that have been used for message communication in sensor networks. Planar networks used for face routing include Gabriel Graph (*GG*), relative neighborhood graph (*RNG*), restricted Delaunay graph (*RDG*), and Delaunay triangulation (*DT*).

2.1 Planar Networks

One of the extensively used planar graphs for communications in sensor network is the Gabriel Graph *GG* formed by a set of point sites $S = \{p_0, p_1, p_2, \ldots p_{n-1}\}$, which was first introduced by Gabriel and Sokal [5]. Formally, two point sites p_i and p_j are connected by an edge $e_{i,j}$ if the disk with diameter ending at p_i and p_j, denoted as $d(p_i, p_j)$, does not contain any other point site. *GG* can be constructed in O(n *log* n) time [8]. It can also be computed locally [11] and for this precise reason *GG* has been extensively used in sensor network. For applications in sensor networks, only those point sites p_i and p_j are considered for possible edge connection if the distance between them is less or equal to the wireless transmission range of sensor nodes.

The Relative Neighborhood Graph (RNG) proposed by Godfried Toussaint [10] is a way of defining a structure from a set of points that would match human perceptions of the shape of the set. This graph is also used to model proximity relations between nodes in two dimensional spaces and has found applications in sensor network. RNG can also be computed locally [2].

A planar graph that contains both Gabriel graph and Relative Neighborhood graph is the Delaunay triangulation [1, 9]. The Delaunay triangulation of a set of point sites $\{p_0, p_1, p_2, \ldots, p_{n-1}\}$ is the triangulation such that no circumscribing circle of any triangle in the triangulation contains any other point site. Many algorithms have been reported to compute Delaunay triangulation [1, 9]. The most popular algorithm for computing Delaunay triangulation is the sweep-line algorithm proposed by Fortune [4] which runs in $O(n \log n)$ time. In the context of routing in sensor networks, locally computable structures are highly suitable. Unfortunately, Delaunay triangulation cannot be computed locally.

L. Gewali (✉) • U. Amatya
University of Nevada, Las Vegas, USA
e-mail: laxmi.gewali@unlv.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_65,

However, a super-set of Delaunay triangulation called Restricted Delaunay Graph [6] can be computed locally. The way it is computed locally is as follows. Consider a set of point sites S = $\{p_0, p_1, p_2. \ldots p_{n-1}\}$ in the plane. Suppose each node p_i computes the Delaunay triangulation of p_i and its one hop neighbor. Let $T(p_i)$ denote the Delaunay triangulation of p_i and its one hop neighbors $N(p_i)$. The triangulation $T(p_i)$ is called the local Delaunay triangulation of p_i. The network obtained by the union of all $T(p_i)$'s is not necessarily planar and it may not be even a triangulation graph. Let G_u denote the union of all $T(p_i)$ for all p_i in S. The graph G_u is not necessarily planar and may not contain all Global Delaunay edges of S. A method of extracting a planar graph from G_u with 1-hop information exchange was proposed by Gao et al. [6].

2.2 Routing

Routing is the process of selecting a path in a network for sending information from the source node to the destination node along the network [2, 11]. Routes are constructed in sensor networks using appropriate planar graphs that include Gabriel graphs, relative neighborhood graphs, and restricted Delaunay graphs. In general, if a source node s wants to send a message to a destination node t which is outside the range of s, then s needs to send the message through a sequence of relay nodes. This sequence of relay nodes together with source and destination nodes define a route connecting s to t. The process of generating these routes is called *routing*. Some of the well-known route construction techniques are greedy forward, face routing and hybrid greedy *face routing* [2, 11]. Greedy forward routing is a very simple yet one of the most powerful routing algorithms. It constructs route locally in a sequence of steps. In greedy routing, all of the adjacent nodes that are within transmission range are first detected. The next node to forward the message is selected from among the adjacent nodes which are nearer to the target node than the current node. The node that is closest to the target node gets the message. This process is repeated until target is reached. Sometimes, a message gets stuck while using greedy forward routing algorithm. This occurs when the node itself becomes the shortest node rather than its adjacent nodes. Hence the message gets stuck at a local minimum.

A message delivery method which is guaranteed to work if there is some path connecting source node s to target node t is based on the traversal of the faces of the planar graph formed on the underlying sensor network. A path construction algorithm based on this approach is known as face routing [2, 11]. We can give a brief description of face routing algorithm by considering an example planar graph constructed on the sensor network. The planar graph could be either a Gabriel Graph (*GG*), or a Relative Neighborhood Graph (*RNG*) induced by the sensor nodes. In Figure 1, a Gabriel graph of eighteen nodes S = $\{s, p_1, p_2, \ldots, p_{15}, p_{16}, t\}$ is shown where there are nine faces that includes the outer unbounded face. Node s and node t are the source node and target node, respectively.

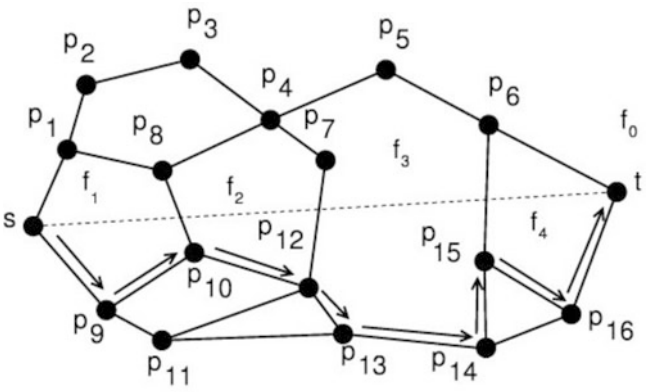


Fig. 1 Illustrating Face Routing

The source node s knows the position of itself, its 1-hop neighbors, and the position of the target node t. If t is within the transmission range of s, then the message is delivered directly. Otherwise, the algorithm construct the correct face f_c of the Gabriel graph such that (i) s is a node of f_c and (ii) f_c is intersected by the guiding line segment $e_g = (s, t)$. The algorithm determines the transition edge e_w of f_c. Transition edge e_w of f_c is the line segment of f_c that is intersected by the guiding segment e_g. In Figure 1, (p_8, p_{10}) is the transition edge. After constructing the current face f_c, the message traverses counterclockwise along the edges of f_c and stops at the transition edge. The message is then delivered to the other face (f_2 in Figure 1) incident on the transition edge. In the next iteration, one of the nodes of the transition edge becomes the source node and the other face incident on e_w becomes the current face f_c. This process of traversing the faces is continued until the target node t is discovered to deliver the message. In Figure 1, the constructed route is shown by directed segments.

3 Identification of Redundant Nodes

In this section, we describe the main contribution of the paper. We consider the problem of removing redundant or pseudo-redundant nodes from a set of given nodes. We call this process *node filtering*.

3.1 Node Filtering

Consider a set of nodes S = $\{p_0, p_1, p_2. \ldots p_{n-1}\}$ used for face routing in a sensor network. Two nodes close to each other are called *equivalent* if their transmission ranges

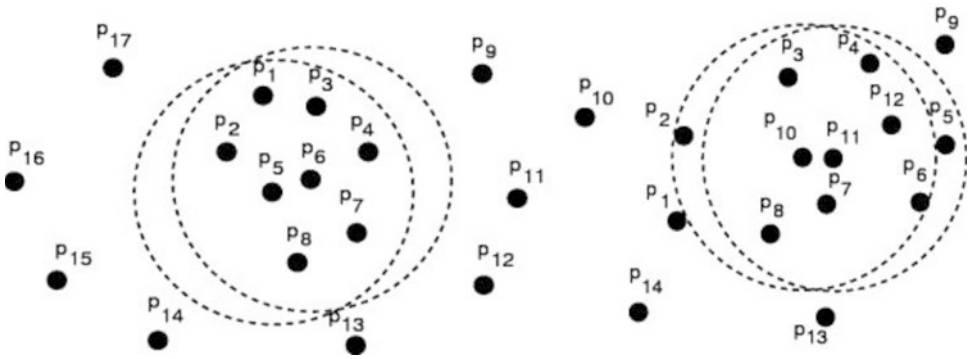


Fig. 2 Illustrating Equivalent Nodes

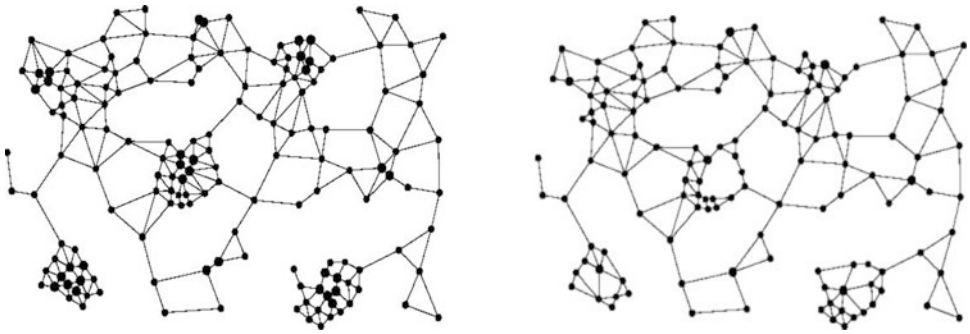


Fig. 3 Illustrating Node Filtering

cover the same sub-set of nodes. Recall that all nodes are assumed to have an identical transmission range which is taken, without loss of generality, as 1. We can illustrate the notion of equivalent nodes with a specific example. In left side of Figure 2, the transmission disks of two nodes p_5 and p_6 are shown with dashed circles. This shows that nodes p_5 and p_6 cover the identical set of nodes p_1, p_2, p_3, p_4, p_7 and p_8.

In right side of Figure 2, two nodes p_{10} and p_{11} are shown which are not equivalent even though they are very close to each other. There can be many nodes equivalent to each other in some rare distributions that contain clustered nodes in some pocket regions.

Definition 3.1: (*Compressed Gabriel Graph*) Consider a Gabriel Graph *G(V,E)* of a set of sensor nodes. Let C_1, C_2, . . .,C_k be the set of equivalent nodes in *G(V,E)*. The nodes in V not in the equivalent sets are referred to as *background nodes* and the set of these nodes is denoted by V_B. The set of nodes obtained by adding to V_B exactly one member from each equivalent set is the compressed set of nodes, V_C. The resulting Gabriel graph of V_C, denoted by $G_C(V_C, E_C)$, is the compressed Gabriel graph. Figure 3 shows the original Gabriel graph and its compressed version for indicated transmission range.

It is very interesting to look for the preservation of connectivity when the network is compressed. The following theorem settles the connectivity issue.

Theorem 1 *If background nodes v_i and v_k are connected in Gabriel graph G(V,E), then they are also connected in the compressed Gabriel graph $G_C(V_C, E_C)$.*

Proof (Omitted).

3.2 Identification of Floating Chains and Removable Clusters

In this subsection we introduce the notion of floating chains and removable 'external clusters'. These notions can be used to identify node clusters that consist of linear chains, trees, and entire 2-connected components. Such clusters can be potentially removed or deactivated from the network without affecting the connectivity between source and target nodes. The set of nodes that can be deactivated or removed depends on the relative positioning of the source and target nodes.

Consider a sensor network formation as shown in Figure 4, where source node s and target node t are as indicated. This network has four chains and trees labeled as ch_1, ch_2, ch_3, and ch_4. It has 3 bounded faces F_1, F_2, F_3 *and* 1 unbounded face F_0. For this network, with indicated positioning of source node and target node, removal of nodes in ch_3, and ch_4 does not affect the delivery of message from source node to target node. The only chains whose presences affect the connectivity between source and target nodes are ch_1 and ch_2.

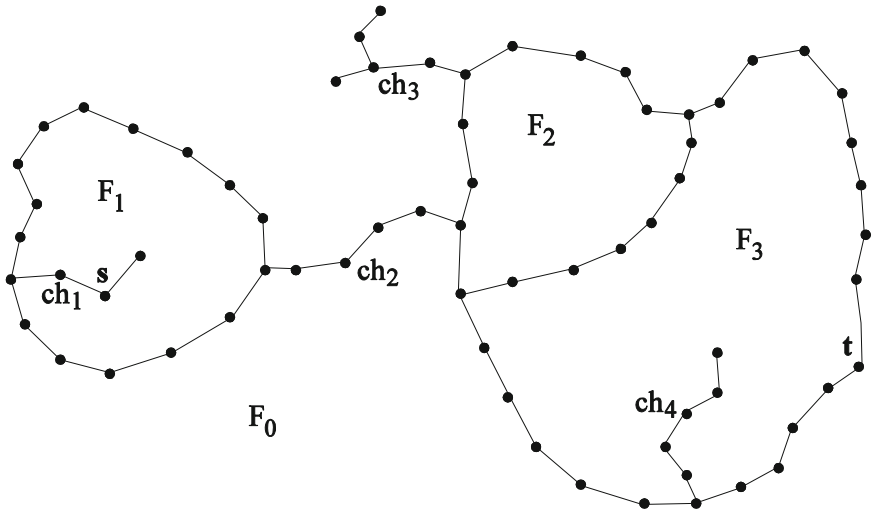


Fig. 4 Characterizing Solo-Faced Chains

Definition 3.1 An edge e_1 is called *solo-faced* if e_1 is adjacent only to one face. All five edges in chain ch_4 are solo-faced edges. A maximal sequence of solo-faced edges is called a *solo-faced chain*. All four chains labeled ch_1, ch_2, ch_3, ch_4 in Figure 4 are solo-faced chains.

A solo-faced chain that can be removed without affecting the connectivity between source and target nodes is characterized in the following observation.

Observation 3.1 If a solo-faced chain ch_i has a node of degree 1 and does not contain source or target node then such a chain can be removed without affecting the connectivity between the source and target nodes. Such chains are referred to as *floating chains* with respect to a given pair of source and target nodes. In Figure 4, chains ch_3 and ch_4 are floating chains. Furthermore, some edges (and not all) in a solo-faced chain containing source and/or target node may be removable. Edge ch_1 containing source node s (Figure 4) is such that one of its edges is removable.

To describe an algorithm for removing floating chains from the network we need to adopt the terms and notations defined in representing a planar graph by doubly connected edge list data structure (DCEL). Interested readers can see the detail of DCEL data structure in [1]. DCEL stores records for edges, faces, and vertices with appropriate attributes. Each edge is represented as a pair of half-edges which are called *twin* of each other. Each half-edge has four attributes: (i) *twin half-edge*, (ii) *previous half-edge*, (iii) *next half-edge*, (iv) *incident node*, and (v) *incident face* with obvious meaning. The main attribute of a face is its *bounding half edge*. Finally the main attribute for a node is its *incidence half edge*. In term of these notations, an algorithm for removing floating chains is listed as Algorithm 1.

In some situations, the connectivity of the network and the relative positions of source and target nodes are such that a whole bi-connected cluster can be removed. This is illustrated in Figure 5. For the purpose of continuity we recall the definition of bi-connected component of a network used in standard algorithm literature. A graph $G(V,E)$ is called *bi-connected* if there are two distinct paths between any pair of vertices in $G(V,E)$. In a connected sensor network without floating chains, there could be several bi-connected components as shown in Figure 5, where there are five bi-connected components labeled as A, B, C, and D. Furthermore, a solo-faced chain connecting two bi-connected components is called *bridge chain*.

One can imagine a 'cluster tree' where nodes are bi-connected components and bridge chains are links. In such a tree, leaf nodes are referred to as *external clusters*. In Figure 5, B and D are external clusters. This observation immediately leads to the following lemma.

Lemma 3.1 In a connected sensor network, external bi-connected components can be removed or deactivated without compromising the connectivity between source node s and target node t.

Algorithm 1 Removing a Floating Chains

```
RemoveFloatingChain(Vertex v_c, DCEL D_1) {
   v_c is a degree 1 node
   HalfEdge e_1, e_2;
   Vertex v_1;
   while v_c is not in {s, t} do
         e_1 = v_c.getIncHalfEdge();
         e_2 = e_1.getTwin();
         v_1 = e_2.getStartNode();
         remove v_c, e_1, and e_2 from D_1;
    end while
} // end RemoveFloatingChain
```

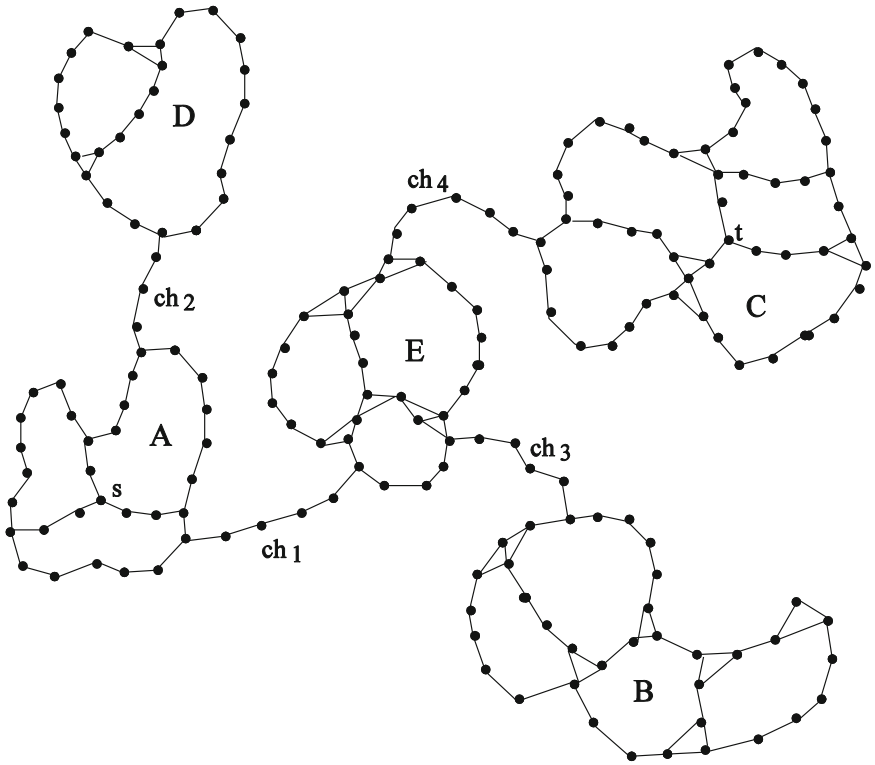


Fig. 5 Characterizing External Bi-Connected Components

A formal algorithm for removing external clusters can be written by modifying the standard bi-connected component separation algorithm [3]. However, for sensor network formed as Gabriel graph (*GG*), a simpler approach can be used to remove external clusters. One can represent the *GG* as a doubly connected edge list (*DCEL*) data structure [1] and follow the half edges to identify floating chains. The bridge edges can be identified by following the boundary of unbounded face. A route *R* connecting source node to target node can be inspected to identify external bi-connected clusters. The details are omitted.

4 Discussion

We reviewed planar graphs used for modeling sensor networks. We reviewed the techniques of face routing for message delivery from a source node to a target node in sensor network. We proposed techniques for filtering unnecessary nodes in a sensor network. In particular we showed that floating nodes can be identified by simply following the half edges bounding the faces. The external clusters can be identified and removed by following the half edges of the unbounded face. We articulated how the removal of redundant nodes, floating chains, and external clusters do not affect the connectivity between source and target nodes. Removal and/or deactivation of unnecessary nodes can potentially increase the life of sensor network. Message packets delivered to the target node could potentially trace the outer boundary of external cluster and traverse the corresponding bridge edge more than once. For such situations the removal of external cluster increases the battery life of some nodes. Furthermore, the proposed technique improves the performance of face routing algorithm by constructing shorter routes.

Bibliography

1. de Berg Mark, Van Kreveld Mark, Overmars Mark, and Schwarzkopf Otfried (1997) Computational Geometry: Algorithms and Applications, Springer. New York.
2. Bose P, Morin P, Stojmenovic I, Urrutia J (2001) Routing with guaranteed delivery in ad hoc network. Wireless networks. 609-616.
3. Cormen T, Leiserson CE, Rivest RL, Stein C (2009) Introduction to algorithms. MIT Press. Massachusetts.
4. Fortune S (1986) A sweep line algorithm for Voronoi diagram, Algorithmica. 2:153-174.
5. Gabriel KR, Sokal RR (1969) A new statistical approach to geographic variation analysis. Systematic zoology 18(3): 259-270.
6. Gao J, Guibas LJ, Hershberger L, Zhang L, Zhu (2001) A geometric spanner for routing in mobile networks. IEEE journal on selected areas in communications. 23:174-185.
7. Gewali L, Hada U (2013) Node filtering and face routing in sensor network. Proceedings of ITNG 2013: 686-691.
8. Matula DW, Sokal R (1980) Properties of Gabriel Graphs Relevant to Geographic Variation Research and the Clustering of Points in the Plane. Geographical analysis 12(3): 205-222
9. O'Rourke J (1994) Computational geometry in C. Cambridge University Press, New York.
10. Toussaint G (1980). The relative neighborhood graph of a finite planar set. Pattern recognition 12: 261-268.
11. Zhao F, Guibas L (2004) Wireless sensor network. Morgan Kaufmann. New York.

Distributed Processing Applications for UAV/drones: A Survey

Grzegorz Chmaj and Henry Selvaraj

1 Introduction

Distributed processing systems contain multiple devices (called *nodes*) equipped with processing capabilities (based on general purpose processor, ARM chips or any other processing unit) and communication unit providing the network communication. Nodes join the designated system, become its members and contribute with their processing power and any other capabilities they possibly might have – such as peripherals attached. The resources in the distributed system are limited [1]. All member nodes along with control element, communicating over common network layer – constitute the distributed processing system. Such system can process a common task that is internally split into sub-tasks (called *blocks*) that are further processed by different nodes – and the result is yield by the control unit. As the task is inputted to the control unit, and result appears there as well, the whole system can be perceived as a virtual machine – having the computation and communication infrastructure inside. Tasks processed by distributed processing systems can be of various types. Popular distributed processing systems such as grids [2] and public computation systems [3] usually perform the computing (such as signal processing, image analysis, etc.) and the result is the processing result, or some answer (e.g. "object found in the processed image"). The evolution of distributed processing systems brought a new task types – that produce the results that immediately affect the distributed processing system itself, modifying its operating parameters.

Unmanned Aerial Vehicles – UAVs (also called *drones*) are the flying objects that are not equipped with any pilot on board. They can be either controlled from the ground station, or by autonomous onboard control algorithms. Recently the small-sized drones are gaining very big attention both in research and commercial world. The researchers focus on optimization of onboard operational algorithms, swarming, navigation systems and many others, while industry companies try to employ drones for commercial applications, such as surveillance [4] or package deliveries [5]. The strong interest about UAVs and UAS (Unmanned Aerial Systems) is also expressed by the governments considered as long-term future systems with high potential [6]. Multiple drones communicating with each other and being commonly controlled are called *swarms* or *teams*. UAVs participating in such groups are often equipped with processing capabilities, various sensors (such as cameras) and connect to the common communication network. This way, such drone swarms fulfill the principles of distributed processing system – in which drones are the nodes, and the system fully manages by itself using its internal algorithms and mechanisms.

2 The goal of the survey

This paper describes the applications of distributed processing systems for UAV/drones swarms and teams. We focus on computer engineering aspects (and related), the implementation of collaboration mechanisms between nodes of distributed systems, communication, and principles of distributed processing that are used in analyzed distributed processing systems built over groups of UAVs. However, system participants are not limited to UAVs, structures including both UAV and UGV (Unmanned Ground Vehicles) participants are included in the survey too. Systems analyzed in this paper do not limit UAV size – we include systems using large drones, medium-sized UAVs, mini helicopters, micro-UAVs and others. For the purpose of this survey, we classify the distributed processing applications into following categories: 1) general purpose

G. Chmaj
University of Nevada, Las Vegas, NV, USA
e-mail: grzegorz.chmaj@unlv.edu

H. Selvaraj
Department of Electrical and Computer Engineering,
University of Nevada, Las Vegas, NV, USA
e-mail: henry.selvaraj@unlv.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_66,

distributed processing applications, 2) object detection, 3) tracking, 4) surveillance, 5) data collection, 6) path planning, 7) navigation, 8) collision avoidance, 9) coordination, 10) environmental monitoring.

3 Distributed processing applications for UAVs

3.1 General purpose distributed processing applications

Many solutions dedicated to distributed processing in UAV systems are meant to provide under-the-hood functionalities – where other applications are built on top of them. Such architecture designated to manage the resources in the team of autonomous vehicles (aerial UAVs and ground UGVs) was presented in [7]. DFRA – Distributed Field Robot Architecture uses object-oriented approach and implements the interaction of software objects / agents through a decentralized group of vehicles. Each resource present at the vehicle member of the team can be discovered by any other vehicle or the human operator. DFRA was implemented using Java and Jini middleware and applied to the team containing both UAV and UGV units. More complex solution for the software platform was presented in [8]. Authors propose the modular infrastructure aimed to coordinate distributed control, vision control and communications. The communications layer is implemented using control flooding (broadcasting information) in which incremental data passing is used (UAV receiving the message from broadcaster, receives not only data related with broadcaster, but also sender's knowledge about other UAVs). Tasks are distributed among vehicles and assigned based on distributed collaboration of UAVs.

3.2 Object detection

The most popular application of UAV swarms implementing distributed processing is the object detection – based on the data gathered by data sources such as cameras and radars. This application, also called ATR (automatic target recognition), strongly employs the image processing and signal processing algorithms.

The use of Bayesian search (and also localization) in a distributed manner was presented by the authors of [9]. The search and localization algorithm uses the process of filtering multiple observations coming from the pointed-down cameras attached to vehicles. In the target localization process (single objects are considered), the probability of a true detection of an object is a function of the image resolution (at the target's location), and the visibility of the target. The multiple-target search was presented in [10] – where, similarly, the team of UAVs in which each one is equipped with the camera is employed. The monitoring area is divided into regions, and the probability of target existence in each of the regions is calculated. The distributed UAV platform uses the sharing of measurement information, information updates due to environmental factors and changes, keeping in mind the limited communication and sensing capabilities. Based on the existence probability and the shared data fusion, the distributed probability map updating model is implemented. The important aspect regarding the distributed processing algorithms is that the UAVs communicate only the neighboring vehicles in their range – i.e. the overlay communications network is not used in this case.

3.3 Tracking

Tracking is the application related to object detection – often times the mission goal states, that once the target is detected, then it is being tracked by the same UAV team. The object detection and tracking can be implemented as one application. The high level definition of the distributed tracking system was given in [11]. The description contains the technical details of such elements as: vehicle, object tracker, flock guidance and geolocation – and the relations between them. The implementation details and parameters are also provided, along with possible algorithms. The example of the search & tracking application was presented in [12]. The proposed system can switch between search & tracking modes and still use the information that was gathered during past operation. It deals with the scenario where the tracked target is lost – then the search is performed again and the recursive Bayesian filtering is used. Vision measurements were used in [13] and [14] to build the collaborative distributed tracking system. In [13] the information gathered by multiple UAVs is fused along with the goal of minimizing the data transmitted between vehicles. Cameras are directed to POI (Point Of Interest) with the necessary adjustment according to the vehicle/target movement. The geolocation of POI is calculated by vehicle, also including the vehicle/target movement, however due to the collaborative nature of the system, the dynamics of POI is recalculated so it is not a function of any UAV state.

3.4 Surveillance

Area surveillance operated by teams of UAVs is considered as one most perspective distributed processing applications, remaining one of most controversial at the same time. The use of dynamic task reconfiguration (also mentioned in section 3.1) was employed to perform the surveillance task through the team of heterogeneous UAVs equipped with various sensors. Multiple tasks require significant computing

power so maximizing the use of processing resources brings the valuable profit. Another approach to the distributed surveillance application – perimeter surveillance – is described in [15]. The proposed approach of the cooperative perimeter surveillance employs the team of UAVs (each having limited connectivity capability) and allows insertion/deletion of team members. Even though the distributed processing fashion is used for the surveillance process, the results are communicated to the ground station that constitutes the central element of the system. Vehicles communicate only with their neighbors met at the same location – so depending on the relations of positions of UAVs there are many communication schemes – what adds the another aspect to the distributed processing problem. The algorithm implementing the distributed processing perimeter surveillance includes such parameters as: perimeter length, number of UAVs to the 'left' of given vehicle, and to the 'right' of a given vehicle. Another aspect of the surveillance is the prioritization of area sections taken into the consideration.

3.5 Data collection

Wireless sensor networks (WSN) are the structures of multiple data sources that are spread over the area for the data gathering purposes – and captured information is sent to the central unit (such as computing center). They usually register environmental quantities such as temperature, air pressure, wind, etc. – but are not limited to. The technological advance in WSNs first allowed the bidirectional sensor-central unit communication – sensors started to be capable not to only send gathered data to the central unit, but also to receive various information, including the control messages. Second, it allowed to consider UAVs as the autonomous sensors – or more to say – the units handling multiple sensors onboard, but still filling the WSN unit paradigm. The automated sensing platform employing multiple UAVs was presented in [16]. The data gathered is analyzed at UAVs, to process the noise and other uncertainties introduced by the airborne operation of UAV. Then the data fusion is performed, also in cooperative distributed manner at UAVs – unlike other approaches in which it takes place in the central unit (usually ground-based), so the result is provided by the team of UAVs autonomously. The problem of decentralized data fusion is also deliberated in [17], where authors put special attention at decentralization of the presented sensing network. Therefore, no common communication unit is used, fusion is done in a distributed manner at UAVs and sensors know only about neighboring network connections. Decentralization was implemented using Kalman filter that requires the fully connected network (i.e. each unit in the network is connected to every other one), that in practice can be implemented only for small sized systems (or the concept of overlay network can be used). Authors used the tree algorithm and showed that it will be sufficient for this application, regardless the full-connectivity requirement.

3.6 Path planning

The swarms of UAVs operate in diversified environment – that can contain various ground and aerial obstacles that need to be taken into the consideration in the path planning process. Multiple obstacles on a way of multiple collaborative UAVs, each equipped with sensor(s) form the distributed processing problem. The generation of optimal path is usually not possible in the feasible time, thus the path planning applications are trying to produce the solution as close to optimal as possible, but in a very short time. Authors of [18] presented the algorithm that uses the particle swarm optimization and delivers the path of quality dependent on the computation time. The UAVs that are the members of the distributed processing structure concerned, are considered to have the limited-sensor and limited-communication capabilities. The outcome of the UAV-based distributed system is a set of paths, each assigned to single UAV – so the swarm can reach the geographical goal without collision with ground and aerial obstacles, and the other UAVs as well. Each UAV computes the path for itself, and if the communication distance allows – it shares its plan with others. Vehicles sharing information during the flight, update their paths if necessary (e.g. because of predicted collision with other UAV). In [19], the team of UAVs forms the monitoring structure that supplies the information for path planning adaptive algorithms. In this solution, also the exceptional events are supported, such as: vehicle leaves or joins the network, sensor enters/leaves the network. UAVs communicate with each other, so the path is created in the distributed collaborative manner.

3.7 Navigation

The path planning is closely related to the application of navigation – both of these involve directing the UAV in some direction or position. Usually, the path planning is considered as flying from one position to another, while navigation is considered as determining the actual position of the vehicle(s). Nowadays GPS satellites provide a way to calculate the position of the GPS signal receiver, however this calculation might be not accurate enough, or the GPS receiver might be impossible to place on the UAV (e.g. due to energy consumption). Therefore various variants of navigation are still in consideration. The distributed processing navigation framework was presented in [20] where the data

from image sensors is the base for applied algorithms. A swarm of vehicles monitors the environment and takes multiple targets in the consideration – the image data is processed in the distributed manner and each target is labeled with a signature (containing various elements, such as position on the image focal plane, time, and optionally the position obtained from GPS). Data from multiple vehicles is fused to get the greater picture of the area. The approach presented in [21] uses the multiple ground stations that cooperate with UAVs and together form the distributed processing system (unlike the [20], where all the processing is done airborne). Each vehicle is associated with a ground station, and gathers the data from sensor that it is equipped with – and the images (and other sensor data) are filtered and fused by the ground station (using Extended Kalman Filter) what produces the detailed information about vehicles' motion. Next, the information about all UAVs is fused and the state of each vehicle is estimated.

3.8 Collision avoidance

The algorithms for the collision avoidance must be implemented in the cooperative swarm of UAVs, as the vehicles operate in close vicinity of each other and some of vehicles might need to change the spatial position in the swarm formation. Collision avoidance is often used interchangeably with obstacle avoidance. For the purpose of this paper we consider collision avoidance as avoiding the physical contact between UAVs in one swarm. The comprehensive collision avoidance study, including various types of vehicles was presented in [22]: two cases for AGVs (Autonomous Ground Vehicles) and one for UAV. The distributed application for UAVs considers that vehicles share their information between each other about location and anticipated direction of movement. Communication algorithm keeps these information up to date, so it's ready for use in critical moments. The application works in a distributed manner, so each UAV is controlled by separate agent that implements several avoidance algorithms, combining cooperative and non-cooperative methods. The CSM (Collision Solver Manager) unit is defined to manage both collision detection and collision solving – it manages the results from individual collision solvers to produce the final avoidance result. There are many solutions presented in the literature that address the collision avoidance aspect – but in non-distributed manner. The [23] uses lateral acceleration to avoid detected collisions, multiple collisions are also supported. System presented in [24] is fully decentralized, but assumes that the data about other vehicles is known to each UAV (obtained from radar, or other source of information) – what can be problematic in real systems. Small amount of work in this topic exposes the potential of distributed processing applications dedicated to mutual collision avoidance.

3.9 Coordination

While many of applications mentioned in this paper operate on high level and realize the actual mission goal, it's impossible to disregard the coordination application that serves on the basic level and is indirectly used by other, higher level applications. The coordination mechanisms may provide the following: team formation, attitude alignment, rendezvous problem, coordinated decision making, flocking, coupled oscillators, position synchronization [25], [26]. The decentralized optimization of UAV swarm control was presented in [27], showing the methods to coordinate vehicles with multiple decision markers. The communication for solution passing in the system is also a part of the model and includes the uncertainty of messages delivery. Consensus problems in multi-node systems were discussed in [25] and lists the shared knowledge as one of main problems in for distributed control. The distributed processing system must adjust to the changes – consensus is the convergence to a common value. Authors list the following aspects as the key ones to address in consensus problems: finding the equilibrium state to which the consensus protocol converge, information uncertainty, communication delays and others.

3.10 Environmental monitoring

The environmental monitoring is the wide research field for single UAV solutions, where the monitoring of the environment is realized just by one vehicle [28]. Later the approach was extended into multi-vehicle systems. The partnership of University of Colorado, University of Alaska, MIT and University of Oklahoma proposed the Center of Collaborative Mobile Sensing Systems to design the distributed system for environmental monitoring [29]. This work describes the design aspects of environment monitoring with distributed systems, focusing on: wildfire (fire-fighting operations, increased capabilities of wildfire modeling and prediction); polar (heterogeneous mixes of UAVs equipped with sensor used for data acquisition); storm (in-situ data acquisition in severe storms – in altitudes from the ground to the cloud) [29]. The following areas were identified as the critical to implement environmental monitoring systems: assured ad hoc communication; collaborative guidance, navigation and control; data fusion and visualization; regulatory issues and societal response; robust airborne platforms; sensor integration.

Fire detection and monitoring. The need of monitoring of frequent delivery of high-quality information was addressed in [30] and authors propose the team of LASE (low attitude short endurance) to cooperatively monitor the propagation of forest fires and fire perimeter. Each vehicle can gather that much of information to navigate autonomously, it also has IR camera and has limited communication range (can communicate only with UAVs within the range of both). The described approach uses the centralized way, in which UAVs send the data to the ground station, but also the vehicles share the information while rendezvous meeting during perimeter search. The solution for both fire detection and extinguishing was presented in [31]. It contains the monitoring and planning station where all the control is concentrated, but also the image processing algorithms. The communication layer implements the BBCS (Black Board Communication Systems) that implements the distributed shared memory (blackboard) in which each network node has its own local copy of the blackboard portion it is accessing. The task allocation is done either by manual allocation or distributed task allocation module that manages the allocation autonomously in the group of robots.

Pollution detection & Water management. Unlike other objects or phenomenon, the pollution can't be detected in a distance – therefore the UAV has to reach the immediate vicinity of the pollution to detect. The system presented in [32] uses UAVs equipped with attractive beacon that activates when the sensor is in contact with chemical pollution. Once it happens, other vehicles are informed and follow the attractive beacon to reach the pollution area. Next, the swarm of UAVs encloses the contaminated area and follow the cloud.

The work presented in [33] presents the use of distributed team of UAVs applied to water management and distributed irrigation control. Vehicles measure the electromagnetic radiation and are equipped with imagers operating at different wavelength. System operates in the distributed manner with the ground-based station that serves as the data aggregator and the image processing unit.

4 Conclusion

Many collaborative teams of UAVs run the applications that are using the principles of distributed processing systems. Depending on the specific requirements, limitations and goals – various approaches are implemented. The research mentioned in this survey shows, that cooperative swarms of UAVs are demonstrating great usefulness and have much better properties and operational parameters than applications running on single UAVs. The work described above often concentrates on the mission-specific aspects, leaving the distributed processing 'under-the-hood' layer as the secondary area, sometimes full of assumptions and issues to solve. Therefore we consider the work on the distributed processing layer applicable to UAV teams as the promising future research field.

References

1. Zydek, D., Chmaj, G., Chiu, S.: Modeling Computational Limitations in H-Phy and Overlay-NoC Architectures, The Journal of Supercomputing. (2013) doi: 10.1007/s11227-013-0932-9
2. Baker, M., Buyya, R., Laforenza, D.: Grids and Grid technologies for wide-area distributed computing, Software: Practice and Experience, vol. 32, Issue 15, pp. 1437–1466. (2002)
3. Chmaj, G., Walkowiak, K.: Decision Strategies for a P2P Computing System, Journal of Universal Computer Science, vol. 18, no. 5, pp. 599–622. (2012), doi: 10.3217/jucs-018-05-0599
4. Wada Akihisa, Yamashita Toshiaki, Maruyama Masaaki, Arai Takanari, Adachi Hideo, Tsuji Hirokazu,: A Surveillance System Using Small Unmanned Aerial Vehicle (UAV) Related Technologies, Special Issue on Solving Social Issues Through Business Activities, NEC Technical Journal, Vol.8, No.1. (2013)
5. Mayerowitz, S.: Amazon.com sees delivery drones as future, Phys.org, The Associated Press, 2013
6. Winnefeld, J.A., Kendall, F.: Unmanned Systems Integrated Roadmap FY2013-2038, Department of Defense, 14-S-0553, DIANE Publishing Company. (2014)
7. Long, M., Gage, A., Murphy, R., Valavanis, K.: Application of the Distributed Field Robot Architecture to a Simulated Demining Task, Proceedings of the 2005 IEEE International Conference on Robotics and Automation, pp. 3193-3200 (2005), doi: 10.1109/ROBOT.2005.1570602
8. Ryan, A., Xiao, X., Rathinam, S., Tisdale, J., Zennaro, M., Caveney, D., Sengupta, R., Hedrick, J.K.: A Modular Software Infrastructure for Distributed Control of Collaborating UAVs, Proceedings of the AIAA Conference on Guidance, Navigation and Control, (2006)
9. Tisdale, J., Ryan, A., Kim, Z., Tornqvist, D., Hedrick J.K.: A multiple UAV system for vision-based search and localization, American Control Conference - ACC, pp. 1985-1990 (2008), doi:10.1109/ACC.2008.4586784
10. Hu, J., Xie, L., Xu, J.: Vision-Based Multi-agent Cooperative Target Search, 12th International Conference on Control Automation Robotics & Vision (ICARCV), pp. 895-900, 2012, doi:10.1109/ICARCV.2012.6485276
11. Wheeler, M., Schrick, B., Whitacre, W., Campbell, M., Rysdyk, R., Wise, R.: Cooperative Tracking Of Moving Targets By A Team Of Autonomous UAVs, 25th Digital Avionics Systems Conference, IEEE/AIAA, pp. 1-9 (2006), doi:10.1109/DASC.2006.313769
12. Furukawa T., Bourgault, F., Lavis, B., Durrant-Whyte, H.F.: Recursive Bayesian Search-and-Tracking Using Coordinated UAVs for Lost Targets, Proceedings 2006 IEEE International Conference on Robotics and Automation, pp. 2521-2526 (2006), doi:10.1109/ROBOT.2006.1642081
13. Campbell, M.E., Whitacre, W.W.: Cooperative Tracking Using Vision Measurements on SeaScan UAVs, IEEE Trans. Contr. Sys. Techn. vol. 15, no. 4, pp. 613-626, (2007)
14. Bethke, B., Valenti, M., How, J.: Cooperative Vision Based Estimation and Tracking Using Multiple UAVs, Advances in Cooperative Control and Optimization Lecture Notes in Control and Information Sciences, vol. 369, pp. 179-189 (2007)
15. Kingston, D., Holt, R., Beardy, R., McLain, T., Casbeer, D.: Decentralized Perimeter Surveillance Using a Team of UAVs, IEEE Transactions on Robotics, vol.24, no.6, p.1394-1404 (2008)

16. Teacy, W.T.L., Nie, J., McClean, S., Parr, G., Hailes, S., Julier, S., Trigoni, N., Cameron, S., Collaborative Sensing by Unmanned Aerial Vehicles, Proceedings of the 3rd International Workshop on Agent Technology for Sensor Networks
17. Durrant-Whyte, H., Stevens, M.: Data Fusion in Decentralised Sensing Networks, Technical report, Australian Centre for Field Robotics, The University of Sydney, 2001
18. Sujit, P.B., Beard, R.: Multiple UAV Path Planning using Anytime Algorithms, In proceeding of: American Control Conference, (2009)
19. Cunningham, C.T., Roberts, R.S.: An Adaptive Path Planning Algorithm for Cooperating Unmanned Air Vehicles, IEEE International Conference on Robotics and Automation, vol. 4, pp. 3981-3986 (2001)
20. Deming, R.W., Perlovsky, L.I.: Concurrent multi-target localization, data association, and navigation for a swarm of flying sensors, Journal Information Fusion, vol. 8, Issue 3, pp. 316-330 (2007), 10.1016/j.inffus.2005.11.001
21. Rigatos, G.G.: Distributed filtering over sensor networks for autonomous navigation of UAVs, Intelligent Service Robotics, vol. 5, Issue 3, pp. 179-198 (2012)
22. Vrba, P., Mařík, V., Přeučil, L., Kulich, M., Šišlák, D.: Collision Avoidance Algorithms: Multi-agent Approach, Holonic and Multi-Agent Systems for Manufacturing Lecture Notes in Computer Science, vol. 4659, pp. 348-360 (2007)
23. Manathara, J.G., Ghose, D.: Reactive collision avoidance of multiple realistic UAVs, Aircraft Engineering and Aerospace Technology, vol. 83, no. 6, pp. 388-396 (2011), doi:10.1108/00022661111173261
24. Lalish, E., Morgansen, K.A.: Decentralized Reactive Collision Avoidance for Multivehicle Systems, 47th IEEE Conference on Decision and Control, pp. 1218-1224 (2008), doi:10.1109/CDC.2008.4738894
25. Ren, W., Beard, R.W., Atkins, E.M.: A Survey of Consensus Problems in Multi-agent Coordination, Proceedings of the 2005 American Control Conference, vol. 3, pp. 1859-1864 (2005), doi:10.1109/ACC.2005.1470239
26. Ren, W., Cao, Y.: Overview of Recent Research in Distributed Multi-agent Coordination, Distributed Coordination of Multi-agent Networks, pp. 23-41 (2011), doi:10.1007/978-0-85729-169-1_2
27. Inalhan, G., Stipanovid D., Tomlin, C.J.: Decentralized Optimization, with Application to Multiple Aircraft Coordination, Automatica, vol. 40, no. 8, pp.1285-1296 (2004)
28. Harris, A., Sluss Jr., J.J., Refai, H.H.: Alignment And Tracking Of A Free-Space Optical Communications Link To A UAV, The 24th Digital Avionics Systems Conference, vol. 1 (2005), doi:10.1109/DASC.2005.1563300
29. Argrow, B., Lawrence, D.: UAV Systems for Sensor Dispersal, Telemetry, and Visualization in Hazardous Environments, AIAA Aerospace Sciences Meeting and Exhibit (2005)
30. Casbeer, D.W., Kingston, D.B., Beard, R.W., Mclain, T.W.: Cooperative forest fire surveillance using a team of small unmanned air vehicles, International Journal of Systems Science, pp. 1-18 (2005)
31. Viguria, A., Maza, I., Ollero, A.: Distributed Service-Based Cooperation in Aerial/Ground Robot Teams Applied to Fire Detection and Extinguishing Missions, Advanced Robotics 24, pp. 1–23 (2010)
32. Scheutz, M., Schermerhorn, P., Bauer, P.: The Utility Of Heterogeneous Swarms Of Simple UAVs With Limited Sensory Capacity In Detection And Tracking Tasks, Proceedings 2005 IEEE Swarm Intelligence Symposium, pp. 257-264 (2005), doi:10.1109/SIS.2005.1501630
33. Chao, H., Baumann, M., Jensen, A., Chen, Y., Cao, Y., Ren, W., McKee, M.: Band-reconfigurable Multi-UAV-based Cooperative Remote Sensing for Real-time Water Management and Distributed Irrigation Control, International Federation of Automatic Control World Congress, 2008

UAV Cooperative Data Processing Using Distributed Computing Platform

Grzegorz Chmaj and Henry Selvaraj

1 Introduction

Distributed processing systems (DPS) find many applications in various systems of different scale. In this paper we present a general structure for a distributed processing system with the focus on multiple-UAV cooperative teams realizing a single mission. The contribution of this paper is the novel concept of distributed processing system in which the management is moved to the nodes that can perform various roles simultaneously. We also present the system design details, including the definition of the task, DSD (Distributed System Data) idea, the energy consumption model and experimentation results that show the value of proposed ideas.

2 Literature overview

The distributed processing systems find many applications on various architectures. Authors of [1] presented the way to model the multiscale distributed computing, using the MML (Multiscale Modeling Language). The presented method is later used for nanotechnology and biophysics applications. One of important aspects for distributed processing is the resource discovery – centralized systems are easy to implement, but introduce the single point of error and can face the bottleneck problems. [2] contains the classification of decentralized resource discovery mechanisms, also in our previous work [3] we presented the use of DHT (Distributed Hash Table) as the resource localization mechanism for the distributed system having nodes equipped with multiple data sources. The efficiency of distributed processing system is often measured by the electrical energy consumption – same way as we model the efficiency in this paper. In [4] the energy evaluation of multi-processor system is presented, along with the experiments comparing two approaches of realizing the multi-processor computations. The specific computational needs can be addressed using the specialized techniques such as CUDA [5], this way increasing the efficiency of the system. For the UAV-based distributed processing system presented in this paper we use the object detection mission as the example. However these kind of UAV teams can realize various other applications, such as area surveillance, object tracking, fire detection etc. [6].

3 System description

The goal of this paper is to present a model of distributed processing operation using UAVs equipped with multiple data sources (such as cameras, radars, etc.). We use the general paradigm of distributed processing system and propose an architecture that is also applicable to other platforms (UGVs, combination of UAVs and UGVs, and systems incorporating elements of other types). Unmanned Aerial Vehicles (interchangeably called *nodes* in this paper, according to the terminology used for distributed processing systems) communicate using the common overlay network, to which the interface is defined. The transmission level and all the abstraction behind the overlay network is out of the scope of this paper. The nature of distributed processing includes the task division into fragments that are further sent for processing to selected nodes. The results are collected at the designated node, and then combined into the final result.

3.1 Architecture overview

Distributed processing systems usually are managed using some designated units, often the management is centralized

G. Chmaj
University of Nevada, Las Vegas, NV, USA
e-mail: grzegorz.chmaj@unlv.edu

H. Selvaraj
Department of Electrical and Computer Engineering,
University of Nevada, Las Vegas, NV, USA
e-mail: henry.selvaraj@unlv.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_67,

and performed by a single element specialized into that role. This element performs the task management. In this paper we propose another approach: we let every node propose its task. This way each node can manage its own task, communicate with other nodes as the task manager and collect the results. *DIAPS* (DIstributed Aerial Processing System) contains V UAV denoted as: $v = 1, 2, \ldots, V$ (1). Each of them can possess multiple properties and can perform various roles. We propose flexible roles management in which roles are not strictly assigned to UAV nodes, like it is implemented in most UAS (Unmanned Aerial Systems) and DPS. In our approach, there are R roles defined denoted as $r = 1, 2, \ldots, R$ (2) and each node must have at least one role assigned and is not limited to their maximum number (3), (4).

$$q_{v,r} = 1 \text{ if role } r \text{ is present on node } v, \quad 0 \quad \text{otherwise} \tag{3}$$

$$\sum_{r=1}^{R} q_{v,r} > 0 \quad v = 1, \quad 2, \ldots, V \tag{4}$$

In this paper, we consider three basic roles defined for the processing system:

- *control node* (*CN*) – node holding this role is supplying the system with the management services, such as asset localization.
- *task manager* (*TM*) – this role is played by a node that registered the task to the CN
- *computation* (*CP*) – the node is processing blocks, either received from other nodes (task owners) or those who are part of its own task

As mentioned earlier, nodes can register the task in the system and play the role of *task manager*. Thus there are Z tasks processed in the system $z = 1, 2, \ldots, Z$ (5), each divided into a given number of blocks $b = 1, 2, \ldots, B_z$ (6). Blocks' affiliation is denoted by variable x_{bz} (7):

$$x_{bz} = 1 \textit{ if block } b \textit{ belongs to task } z;\ 0 \text{ otherwise}. \tag{7}$$

Each UAV node holds some computation power provided by *processing units* (*PU*) denoted as $p = 1, 2, \ldots, P$, also a vehicle has G hardware sockets installed (8) denoted as $g = 1, 2, \ldots, G$. Each *pu* has the type defined and is installed in one of vehicle's sockets (9). Each socket must be equipped with processing unit (10).

$$w_{v,g} = 1 \textit{ if socket } g \textit{ is installed on node } v;\ 0 \text{ otherwise}. \tag{8}$$

$$u_{p,g} = 1 \textit{ if processing unit } p \textit{ is installed in socket } g,\ 0 \textit{ otherwise} \tag{9}$$

$$\sum_{p=1}^{P} \sum_{g=1}^{G} w_{v,g}\, u_{p,g} = \sum_{g=1}^{G} w_{v,g} \quad v = 1, \quad 2, \ldots V \tag{10}$$

Unmanned Aerial Vehicles gather data from their surroundings and environment using *data sources* (*DS*) denoted as $d = 1, 2, \ldots, D$ (thus there are D data sources in total present in the described system). Often times UAV's data sources are called *sensors*, however in our work we use only data source, to avoid the confusion with other systems, such as sensor networks. Each vehicle has $\sum_{d=1}^{D} k_{v,d}$ data sources installed, however its number can be zero as well (11, 12).

$$k_{v,d} = 1 \text{ if data source } d \text{ is present on node } v; \quad 0 \text{ otherwise} \tag{11}$$

$$\sum_{d=1}^{D} k_{v,d} \geq 0 \quad v = 1, 2, \ldots, V \tag{12}$$

3.2 Task definition

The proposed UAV-based distributed processing system is designed to perform mission of various kinds. The objectives of the missions and all related and required data is defined in the task object whose internal structure is proposed as follows:

Task is the structure that contains the following elements (Fig. 1):

- *task raw data* – binary/ASCII data that is to be processed during the mission
- *defined blocks* – blocks that are defined by the task designer
- *online blocks* – definitions of the online blocks

Blocks are the fragments of the task that can be separately processed on designated vehicle. We distinguish three types of blocks, denoted as, b_{t2}, b_{t3}:

- *block offline* – the type of block that is created as the division of task raw data ($b_{t1} = 1$)
- *block defined* – the type of block, for which all parameters are specified manually by the task designer ($b_{t2} = 1$)
- *block online* – blocks that are processed in a loop until the task processing is finally finished ($b_{t3} = 1$).

Blocks offline and blocks defined are processed in order to produce one-time result. They might be requested to be processed again as well, but such a request must be placed by task manager. Blocks online are processed repeatedly (thus they are also defined this way) during the task (mission) operation and are able to constantly provide the result of their processing. Each block must have the type assigned (13).

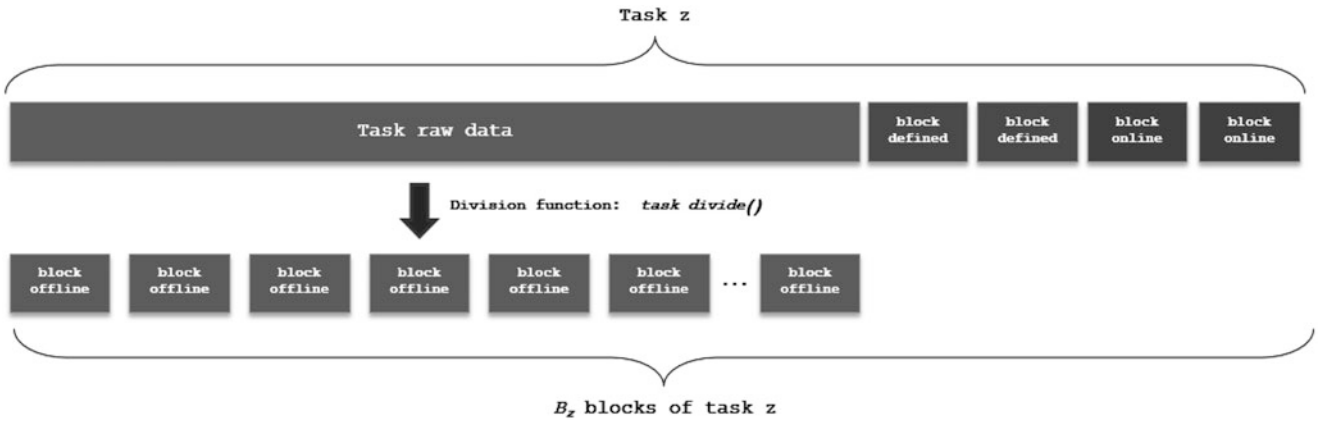


Fig. 1 The structure of the task

$$\sum_{b=1}^{B_z} (b_{t1} + b_{t2} + b_{t3}) = B_z \tag{13}$$

Block has the internal structure defined, that contain the combination of data, data sources and functions. Data is assigned either at the stage of task raw data division using *task_divide*() function (for blocks offline) or manually by the task designer (for blocks defined and blocks online). Data sources specify which data source up-to-date values are required for block processing, given block may require any number of data source values, but also can be defined without such requirement (14, 15). For blocks offline, the *task_divide*() function defines these data sources and each block offline holds the same requirement. Blocks online and blocks defined are specified manually, thus the data sources assignment can be different for each block. Functions identify which functions are used to process the particular block, and there are F functions known to the system, denoted as $f = 1, 2, \ldots, F$. Similarly to data sources, *task_divide*() function assigns functions for blocks offline, and blocks online and blocks defined are set up individually. Each block must have at least one processing function assigned (16, 17).

$$l_{b,d} = 1 \text{ if block } b \text{ requires data source } d;\ 0 \text{ otherwise} \tag{14}$$

$$\sum_{d=1}^{D} l_{b,d} \geq 0 \quad b = 1, 2, \ldots, B_z \tag{15}$$

$$n_{b,f} = 1 \text{ if block } b \text{ requires function } f;\ 0 \text{ otherwise} \tag{16}$$

$$\sum_{f=1}^{F} n_{b,f} > 0 \quad b = 1, 2, \ldots, B_z \tag{17}$$

3.3 Implementation of functions and Processing of blocks

As stated by (16) and (17), block contains the identifier(s) of functions that need to be run for the block in order to obtain the result (19). The identifiers are not the actual functions though. A node must know the function implementation – i.e. the binary implementation i of the function f must be present locally on the vehicle v (21). Functions implementations are denoted as $i = 1, 2, \ldots, F$. This requirement raises the problem of blocks to nodes assignment – blocks have different functions requirements (20) and UAVs possess different functions implementations (18). This problem is solved by separate algorithm *TM_BA* run by task manager.

$$h_{v,f} = 1 \text{ if function } f \text{ is implemented on node } v;\ 0 \text{ otherwise} \tag{18}$$

$$y_{v,b} = 1 \text{ if block } b \text{ is processed on node } v;\ 0 \text{ otherwise} \tag{19}$$

$$n_{b,f} = 1 \text{ if block } b \text{ requires function } f;\ 0 \text{ otherwise} \tag{20}$$

$$\sum_{f=1}^{F} \sum_{b=1}^{B_z} h_{v,f} y_{v,b} = \sum_{b=1}^{B_z} y_{v,b} \tag{21}$$

The matrices of $q_{v,r}$, $w_{v,g}$, $l_{b,d}$, $n_{b,f}$ and $h_{v,f}$ constitute the *distributed system data* (DSD), that is the base for the system operation. Task raw data is the terrain data that is divided into the regions, processed in distributed manner by multiple UAVs in a team. The location of presented matrices depend on the roles: $q_{v,r}$ is stored at *CN* vehicle, $w_{v,g}$, $l_{b,d}$ and $n_{b,f}$ are stored at *TM* node. The knowledge state also depend on the function: $w_{v,g}$ matrix is constant during the mission, so it can be

uploaded to *CN* before mission start (if *CN* is determined) otherwise *CN* will obtain this data from other UAVs. Matrix $q_{v,r}$ is constant if the assignment of *CN* roles is constant, if *CN* roles are assigned dynamically, then this matrix will be dynamically updated as well. Matrices $l_{b,d}$ and $n_{b,f}$ are stored only at *TM* node, that manages related task.

3.4 Energy emission model

The presented algorithms and DIAPS design focus on the efficiency of operation which is expressed as the electrical energy consumption. We choose this metric, as it is the essential factor for UAV based distributed processing systems, because the energy available at each vehicle is limited (either for battery powered units or the ones using combustion engines). The additional motivation is that electrical energy consumption is the base for many other system parameters, such as processing speed, team range of operation, time of operation and many others that are dependent on electrical energy efficiency.

The model defines multiple elements of energy usage:

$e_v^{1,p,f}$ = const – the energy required to process 1kB of data in offline/defined block using function *f* at PU *p* installed on node *v* (size-related)

$e_v^{2,p,f}$ = const – the energy required to process offline/defined block at PU *p* installed on node *v* using function *f* (size-unrelated)

$e_b^{3,p,f}$ = const – the energy required to process block online for time period (slot) *t*

e^4 = const – the energy required to send 1kB of data between two UAVs

s_b^1= size of data enclosed in block b

s_m^2= size of the message m

$a_{v,b} = 1$ if block *b* is processed on node *v*; 0 otherwise.

$t = 1, 2, \ldots, T$ – time period (slot) of the defined length. This index is used for the modeling and the length of *t* determines the accuracy of OPEX modeling. In this paper we assume the *t* being the length of 1 second.

$j_{v,b,t} = 1$if online block *b* is run on the node *v* at the time slot *t*; 0 otherwise.

$y_{b,p} = 1$ if block *b* is processed at socket *p*; 0 otherwise.

c – node index: $c = 1, 2, \ldots, V$

m – message index: $m = 1, 2, \ldots, M$

$k_{m,v,c} = 1$ if message *m* is being transmitted from node *v* to node *c*

The total OPEX (operational expenditure) is:

$$OPEX = \sum_{v=1}^{V}\sum_{b=1}^{B_z}\sum_{f=1}^{F}\sum_{p=1}^{P}\sum_{w=1}^{W}\sum_{g=1}^{G}\sum_{t=1}^{T}(y_{b,p}w_{v,g}s_b + a_v e_v^{2,p,f} y_{b,p} w_{v,g} + tj_{v,b,t} y_{b,p} w_{v,g} e_b^{3,p,f}) + \sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{v=1}^{V}\sum_{c=1}^{V} k_{m,v,c,t} e^4 s_m^2$$

3.5 System operation

DIAPS uses several internal algorithms at the various logical areas of the system. Fig. 2. presents the top-level operational algorithm of the DIAPS, with the logical areas pointed out.0 Team level shows the behavior of the UAV-based distributed system from the mission point of view. Task manager includes the actions taken by the vehicle that implements the *TM* role, and computation node area shows actions executed by *CP* node. Depending on the DIAPS configuration, *TM* node can also implement *CP* role, or they can be placed on separate nodes.

The diagram shows operations for various roles, roles can be executed concurrently (the whole system runs as full parallel architecture) – therefore some simplifications had to be done in order to show the DIAPS operation as a whole. The main internal algorithms are:

- *TM_BA* – run within *TM* role, selects the block to be sent as the response to block request message (sent by computation node within *CP* role). *TM_BA* requires the requestor's processing and DS capabilities enclosed in the block request message.
- *AN_CN* – the automatic assignment of the *CN* role. This algorithm specifies the criteria that a node must meet in order to take *CN* role. This algorithm is executed periodically at each node, however the mission designer can set such criteria that only one UAV will hold the *CN* role during the whole mission.
- *CP_DS* – algorithm used to obtain the *DS* values needed for the block's computation. It must be implemented, as various vehicles can be equipped with same data sources, and *CP* node must select the most convenient one.
- *CP_PU* – algorithm that assigns the block to the available *PU*. This algorithm works at the stage of block requisition – given *PU* node sends the specific information in the block request message, according to resources available locally at that time.

In the work presented hereby, we describe the general functionalities of the main internal algorithms, however their details are out of the scope of this paper.

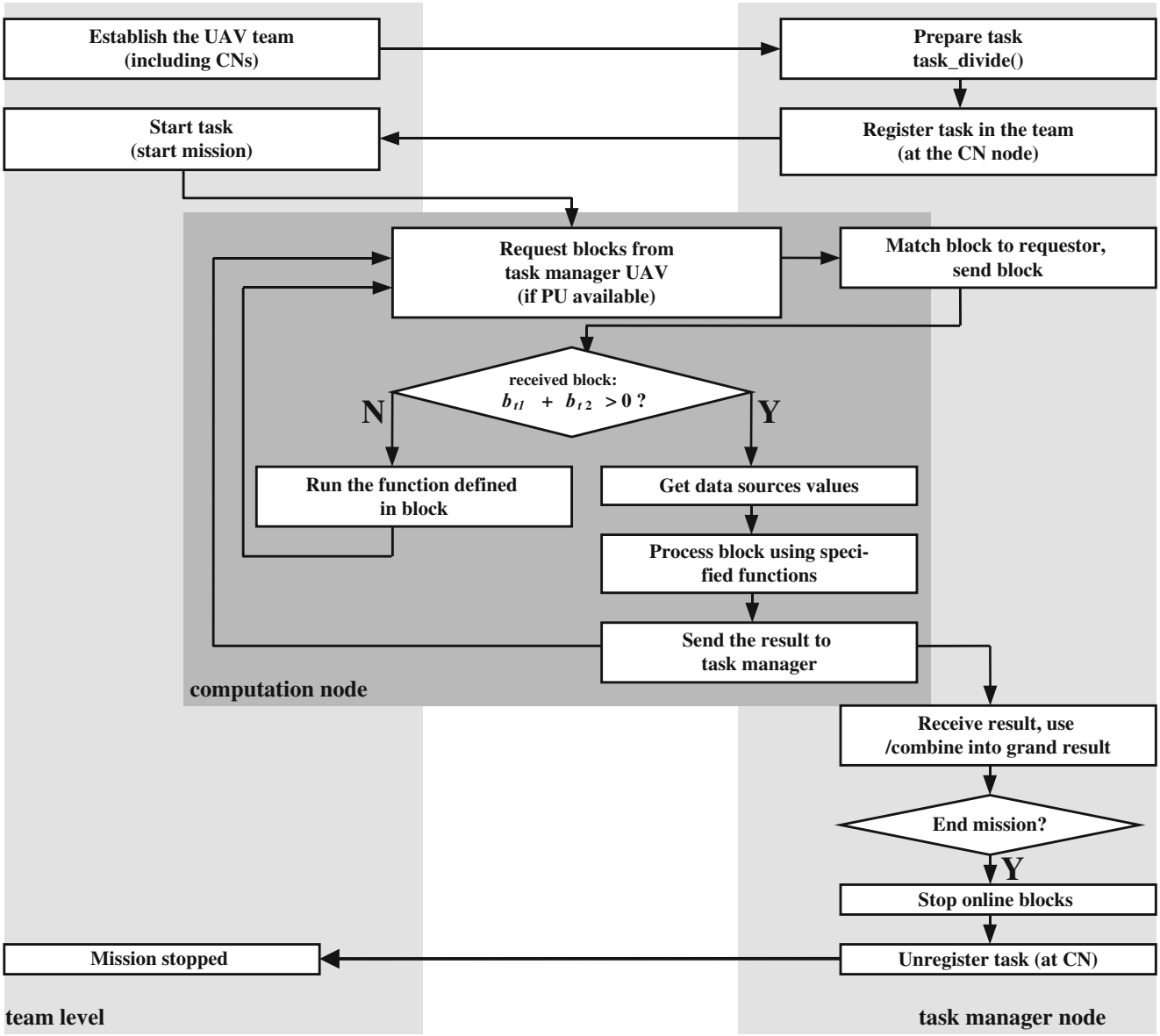


Fig. 2 The algorithm diagram for DIAPS

4 Experimentations

Based on the architecture proposed in previous sections we performed a series of experiments in which we use the following application of UAV-based distributed processing system to realize the object detection mission. The diagram of the simulated DIAPS is shown in Fig. F4, and table 1 contains system elements with values.

Fig. 6 shows the DSD for the DIAPS being the subject of experiment, for $l_{b,d}$ and $n_{b,f}$ blocks 7-38 are offline blocks that have the same *DS* and functions requirements according to *task_divide*() function. We built the object-oriented simulation system DIAPS-SIM, in which we synthesized the functions implementations using *SystemC*. The DIAPS-SIM is written in Ruby 1.9.3 and executes binary functions implementations as threads. Each UAV is modeled as the separate object running concurrently with the local algorithms applied – this way it closely resembles the real UAVs team. The mission being the subject of the research, the task raw data – the terrain map – has the size of 8 MB. The size of the block *size*(1) is set to 256kB, thus there are 32 offline blocks. We use Cortex A5 TSMC 40LP as *PU* units.

Hough transform is used as the example of image processing function used for pattern recognition. Based on [7], [8] and [9], we estimate the energy consumption as:

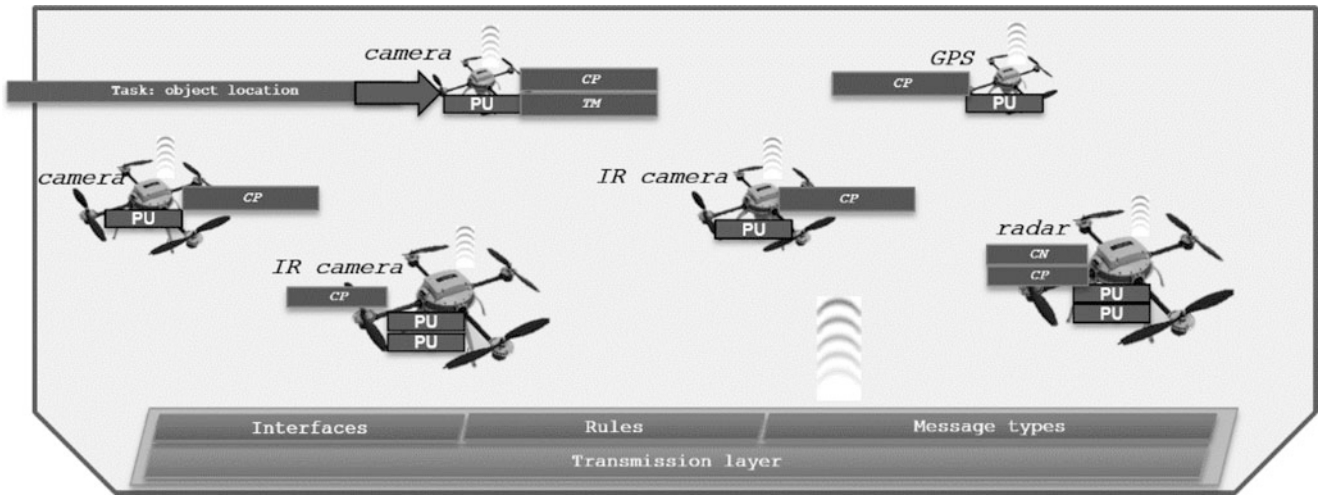


Fig. 3 The schematic of DIAPS for object location mission

$q_{v,r}$ matrix:

r	1	2	3	4	5	6	<-v
1	1	1	1	1	1	1	
2	0	0	0	0	0	1	
3	0	0	1	0	0	0	

$w_{v,g}$ matrix:

g	1	2	3	4	5	6	<-v
1	1	0	0	0	0	0	
2	0	1	0	0	0	0	
3	0	1	0	0	0	0	
4	0	0	1	0	0	0	
5	0	0	0	1	0	0	
6	0	0	0	0	1	0	
7	0	0	0	0	0	1	
8	0	0	0	0	0	1	

$l_{b,d}$ matrix:

b	1	2	3	4	5	6	<-d
1	1	0	0	0	0	0	
2	0	0	1	0	0	0	
3	1	1	0	0	0	0	
4	0	0	1	1	0	0	
5	0	0	0	0	1	0	
6	0	0	0	0	0	1	
7	0	1	0	1	0	1	
8	0	1	0	1	0	1	
9	0	1	0	1	0	1	
10	0	1	0	1	0	1	
...			...				
38	0	1	0	1	0	1	

$n_{b,f}$ matrix:

b	1	2	3	4	<-f
1	1	0	0	0	
2	1	0	0	0	
3	1	1	0	0	
4	1	1	0	0	
5	0	0	1	0	
6	0	0	0	1	
7	1	1	0	1	
8	1	1	0	1	
9	1	1	0	1	
10	1	1	0	1	
...		...			
38	1	1	0	1	

$h_{v,f}$ matrix:

v	1	2	3	4	<-f
1	0	1	0	0	
2	1	0	0	0	
3	1	1	1	1	
4	1	0	0	0	
5	0	0	0	1	
6	0	0	1	0	

Fig. 4 DSD matrices

Table 1 Elements of DIAPS

			$P = G = 8$
Nodes:	$V = 6$	Processing units:	$u_{p,g} = 1\ p = g = 1...8$
Blocks:	$b = 1, 2, \ldots, B_1$ $B_1 = 6 + \left\lceil \frac{size(1)}{s_b^1} \right\rceil$ $\sum_{b=1}^{B_1} b_{t1} = B_1 - 4$ $\sum_{b=1}^{B_1} b_{t2} = 2$ $\sum_{b=1}^{B_1} b_{t3} = 2$	Functions:	$F = 4$ {image filtering – camera picture, image filtering – IR image, radar signal processing, GPS signal interpretation}
Data sources:	$D = 6$ $\sum_{d=1}^{D} k_{v,d} = 1\ v = 1...8$	Roles:	$R = 3$ {*CP, CN, TM*}
Tasks:	$Z = 1$		

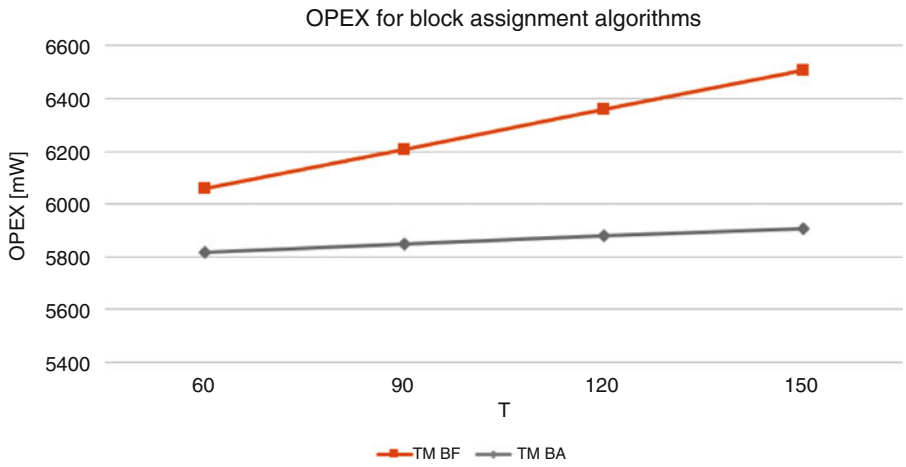


Fig. 5 OPEX for *TM_BF* and *TM_BA*

$e_v^{1,p,1} = 0.08$ mW (1 = Hough Transform), $e_v^{1,p,2} = 0.08$ mW, $e_v^{1,p,3} = 0.05$ mW, $e_v^{1,p,4} = 0.02$ mW for $p = 1, 2, \ldots, P$; $e_v^{2,p,1} = 5$ mW, $e_v^{2,p,2} = 5$ mW, $e_v^{2,p,3} = 3$ mW, $e_v^{2,p,4} = 2$ mW for $p = 1, 2, \ldots, P$; $e_v^{3,p,1} = 2$ mW, $e_v^{3,p,2} = 2$ mW, $e_v^{3,p,3} = 1$ mW, $e_v^{3,p,4} = 1$ mW for $p = 1, 2, \ldots, P$; $e^4 = 0.5$ mW. In our experiment, we compare the energy consumption for the mission using the algorithms set described before and the case where *TM_BA* is replaced with *TM_BF* that assigns blocks regardless the presence of *DS* on given UAV (first-ask-first-granted). The experiment shows that the *DS* locations have strong influence to the energy consumption. Depending on the time mission duration, *TM_BA* saved from 4 % to 9 % of the electrical energy required to complete the mission.

5 Conclusion

In this paper we presented the distributed processing system based on the unmanned aerial vehicles equipped with ARM processors and having the wireless communication capabilities. We have proposed the system design described using the mathematical foundations. We also proposed the key elements of the system such as DSD, listed the necessary algorithms and presented the experimentation results that show the energy consumption for *DS*-location optimized algorithm compared to first-asked-first-granted approach. The results show that the proposed optimization could save up to 9 % of the energy for the mission of duration 150 s. The future work includes the extension of the system model, energy model and further work on the management algorithms in order to lower the energy consumption and increase system efficiency.

References

1. Borgdorff J., Falcone J., Lorenz E., Bona-Casas C., Chopard B., Hoekstra A.: Foundations of distributed multiscale computing: Formalization, specification, and analysis, Journal of Parallel and Distributed Computing, Volume 73, Issue 4, April 2013, pp. 465–483.
2. Lazaro D., Marques J., Jorba J., Vilajosana X.: Decentralized resource discovery mechanisms for distributed computing in peer-to-peer environments, Journal ACM Computing Surveys (CSUR), Volume 45, Issue 4, 2013
3. Chmaj G., Latifi S.: Decentralization of A Multi Data Source Distributed Processing System Using A Distributed Hash Table, International Journal of Communications, Network and System Sciences, vol. 6, no. 10, pp. 451-458, 2013
4. Zydek D., Chmaj G., Chiu S.: Modeling Computational Limitations in H-Phy and Overlay-NoC Architectures, The Journal of Supercomputing, 2013
5. Krol D., Zydek D., Selvaraj H.: Matrix Multiplication in Multiphysics Systems Using CUDA, Adv. in Intelligent Systems and Computing, vol. 240, Springer, 2014, pp. 493-502
6. Chmaj G., Selvaraj H., Distributed processing applications for UAV/drones: a survey, Proceedings of the 23rd International Conference on Systems Engineering ICSEng 2014 (paper accepted)
7. Cortex A5-processor, http://www.arm.com/products/processors/cortex-a/cortex-a5.php
8. Baglietto P., Maresca M., Migliardi M., Zingiria N.: Image Processing on High Performance RISC Systems, Proceedings of the IEEE, vol. 84, issue 7, 1996, pp. 917-930
9. Hollitt C.: Reduction of Computational Complexity of Hough Transforms using a Convolution Approach, 24th International Conference Image and Vision Computing New Zealand, 2009, pp. 373–398.

Implementation of an Efficient Library for Asynchronous Circuit Design with Synopsys

Tri Caohuu and John Edwards

1 Introduction

As the needs of the industry demand ever more complex integrated circuits, it becomes more difficult to supply a single uniform clock signal to the entire device. Currently, the transistors in a clock tree often use an amount of power comparable to the amount used by the transistors implementing the logic. The tree also occupies a large fraction of the chip's area. The problem of clock skew requires elaborate and costly solutions, such as placing multiple phase-locked loops on the same chip.

The alternative is asynchronous circuitry: digital logic without clock signals. Instead of communicating at regular and defined intervals, asynchronous interfaces use control signals to indicate when they are ready to process data. Asynchronous storage elements do not load new values when a clock ticks. Rather, data signals are accompanied by control signals, which notify storage elements when the data signals are valid and should be stored.

Industry standard tools, such as those from Cadence and Synopsys, are designed to deal with problems that arise in synchronous workflows. Timing tools for synchronous circuits tend to focus on solving long paths, but asynchronous circuits are subject to a broader range of timing problems. Common test methods, such as adding scanning to the registers, do not work for asynchronous circuits, which do not have clocked registers.

The asynchronous design methods also pose challenges to integration. The models of delay used in asynchronous circuits are rarely needed for circuits that are constrained by clocks. There is a need for special primitives for encoding data such as used in dual-rail protocols[1] and ternary logic[2]. Integrating these unique features with synchronous tools would be difficult. It would be more effective to begin with asynchronous workflows comparable to the synchronous ones.

In order to use common tools to design asynchronous circuitry, the components used by such circuits first must be integrated with the tools. The first step is to create a design library containing these components.

2 Asynchronous Library

This library should contain basic asynchronous components such as the Muller C-element, join, fork, mux and demux[1]. It can be incorporated into complex digital designs within common EECAD software. This makes it possible to develop a workflow for creating asynchronous circuits with these design tools.

The main feature of the library is a Muller C-element, a simple sequential device often used to control asynchronous pipelines. It has two input signals and one output. When both inputs are high, the output rises and remains high until both inputs are low. The output falls and stays low until both inputs rise again. The output does not change when the inputs are different. It can be used to represent a condition that depends on two prerequisite conditions.

C-element implemented as custom cell will obviously be more efficient than those constructed from logic gates. The implementation chosen was the symmetric static CMOS version, which uses 6 transistor pairs and is more efficient [3] than other static versions.

The transistors gated by the inputs (A and B) are a pull-up and pull-down network. The behavior of the four transistors in a network depends on the transistor gated by the output (C). If it is open, the network is equivalent to two transistors in series. If it is closed, the network is equivalent to two transistors in parallel. Depending on the value of C, the pull-up network will be in series and the pull-down network in parallel, or the other way around. They will behave as either a NAND gate or a NOR gate. The final two transistors are an inverter, changing the function to AND or OR.

T. Caohuu (✉) • J. Edwards
Department of Electrical Engineering, San José State University, San José, CA, USA

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_68,

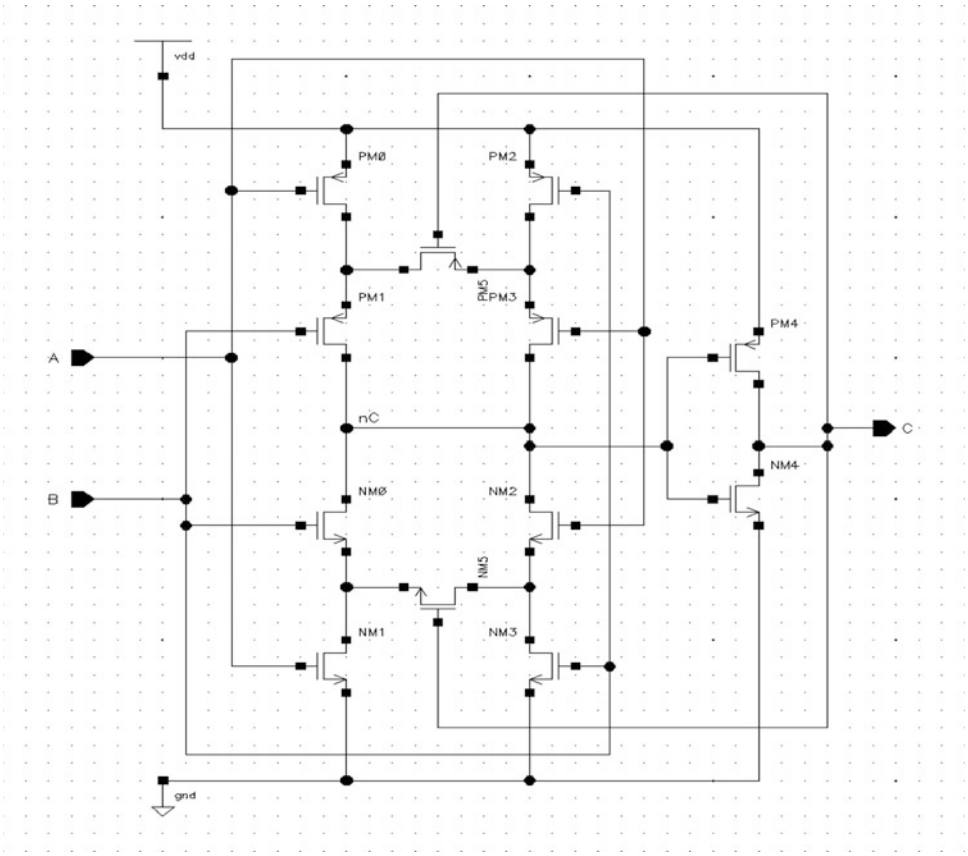


Fig. 1 C-element schematic

This cell was created on a 45 nm process. Transistors with a high threshold voltage were chosen, to minimize leakage current. The widths of the pull-up PMOS transistors were set at 480 nm, and the pull-down NMOS transistors at 240 nm. These were determined by the constraints of the desired cell size. After trying various sizes for the inverter transistors, widths of 480 nm for PMOS and 355 nm for NMOS were found to give approximately equal rise and fall times for a variety of loads.

To minimize cell area, the transistors were ordered using an Euler path technique [5].

The cells were converted to a netlist, which was then characterized by running simulations to measure the cells' electrical characteristics. Their logical and timing behaviors can be determined by this process. The characterizer produces three formats Verilog, VHD, and Liberty. The Liberty format is compatible with synthesis tools, such as Synopsys Design Compiler, and it can be used as a target when synthesizing asynchronous designs. The Verilog and VHDL designs also include timing information, and can be invoked from those languages when running simulations. This allows the accurate testing of asynchronous designs before and after synthesis.

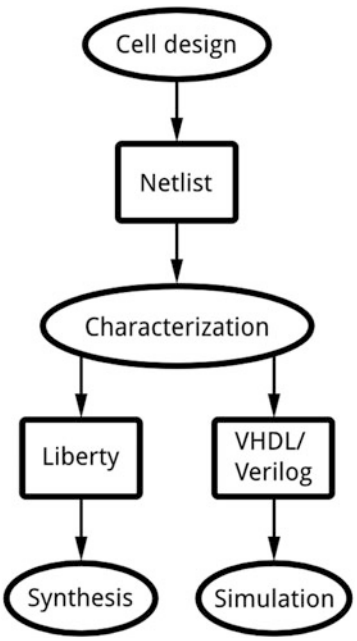


Fig. 2 Library workflow

3 Test Designs

3.1 Asynchronous FIFO

To test the library's performance, the C-element was used as part of an asynchronous FIFO. This FIFO was compared with a similar synchronous FIFO to investigate the relative merits of synchronous and asynchronous circuits for this application. The FIFO is practical for this kind of test, since it is commonly used, particularly in asynchronous circuits. One application is buffering data between different clock domains. An asynchronous FIFO could be used within a mostly synchronous IC.

This FIFO was based on the Muller circuit, a basic asynchronous design pattern. This is a pipeline where propagation is controlled by Muller C-elements.[1] The stages of the pipeline are separated by clock-less latches. Between the latches is only combinational logic. If there is no logic, data simply propagates through, and the pipeline is a FIFO. C-elements drive the control signals, a simple request and acknowledge pair.

It is important that the data signals and the control signals remain synchronized. If there are logic blocks, there must be delays in the control signals that use the same amount of time as the logic.[1]

The pipeline chosen in this design uses a push protocol: the request signal indicates to the next stage that data is available, and the acknowledge signal indicates to the sender that the data is being read. Each of the latches should store a data word whenever the previous latch is sending one and the next latch has already received the one currently stored.

In the push protocol, the request signal indicates to the next stage that data is available. In a pull protocol, it would indicate to the previous stage that there is room for more data. The C-element can be used in a similar manner to control a pull pipeline.

In the push protocol, the request signal is raised to push data to the next stage, and the acknowledge signal is raised once that data is stored in the next latch. In a 2-phase protocol, the request signal falls when the next data word is ready, and the acknowledge signal falls in response.

Since data is sent on both transitions, it must be stored on both transitions of the C-element's output. This is done with a double edge-triggered flip-flop, which is built by combining positive and negative edge-triggered flip-flops and using a multiplexer to select the proper data value. Yun, Beerel, and Arceo point out [4] the complexity of the control circuitry is reduced at the expense of requiring more space for data storage.

Several delays had to be inserted in each pipeline stage to meet timing requirements. The acknowledge signal sent to the previous stage must be delayed until the register has finished loading the data, because once the acknowledgment is sent, the data signals could change, and a hold violation would occur. Similarly, the request signal to the next stage must be delayed until the register's output has stabilized.

Setup violations could occur if the enable signal from the C-element arrived too soon after a transition on the data bus. This solution would be to place a delay block between the C-element and the register. However, this condition was not found to occur during simulations, because the C-element's internal delay was already longer than the setup time. Hence, no delay block was needed.

In case of a pipeline, additional delay proportional to the logic delay for each processing stage is added to the *enable* signal.

Once appropriate delay blocks are chosen, the stage module will adhere to the protocol, even when it is incorporated into a larger design. The asynchronous timing constraints can be met by the library internally, without any other action from the library user or changes to the EDA software.

The FIFO's depth can be changed by adding or removing stages. For this, depths from 4 to 24 were studied. They were implemented as VHDL entities and synthesized. During

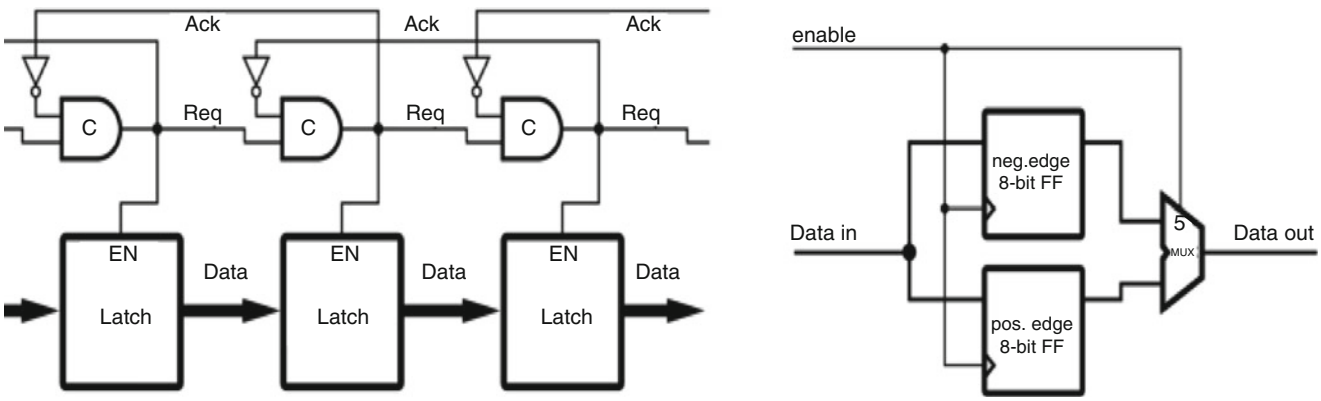


Fig. 3 Async FIFO and Latch Details

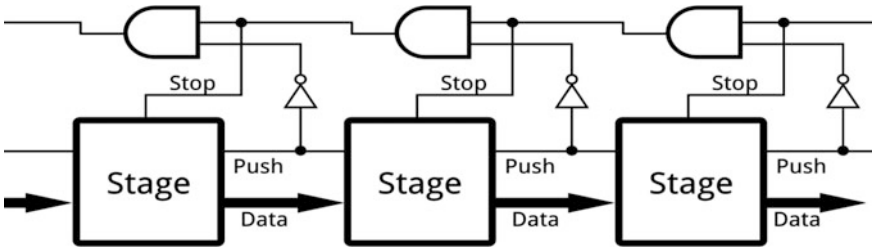


Fig. 4 Block diagram of synchronous FIFO

Table 1 Comparison data for depth 8

	Synchronous	Asynchronous
Maximum throughput	323 MHz	298 MHz
Latency	24.8 ns	12.0 ns
Area	709.0 μm^2	975.2 μm^2
Power (max throughput)	6.300 mW	7.471 mW
Power (200 MHz)	6.141 mW	5.907 mW
Power (100 MHz)	6.038 mW	4.305 mW
Power (40 MHz)	6.096 mW	3.406 mW

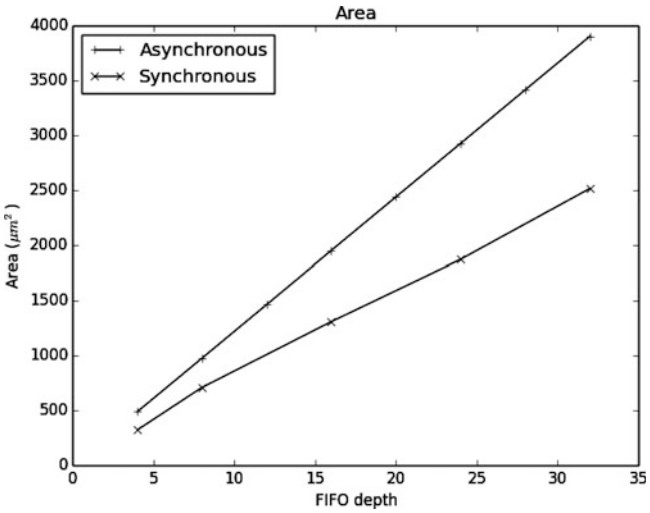


Fig. 5 Area vs. FIFO depth

synthesis, the C-element was provided by the asynchronous library and logic by a set of standard cells.

3.2 Synchronous FIFO

In this case study, a shift register based design was used for the synchronous FIFO. While a memory-based design would have much lower latency, it would differ too much from the asynchronous design architecturally for a more meaningful comparision.

Like the asynchronous design, it contained identical stages, and data words propagated from one stage to the next. The stages communicate with push signals, which indicate that a stage has valid data to send, and stop signals, which indicate that a stage is unable to receive data. As in the asynchronous FIFO, these signals determine whether a stage's register stores or holds. However, these registers are flip-flops, all driven by a single clock signal.

The synchronous FIFO's depth can also be changed by adding more stages. But the stop signals run the entire length of the pipeline, making a longer path if there are more stages. The clock frequency must be reduced to compensate.

This FIFO was also implemented in VHDL and synthesized, using as a synthesis target the same standard cell library used for the asynchronous FIFO.

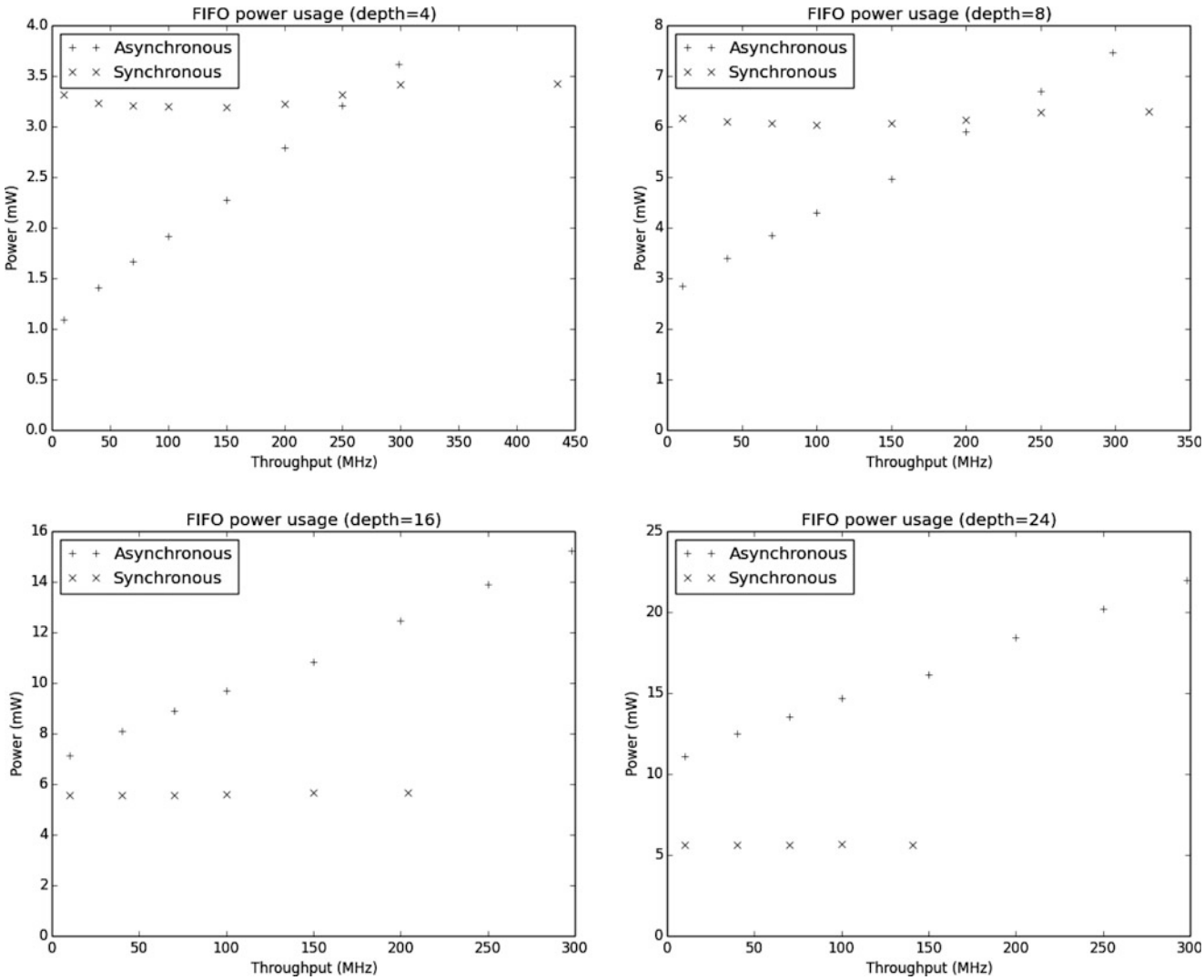


Fig. 6 Power usage vs. throughput at various FIFO depths

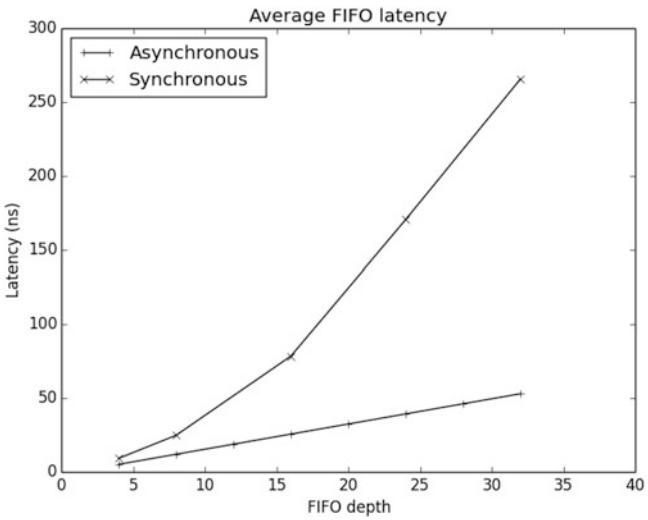


Fig. 7 Latency vs. FIFO depth

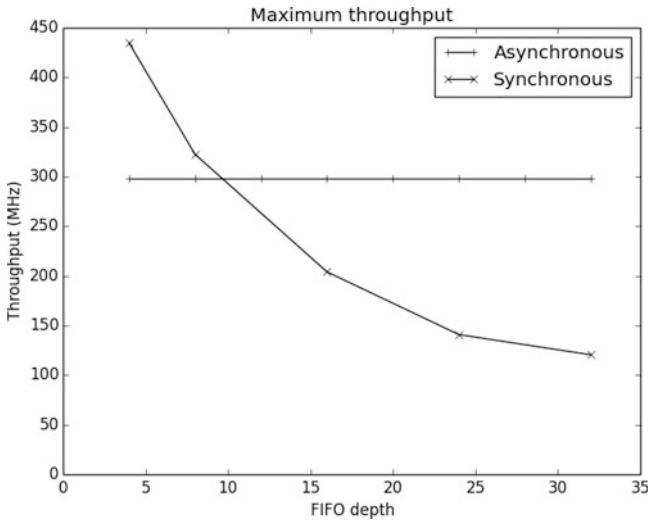


Fig. 8 Throughput vs. FIFO depth

4 Result Comparison and Analysis

The FIFO designs of various depths were compared on the criteria of minimum latency, maximum throughput, and area. The power usage was also measured at various throughput rates. The measurements were produced during the synthesis step. The synthesized designs were then tested to determine maximum performance.

The circuits' performance was quantified by measuring throughput and latency. Throughput was defined as the number of data blocks processed per second. Latency was defined as the time required for one data block to pass from one end of an empty FIFO to the other. For the asynchronous design these were measured by simulations, which used the HDL simulation version of the library.

When taking power measurements of asynchronous designs, the throughput was controlled by setting the switching rate on the write-side request signal. The switching rate of the read-side acknowledge signal was the same, because that is necessary to prevent filling and stalling the pipeline.

The synchronous FIFO clock rates were set to the highest values that did not cause timing violations. The throughput was controlled by the fraction of time the push signal was high. The latency was one clock cycle per stage.

Area varies linearly with depth for both designs. This is to be expected, since both are repetitions of a stage. The linearized curve for area used in the synchronous FIFO rises with a smaller coefficient and uses less area at any depth. This is to be expected, since the asynchronous FIFO has more complex control circuitry. The latency is linear for the asynchronous FIFO, but the synchronous FIFO shows a quadratic dependency.

The asynchronous FIFO's maximum throughput is independent of depth. The synchronous FIFO has an inverse relationship between depth and throughput. Note that they are equal at a depth of 10.

The asynchronous design saves power more effectively when utilization is less than 100 %. At a depth of 8, it required 45 % less power when idle and 25 % less when half idle. The asynchronous design does not show an advantage in power at depths of 16 or more.

5 Conclusions

We have demonstrated that the library performs successfully as indicated by the results obtained. We are able to invoke the Library at the HDL level similar like the case of other synchronous primitive.

While the chosen synchronous design does not reflect the most effective design, the throughput of the asynchronous is more or less constant with respect to the FIFO depth. Moreover, at a smaller depth the asynchronous design show clearly some power advantage. It would be worthwhile to develop more asynchronous primitive components to accommodate the design of asynchronous circuits of higher level of complexity.

References

1. Sparsø, J.; *Asynchronous Circuit Design: A Tutorial*; Technical University of Denmark; p.16-18, March 2006
2. Nagata, Y.; Mukaidono, M.;, "Design of an asynchronous digital system with B-ternary logic," *Multiple-Valued Logic, 1997. Proceedings., 1997 27th International Symposium on*; pp.265-271; May 1997

3. Shams, M.; Ebergen, J.C.; Elmasry, M.I.;, "Modeling and comparing CMOS implementations of the C-element," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol.6, no.4; p.564; Dec. 1998
4. Yun, K., Beerel. P., Arceo, J.; "High-Performance Asynchronous Pipeline Circuits", *ASYNC '96: Proceedings of the 2nd International Symposium on Advanced Research in Asynchronous Circuits and Systems*; 1996
5. Roy, K. "Optimal Gate Ordering of CMOS Logic Gates Using Euler Path Approach: Some Insights and Explanation", *Journal of Computing and Information Technology*" vol. 15. No. 1 pp. 85-92, 2007

A Dynamic System Matching Technique-An Analytical Study

Peter Stubberud, Stephen Stubberud, and Allen Stubberud

1 Introduction

Consider the design and production of an analog system where the desired design is defined by a differential equation which defines a desired relationship between an input signal and an output signal. Production systems will deviate from the desired design due to manufacturing errors in the coefficients of the differential equations describing the production systems. These errors contribute to the errors in the desired system input-output relationship. In an earlier paper, [2], the authors presented a method for combining the measurements from many nominally equal micro-gyroscopes using a technique based on a design technique from electronics called dynamic element matching, see [1]. This technique essentially transforms the system output noise caused by manufacturing errors into an additive (almost) white noise in the system output. This 'spreading of the spectrum' reduces the noise power in the pass band of the analog system. Additional reduction in the effect of the noise can be attained by appropriately filtering the system output. In a more recent paper [3], this technique was generalized by applying dynamic element matching to analog systems. Since the method deals with systems rather than elements, it was called a *dynamic system matching technique* (DSMT). The DSMT proposed in that paper, generates the system output by randomly switching between the outputs of several, nominally identical, production systems. \A heuristic analysis in that paper indicated that the DSMT is effective in reducing the effects of the random coefficient variations in a system output. In a more recent paper [4], a simulation study of the DSMT was developed which, along with some additional analysis, provided further validation of the DSMT. In this paper, a detailed analysis of the noise in the output signal of a DSMT-based system is presented. The results of this study provide not only a validation of the effectiveness of the DSMT, but also provide formulae which can be used to aid in the design of a DSMT system.

P. Stubberud (✉)
Department of Electrical and Computer Engineering, University of Nevada, Las Vegas, Las Vegas, USA
e-mail: stubber@ee.unlv.edu

S. Stubberud
Oakridge Technology, Del Mar, CA, USA
e-mail: scstubberud@ieee.org

A. Stubberud
Department of Electrical Engineering and Computer Science, University of California, Irvine, Irvine, USA
e-mail: stubberud@att.net

2 A Dynamic System Matching Technique

Assume that a system design can be represented by a ***nominal differential equation***

$$\frac{d^n y}{dt^n} + a_{n-1}\frac{d^{n-1} y}{dt^{n-1}} + \cdots + a_1 \frac{dy}{dt} + a_0 y = z \qquad (1)$$

where the n coefficients, a_0, a_1, $\cdots$, a_{n-1}, a_n, are called the ***nominal coefficients*** and the equation represents a ***nominal system***. Now, consider a production version of the system which is defined by the nominal differential equation with the same nominal coefficients except that the coefficient a_0 is replaced by a perturbed coefficient, $a_0 + \Delta a_0$, where the perturbation Δa_0 is a random variable with the first and second order statistics, $E[\Delta a_0] = 0$ and $\mathrm{Var}[\Delta a_0] = \sigma^2_{a_0}$. This system will be called a ***real system***. The input signal to the real system is the same as for the nominal system and the real system output signal is perturbed from $y(t)$ to $y(t) + \Delta y(t)$; therefore, the real system is defined by the differential equation:

$$\frac{d^n (y + \Delta y)}{dt^n} + a_{n-1}\frac{d^{n-1}(y + \Delta y)}{dt^{n-1}} + \cdots + a_1 \frac{d(y + \Delta y)}{dt} + (a_0 + \Delta a_0)(y + \Delta y) = z$$

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_69,

Assume that Δa_0 is small enough that the second order term $\Delta a_0 \Delta y$ is negligible and can be ignored. Under this assumption, the differential equation defining the random perturbation $\Delta y(t)$ is linear and given by:

$$\frac{d^n \Delta y}{dt^n} + a_{n-1}\frac{d^{n-1}\Delta y}{dt^{n-1}} + \cdots + a_1 \frac{d\Delta y}{dt} + a_0 \Delta y = -\Delta a_0 y \quad (2)$$

Now consider a set of N real systems with a common input $z(t)$ applied to each system. The real systems will differ from each other in that the random coefficient perturbations will define a set of N independent identically distributed random variables, Δa_0^i, $i = 1, 2, \cdots, N$. The output signals of the real systems are combined through a switching circuit in a DSMT structure. The switching circuit output signal at a specific time t is the output signal $y(t)$ of one of the N systems which has been randomly selected by the switching circuit. The switching circuit output continues to be the output of that system over a fixed interval of time, T, the ***switching circuit period***, and then it is switched to the output of a different system which has been randomly selected by the switching circuit. A different system is randomly selected for each successive switching circuit period *ad infinitum*. Note that if each system were a ***nominal system***, each of their outputs would be equal and, assuming perfect switching, the output of the switching circuit would equal the output of the nominal system. Because the random perturbations are independent, the noise in the sequence of switched outputs formed by the switching circuit is an 'almost white' sequence. The time correlation in the sequence is due to the fact that the number of different random perturbations is finite and each perturbation will, with non-zero probability, be chosen more than once by the switch. Apparently, the sequence tends to become 'whiter' as N increases.

The linear perturbation $\Delta y(t)$ can be written in the form of a convolution integral

$$\Delta y(t) = (-\Delta a_0) \int_{\tau=-\infty}^{\tau=t} h(t-\tau) y(\tau) d\tau$$

where $h(t)$ is the unit impulse response of the nominal system defined by Equation (1). In the Fourier transform domain, the linear random perturbation is given by $\Delta Y(j\omega) = (-\Delta a_0) H^*(j\omega) Z(j\omega)$ where $\Delta Y(j\omega)$ is the Fourier transform of $\Delta y(t)$, $Z(j\omega)$ is the Fourier transform of $z(t)$, and $H^*(j\omega) = |H(j\omega)|^2$ where $H(j\omega)$ is the Fourier transform of $h(t)$.

Note: To simplify the mathematical development in this paper, but without diminishing the value of the results, only one of the parameters defining the real systems will be considered to be random. For the more complex development involving several, or all, of the parameters being random, the same basic development can be used for each parameter and, assuming independence of the random variations, results for the case of all parameters being random can be generated by combining the results for the individual parameter variations.

3 The Unfiltered Output Noise in a DSMT System

Consider the set of independent random output perturbations (output noise) $\Delta y^1(t), \Delta y^2(t), \cdots, \Delta y^N(t)$ generated by the set of N random coefficient variations, Δa_0^i, $i = 1, 2, \cdots, N$, associated with the N real systems each of which is driven by the same input signal $z(t)$ and each of which is defined by Equation (2). Now, consider a time sequence of pulses $p_k(t)$, $-\infty < k < \infty$ where the pulse $p_k(t)$ has width T, has unity amplitude for $kT \le t < (k+1)T$, and is zero elsewhere. Let $\Delta y^{i_k}(t)$ represent that one of the linear perturbations that is chosen randomly by the switching circuit during the k^{th} switching period. The noise in the output of the switching circuit of the DSMT can then be written as:

$$M(t) = \sum_k \Delta y^{i_k}(t) p_k(t)$$

Combining the outputs of Equations (1) and (2) the perturbation $\Delta y^{i_k}(t)$ can be written as

$$\Delta y^{i_k}(t) = \left(-\Delta a_0^{i_k}\right) \int_{\theta=0}^{\theta=\infty} h^*(\theta) z(t-\theta) d\theta$$

where $h^*(\theta) = \int_{\tau=0}^{\tau=\theta} h(\theta-\tau) h(\tau) d\tau$ Then

$$N(t) = \sum_k \left(-\Delta a_0^{i_k}\right) \int_{\theta=0}^{\theta=\infty} h^*(\theta) z(t-\theta) d\theta\, p_k(t)$$

$$= \left[\int_{\theta=0}^{\theta=\infty} h^*(\theta) z(t-\theta) d\theta\right] \cdot \left[\sum_k \left(-\Delta a_0^{i_k}\right) p_k(t)\right] = N_1(t) \cdot N_2(t)$$

Assuming that $z(t)$ and Δa_0^i, $i = 1, 2, \cdots, N$ are statistically independent, then $M_1(t)$ and $M_2(t)$ are independent random processes. $M(t)$ is the ***output noise*** in a DSMT system.

4 Power Spectral Density and Total Power in the Output Noise

The autocorrelation function of $M(t)$ can be written as:

$$\begin{aligned} R_M(\tau) &= E[M(t)M(t+\tau)] \\ &= E[M_1(t)M_1(t+\tau)] \cdot E[M_2(t)M_2(t+\tau)] \\ &= R_{M_1}(\tau) \cdot R_{M_2}(\tau) \end{aligned}$$

The power spectral density (PSD) of $M(t)$ is given by:

$$\begin{aligned} S_M(\omega) &= \int_{\tau=-\infty}^{\tau=\infty} e^{-j\omega\tau} R_{M_1}(\tau) R_{M_2}(\tau) d\tau \\ &= \frac{1}{2\pi} \int_{v=-\infty}^{v=\infty} S_{M_1}(\omega - v) S_{M_2}(v) dv \end{aligned}$$

where $S_{M_1}(\omega)$ is the PSD of $M_1(t)$ and $S_{M_2}(\omega)$ is the PSD of $M_2(t)$. Because white noise excites all frequencies of the system equally; in this general development, it is assumed that $z(t)$ is white noise with the mean and the autocorrelation function:

$$E[z(t)] = 0 \quad \text{and} \quad R_z(\tau) = K_z \delta(\tau)$$

Under this assumption, the autocorrelation function of $M_1(t)$ is given by:

$$R_{M_1}(\tau) = E[M_1(t)M_1(t+\tau)] = K_z \int_{\theta=0}^{\theta=\infty} h^*(\theta) h^*(\tau+\theta) d\theta$$

and the PSD of $M_1(t)$ is given by:

$$S_{M_1}(\omega) = K_z |H(j\omega)|^4 \quad -\infty < \omega < \infty$$

Assuming that the nominal system has a low-pass frequency response function with bandwidth ω_s, it is approximated by an idealized low-pass frequency response function $H(j\omega) = u(\omega + \omega_s)u(\omega_s - \omega)e^{-j\omega} \quad -\infty < \omega < \infty$

where $u(\cdot)$ is a unit step function. Using this idealized frequency response function, the PSD of $M_1(t)$ is given by:

$$S_{M_1}(\omega) = K_z u(\omega + \omega_s) u(\omega_s - \omega) \quad -\infty < \omega < \infty$$

and the PSD of $M(t)$ can be written:

$$S_M(\omega) = \frac{K_z}{2\pi} \int_{v=\omega-\omega_s}^{v=\omega+\omega_s} S_{M_2}(v) dv \quad -\infty < \omega < \infty$$

The mean and variance of the noise component $M_2(t)$ are:

$$E[M_2(t)] = 0, \quad \text{Var}[M_2(t)] = \sigma_{a_o}^2, \quad -\infty < t < \infty$$

The autocorrelation function of $M_2(t)$ is developed as follows:

$$R_{M_2}(t,\tau) = E[M_2(t)M_2(\tau)]$$

For $t, \tau \in [kT, (k+1)T)$, $M_2(t) = M_2(\tau) = -\Delta a_0^{i_k}$, and

$$R_{M_2}(t,\tau) = E\left[\left(-\Delta a_0^{i_k}\right)^2\right] = \sigma_{a_0}^2$$

For $t \in [jT, (j+1)T)$, $\tau \in [kT, (k+1)T)$ and $j \neq k$, either $M_2(t) = M_2(\tau) = \Delta a_0^{i_k}$ with probably $\frac{1}{N}$ or $M_2(t) \neq M_2(\tau)$ with probability $\frac{N-1}{N}$; therefore for $j \neq k$

$$\begin{aligned} R_{M_2}(t,\tau) &= E[M_2(t)M_2(\tau) \mid M_2(t) = M_2(\tau)] \cdot \Pr[M_2(t) = M_2(\tau)] \\ &\quad + E[M_2(t)M_2(\tau) \mid M_2(t) \neq M_2(\tau)] \cdot \Pr[M_2(t) \neq M_2(\tau)] \\ &= \frac{\sigma_{a_0}^2}{N} + E[M_2(t)M_2(\tau) \mid M_2(t) \neq M_2(\tau)] \cdot \frac{N-1}{N} \end{aligned}$$

If $j \neq k$ and $M_2(t) \neq M_2(\tau)$, then $E[M_2(t)M_2(\tau)|M_2(t) \neq M_2(\tau)] = 0$ and

$$R_{M_2}(t,\tau) = \frac{\sigma_{a_0}^2}{N}$$

Then, for all t and τ, the autocorrelation function can be written as

$$R_{M_2}(t,\tau) = \left(\frac{N-1}{N}\right) \cdot \sigma_{a_0}^2 \cdot \delta(k,j) + \frac{\sigma_{a_0}^2}{N}$$

where $\delta(k, j)$ is a Kronecker delta, that is,

$$\delta(k,j) = \begin{cases} 1 \text{ for } t, \tau \in [kT, (k+1)\tau) \\ 0 \;\; t \in [jT, (j+1)T),\; \tau \in [kT, (k+1)T) \;\; k \neq j \end{cases}$$

The power spectral density of $M_2(t)$ is given by

$$\begin{aligned} S_{M_2}(\omega) &= \int_{-\infty}^{\infty} e^{-j\omega\theta} \left(\frac{N-1}{N}\right) \cdot \sigma_{a_0}^2 \cdot \delta(k,j) d\theta \\ &\quad + \int_{-\infty}^{\infty} e^{-j\omega\theta} \frac{\sigma_{a_0}^2}{N} d\theta \quad -\infty < \omega < \infty. \end{aligned}$$

The first term in $S_{M_2}(\omega)$ is given by

$$\int_{-\infty}^{\infty} e^{-j\omega\theta}\left(\frac{N-1}{N}\right)\cdot\sigma_{a_0}^2\cdot\delta(k,j)\,d\theta = \left(\frac{N-1}{N}\right)\cdot\sigma_{a_0}^2\cdot\int_{-\infty}^{\infty} e^{-j\omega\theta}\delta(k,j)\,d\theta$$

The integral $\int_{-\infty}^{\infty} e^{-j\omega\theta}\delta(k,j)\,d\theta$ is integrated as follows.

$$\int_{-\infty}^{\infty} e^{-j\omega\theta}\delta(k,j)\,d\theta = \sqrt{2}T\left[\int_{0}^{\frac{\sqrt{2}T}{2}} e^{-j\omega\theta}d\theta + \int_{-\frac{\sqrt{2}T}{2}}^{0} e^{-j\omega\theta}d\theta\right]$$
$$-2\left[\int_{0}^{\frac{\sqrt{2}T}{2}}\theta e^{-j\omega\theta}d\theta - \int_{-\frac{\sqrt{2}T}{2}}^{0}\theta e^{-j\omega\theta}d\theta\right] = T^2\mathrm{sinc}^2\omega\frac{\sqrt{2}T}{4}$$
$$= 4\left[\frac{1}{\omega^2}\left(1-\cos\frac{\omega T}{\sqrt{2}}\right)\right]$$

where $\mathrm{sinc}\,x = \frac{\sin x}{x}$. The second term in $S_{M_2}(\omega)$ is the Fourier transform of the constant $\frac{\sigma_{a_0}^2}{N}$ and is given by $\int_{-\infty}^{\infty} e^{-j\omega\theta}\frac{\sigma_{a_0}^2}{N}d\theta = 2\pi\left(\frac{\sigma_{a_0}^2}{N}\right)\delta(\omega)$. This term indicates that the noise component $M_2(t)$ contains a random DC offset. Finally then the power spectral density (PSD) of the noise component $M_2(t)$ is given by:

$$S_{M_2}(\omega) = 4\left(\frac{N-1}{N}\right)\cdot\sigma_{a_0}^2\cdot\left[\frac{1}{\omega^2}\left(1-\cos\frac{T\omega}{\sqrt{2}}\right)\right] + 2\pi\left(\frac{\sigma_{a_0}^2}{N}\right)\delta(\omega)\quad -\infty<\omega<\infty$$

and the PSD of $M(t)$ is given by:

$$S_M(\omega) = K_z\left(\frac{2}{\pi}\right)\left(\frac{N-1}{N}\right)\sigma_{a_0}^2\cdot\int_{v=\omega-\omega_s}^{v=\omega+\omega_s}\frac{1}{v^2}\left[1-\cos\frac{Tv}{\sqrt{2}}\right]dv + K_z\left(\frac{\sigma_{a_0}^2}{N}\right)\int_{v=\omega-\omega_s}^{v=\omega+\omega_s}\delta(v)dv\quad -\infty<\omega<\infty \tag{3}$$

Using this expression, the total power in the unfiltered noise of a DSMT system can be determined by the following double integral:

$$P_M = \int_{\omega=-\infty}^{\omega=\infty} S_M(\omega)d\omega = K_z\left(\frac{2}{\pi}\right)\left(\frac{N-1}{N}\right)\cdot\sigma_{a_0}^2\int_{\omega=-\infty}^{\omega=\infty}\int_{v=\omega-\omega_s}^{v=\omega+\omega_s}\frac{1}{v^2}\left[1-\cos\frac{Tv}{\sqrt{2}}\right]dv\,d\omega + K_z\left(\frac{\sigma_{a_0}^2}{N}\right)\int_{\omega=-\infty}^{\omega=\infty}\int_{v=\omega-\omega_s}^{v=\omega+\omega_s}\delta(v)dvd\omega$$

Interchanging the order of integrations, this can be rewritten:

$$P_M = K_z\left(\frac{2}{\pi}\right)\left(\frac{N-1}{N}\right)\cdot\sigma_{a_0}^2\int_{v=-\infty}^{v=\infty}\frac{1}{v^2}\left[1-\cos\frac{Tv}{\sqrt{2}}\right]\left[\int_{\omega=v-\omega_s}^{\omega=v+\omega_s}d\omega\right]dv + K_z\left(\frac{\sigma_{a_0}^2}{N}\right)\int_{v=-\infty}^{v=\infty}\delta(v)\left[\int_{\omega=v-\omega_s}^{\omega=v+\omega_s}d\omega\right]dv$$

Carrying out the integration of the ω variable first, results in:

$$P_M = K_z\cdot 2\omega_s\cdot\left(\frac{2}{\pi}\right)\left(\frac{N-1}{N}\right)\cdot\sigma_{a_0}^2\int_{v=-\infty}^{v=\infty}\frac{1}{v^2}\left[1-\cos\frac{Tv}{\sqrt{2}}\right]dv + K_z\cdot 2\omega_s\cdot\left(\frac{\sigma_{a_0}^2}{N}\right)$$

Because of the double pole at the origin, the integral in P_M must be evaluated using the contour integral:

$$\oint\frac{1}{z^2}\left(1-e^{i\left(\frac{T}{\sqrt{2}}z\right)}\right)dz = 0$$

where the contour is the infinite semicircle enclosing the upper half of the complex plane, an analytic region. Performing the integration generates the result:

$$\int_{\omega=-\infty}^{\omega=\infty}\frac{1}{v^2}\left(1-\cos\frac{Tv}{\sqrt{2}}\right)dv = \lim_{R\to\infty,\,\varepsilon\to 0} 2\int_{v=\varepsilon}^{v=R}\frac{1}{v^2}\left(1-\cos\frac{Tv}{\sqrt{2}}\right)dv = \lim_{\varepsilon\to 0}\left[-\int_{\theta=\pi}^{\theta=0}\frac{1}{(\varepsilon e^{i\theta})^2}\left(1-e^{i\left(\frac{T}{\sqrt{2}}\varepsilon e^{i\theta}\right)}\right)i\varepsilon e^{i\theta}d\theta\right] = \frac{T}{\sqrt{2}}\pi$$

Finally, then the total power in the unfiltered noise of a DSMT system is given by:

$$P_M = K_z \cdot 2\omega_s \cdot \sqrt{2} \cdot \left(\frac{N-1}{N}\right) \cdot \sigma_{a_0}^2 \cdot T + K_z \cdot 2\omega_s \cdot \left(\frac{\sigma_{a_0}^2}{N}\right)$$

Note that the parameters $K_z, \omega_s,$ and $\sigma_{a_0}^2$ are fixed by the original system design and are not involved in the design of the DSMT system. Note, also, that the second term in P_N represents the power in a random DC bias and is dependent only on N.

5 Total Power in the Filtered Output Noise

Optimally, the output signal of the DSMT system is filtered by an ideal, unity gain, low-pass filter whose pass band is equal to the system pass band, in which case, the total power remaining in the noise signal, after filtering, is given by:

$$P_f = \int_{\omega=-\omega_s}^{\omega=\omega_s} S_N(\omega)d\omega = K_z\left(\frac{2}{\pi}\right)\left(\frac{N-1}{N}\right) \cdot \sigma_{a_0}^2 \int_{\omega=-\omega_s}^{\omega=\omega_s} d\omega \int_{v=\omega-\omega_s}^{v=\omega+\omega_s} \frac{1}{v^2}\left[1 - \cos\frac{Tv}{\sqrt{2}}\right]dv$$

$$+K_z\left(\frac{\sigma_{a_0}^2}{N}\right) \int_{\omega=-\omega_s}^{\omega=\omega_s} \int_{v=\omega-\omega_s}^{v=\omega+\omega_s} \delta(v)dvd\omega$$

Interchanging the order of the integrations, the filtered noise power can be written as:

$$P_f = K_z\left(\frac{4}{\pi}\right)\left(\frac{N-1}{N}\right) \cdot \sigma_{a_0}^2 \cdot \left\{\left(-1 + \cos\left(\sqrt{2}T\omega_s\right)\right) + \left(\sqrt{2}T\omega_s\right) \int_{x=0}^{x=\left(\sqrt{2}T\omega_s\right)} \frac{\sin x}{x}dx - 2\int_{x=0}^{x=\left(\frac{\sqrt{2}T\omega_s}{2}\right)} \sin x \frac{\sin x}{x}dx\right\}$$

$$+K_z \cdot 2\omega_s \cdot \left(\frac{\sigma_{a_0}^2}{N}\right)$$

Note that the second term, the DC term, has not been changed by the filter and is controlled by N and that the first term is controlled by N and $\sqrt{2}T\omega_s$; therefore, an obvious design methodology is to first choose N to reduce the effect of the DC random bias and then choose $\sqrt{2}T\omega_s$ to reduce the effect of the non-DC noise. In examining the effect of filtering on the output noise, the DC random bias term will be ignored. The non-DC total power terms for the unfiltered and filtered noise powers are defined as:

$$P_M' = K_z \cdot 2\omega_s \cdot \sqrt{2} \cdot \left(\frac{N-1}{N}\right) \cdot \sigma_{a_0}^2 \cdot T$$

$$P_f' = K_z\left(\frac{4}{\pi}\right)\left(\frac{N-1}{N}\right) \cdot \sigma_{a_0}^2 \cdot \left\{\left(-1 + \cos\left(\sqrt{2}T\omega_s\right)\right) + \left(\sqrt{2}T\omega_s\right) \int_{x=0}^{x=\left(\sqrt{2}T\omega_s\right)} \frac{\sin x}{x}dx - 2\int_{x=0}^{x=\left(\frac{\sqrt{2}T\omega_s}{2}\right)} \sin x \frac{\sin x}{x}dx\right\}$$

6 A Measure of the Effectiveness of the DSMT

A measure of the effectiveness of using a DSMT system without filtering and with filtering is how the signal-to-noise ratio changes when these methods are used. The total power in the output signal $y(t)$ of the idealized nominal system with a white noise input is easily seen to be $P_y = K_z \cdot 2\omega_s$. If a real system is chosen at random, the total power in the noise due to the random parameter perturbation Δa_0 is easily shown to be $P_{\Delta y} = K_z \cdot 2\omega_s \cdot \sigma_{a_0}^2$. The signal-to-noise ratio is given by:

$$\frac{P_y}{P_{\Delta y}} = \frac{K_z \cdot 2\omega_s}{K_z \cdot 2\omega_s \cdot \sigma_{a_0}^2} = \frac{1}{\sigma_{a_0}^2}$$

The total power in the unfiltered DSMT noise signal is:

$$P_M = K_z \cdot 2\omega_s \cdot \sqrt{2} \cdot \left(\frac{N-1}{N}\right) \cdot \sigma_{a_0}^2 \cdot T + K_z \cdot 2\omega_s \cdot \left(\frac{\sigma_{a_0}^2}{N}\right)$$

Apparently, the DSMT reduces the random DC bias by a factor of N and, examining the PSD in Equation (3), the DSMT spreads the rest of the noise over the infinite spectrum. Comparing the total noises of a single real system and an unfiltered DSMT system, it is seen that

$$P_M = \left[\frac{\sqrt{2}(N-1)T + 1}{N}\right]P_{\Delta y}$$

which implies that if $T < \frac{1}{\sqrt{2}}$, the noise power in the unfiltered DSMT system will be less than the noise power in a single real system. To examine how filtering further reduces the output noise power, the partial noise terms P_M' and P_f' are compared by examining the ratio:

$$\frac{P_f^{'}}{P_M^{'}} = \frac{\left(-1 + \cos\left(\sqrt{2}T\omega_s\right)\right) + \left(\sqrt{2}T\omega_s\right) \int_{x=0}^{x=\left(\sqrt{2}T\omega_s\right)} \frac{\sin x}{x} dx}{\left(\sqrt{2}T\omega_s\right) \cdot \frac{\pi}{2}}$$

$$\times \frac{-2 \int_{x=0}^{x=\frac{\left(\sqrt{2}T\omega_s\right)}{2}} \sin x \frac{\sin x}{x} dx}{\left(\sqrt{2}T\omega_s\right) \cdot \frac{\pi}{2}}$$

As an example, let $\sqrt{2}T\omega_s = \frac{1}{10}$, then $\frac{P_f^{'}}{P_N^{'}} \cong \frac{\sqrt{2}T\omega_s}{\pi} \cong 0.0318$ which represents an approximate 15 db decrease in the non-DC component of the system noise. If the switching frequency is denoted f and if the maximum frequency in the system output signal is denoted f_s, then:

$$f = \frac{1}{T} = 10\sqrt{2}\omega_s = \sqrt{2} \cdot 20\pi \cdot f_s \cong 88.876 \cdot f_s$$

Thus, a switching frequency of about 90 times the maximum frequency in the output signal, reduces the power in the non-DC component of the DSMT noise by 15 db.

7 Conclusions

In a series of earlier papers, the authors introduced the concept of a dynamic system matching technique (DSMT), as a generalization the dynamic element matching technique which is used in the design of electronic systems. In those papers, heuristic analyses and a simulation were used to argue that the DSMT can be used to reduce the effects of noise in a system output due to manufacturing errors. In this paper, a DSMT technique was developed for an idealized nominal system with a white noise input. A detailed analysis of the noise in the output of that DSMT system was generated which allows a comparison of the noise powers in a non-DSMT system, an unfiltered DSMT system and a filtered DSMT system. In particular, detailed analytical expressions for the power spectral densities and the total powers of the noises in the unfiltered and filtered outputs of a DSMT system were developed. These results show conclusively that for the idealized system with a white noise input, the DSMT reduces the noise power and that filtering the DSMT output further reduces the noise power.

References

1. Stubberud, P.A., Bruce, J.W.: An Analysis of Dynamic Element Matching Flash Digital-to-Analog Converters. In: IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, vol. 48, No. 2, pp.205-213 (2001)
2. Stubberud, P.A., Stubberud, A.R.: A Dynamic Element Matching Technique for Improving the Accuracy of MEMS Gyroscopes. In: Proceedings of the 20th International Conference on Systems Engineering, Coventry University, Coventry, pp.418-422 (2009)
3. Stubberud, P.A., Stubberud, A.R.: A Dynamic Element Matching Technique for Improving the Accuracy of Subsystems. In: Proceedings of the 21st International Conference on Systems Engineering, University of Nevada, Las Vegas, Las Vegas, pp.142-146 (2011)
4. Stubberud, P.A., Stubberud, S.C., Stubberud, A.R.: A Dynamic System Matching Technique-A Simulation Study. In: Proceedings of the 22nd International Conference on Systems Engineering, Coventry University, Coventry, pp.202-206 (2012)

On the effect of High Power Amplifier Non-linearity on the Ergodic Capacity of Multihop MIMO-OFDM Amplify-and-Forward Relay Network

Ishtiaq Ahmad, khaled Ali Abuhasel, and Ateeq Ahmad Khan

1 Introduction

Future-generation wireless networks need to provide ultra-high data rate services in-order to meet the demands of high-bandwidth multimedia applications over cellular networks. OFDM has the ability to effectively combats frequency selectivity, mitigates intersymbol interference and achieves high spectral efficiency. On the other hand, MIMO system can provide two types of gains: spatial multiplexing or capacity gain, and diversity gain. Therefore, the combination of MIMO and OFDM (MIMO-OFDM) is the most potent solution that can provide such high data rates at high spectral efficiencies [1]. Also, to enhance the Quality of Service (QoS) performance and extend the coverage of the communication system, multi-hop functionalities have been integrated into the next generation wireless networks [2]. The multi-hop relaying communication system consists of Base Station (BS), one or more Relay Stations

(RS's) and Mobile stations (MS's) as shown in Fig. 1. The communication channel between BS and RS is called relay link, while the channel between the RS and MS is called access link. RS's can be classified broadly into three categories based upon their forwarding schemes: Amplify-and-Forward (AF), Decode-and-Forward (DF) and Demodulate-and-Forward relaying protocols (DemF). In this work, we consider AF relaying protocol for the analysis. In AF relaying protocol, the RS simply scales the received version and transmits an amplified version to the next RS or MS.

Nonlinearity in wireless communication systems usually arises in the RF front ends due to nonlinear devices such as power amplifiers, low-noise amplifiers, mixers, etc. In this work, we focus on the transmitter-side nonlinearities induced by the HPA, the last stage in the communication chain at the BS and the relay RS's. It is well-known that OFDM has a large peak-to-average-power ratio (PAPR) which makes it very sensitive to high-power amplifier (HPA) nonlinearities at the RF stage of the transmission chain [3]. Furthermore, higher-order modulation schemes (like 16-QAM, 32-QAM and above) are more susceptible to HPA nonlinearities than lower-order modulation schemes. Fig. 2 shows the effect of HPA nonlinearity on the constellation diagrams of 16-QAM signals. There are two important effects of the HPA nonlinearities introduced in the transmitted OFDM signals: Out-of-band distortion and in-band distortion. The out-of-band distortion arising from the spectral broadening effect of the HPA affects other operators operating in the adjacent frequency bands [4], whereas the in-band distortion degrades the own bit error rate (BER) performance and capacity of the cellular operator [5,6,7].

The effects of HPA nonlinearity on the capacity of communication systems have been provided in [5–7]. All these works, and many others in the literature, however presented the analysis for the single hop communication systems. In this article, we present the effects of amplifier nonlinearity on the ergodic capacity of multihop MIMO-OFDM AF relay networks. We derive closed form expressions for the ergodic capacity of the nonlinear system. It is observed that due to HPA nonlinearity, multihop MIMO-OFDM AF relay network experienced more capacity loss as more relay hops are involved.

2 Nonlinear MIMO-OFDM Relaying System Model

We consider a multihop MIMO-OFDM system model where a BS, RSs and MS are all equipped with N transmitting and N receiving antennas and that the transmitted OFDM signal has n subcarriers. The transmitted signal from a BS passes

I. Ahmad (✉) • A.A. Khan
Department of Electrical Engineering, Salman Bin Abdulaziz University, Kharj, Saudi Arabia
e-mail: i.ahmad@sau.edu.sa; aa.khan@sau.edu.sa

k.A. Abuhasel
Department of Mechanical Engineering, Salman Bin Abdulaziz University, Kharj, Saudi Arabia
e-mail: k.abuhasel@sau.edu.sa

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_70,

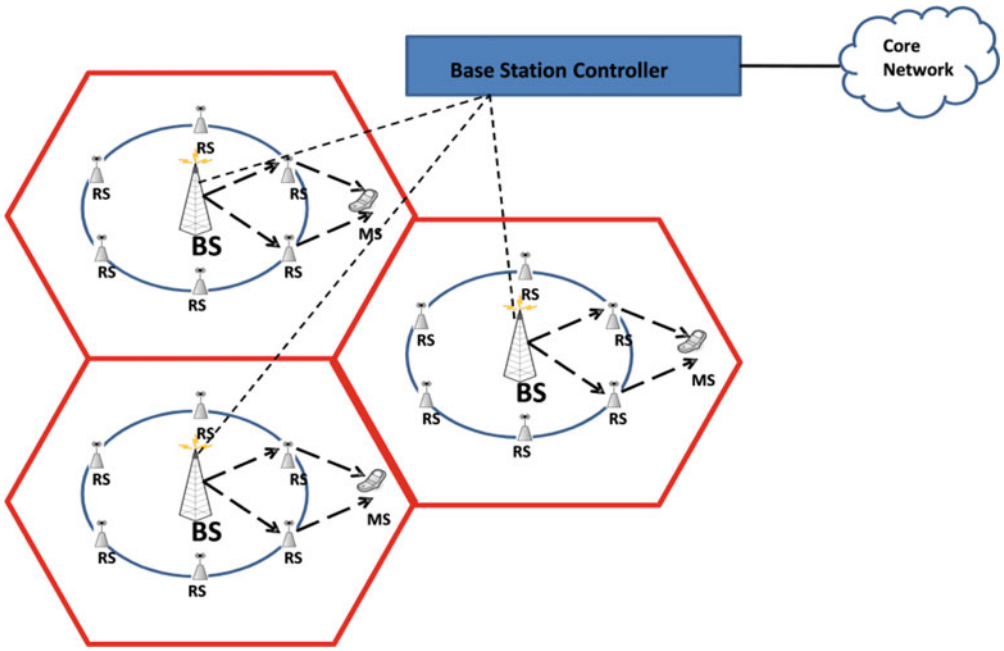


Fig. 1 Multi-hop relaying system architecture

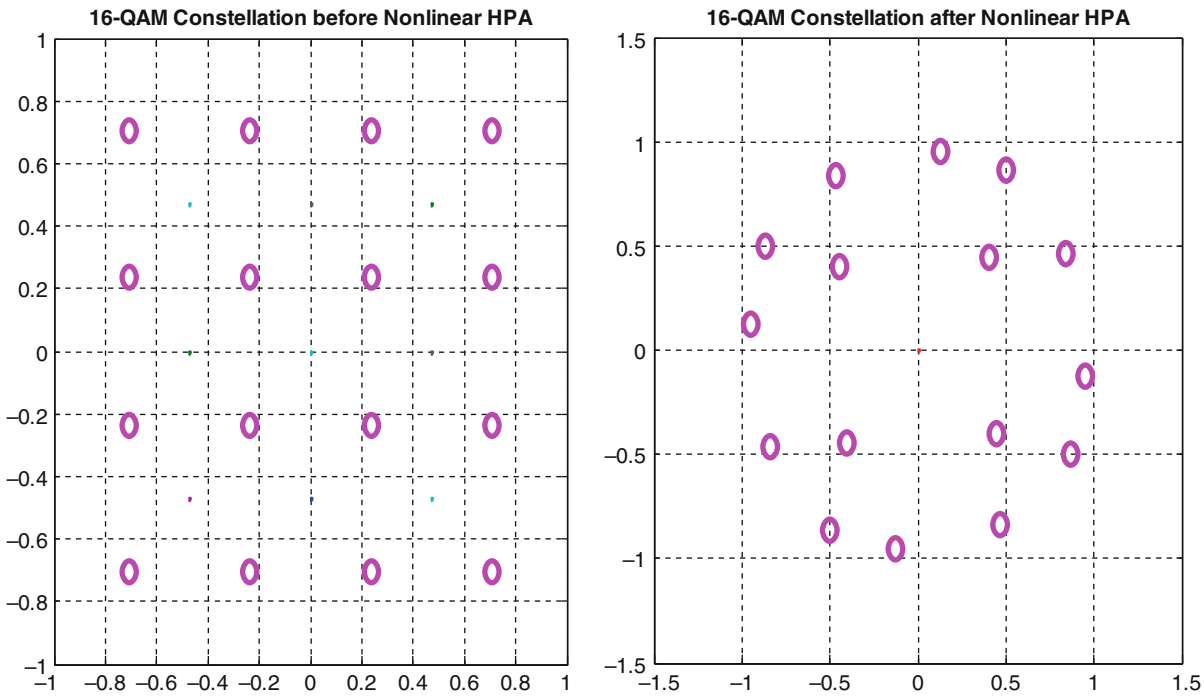


Fig. 2 Effect of HPA nonlinearity on the constellation diagrams of 16-QAM signals

through a single-hop MIMO channel $\mathbf{H}_0$, and R multi-hop MIMO relaying channels $\mathbf{H}_1, \ldots, \mathbf{H}_R$, associated with R fixed or mobile relaying nodes, to the destination node as shown in Fig 3.

The MIMO channel matrix for each i-th hop transmission $\mathbf{H}_i$, $i = 0, 1, \cdots, R$, is an $nN \times nN$ block diagonal matrix, with the l^{th} block diagonal entries $\mathbf{H}[l]_i$ corresponding to the fading on the l^{th} OFDM subcarrier, $l = 0, 1, \ldots, n-1$,

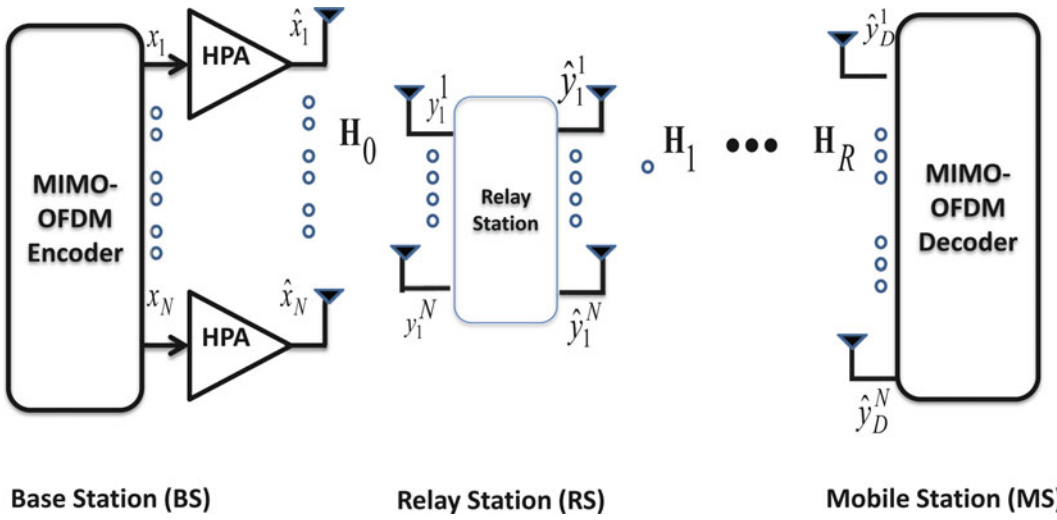


Fig. 3 Nonlinear multihop MIMO-OFDM AF relay system model

modeled as independent and identically distributed (iid) random variables taken from zero mean complex Gaussian distribution, with unit variance. The set of random channel matrices $\{\mathbf{H}_0, \ldots, \mathbf{H}_R\}$ are assumed independent. We can express the transmitted signal in polar coordinate as:

$$x(t) = r(t)e^{j\theta(t)} \tag{1}$$

where $r(t)$ is the amplitude and $\theta(t)$ is the phase of the input signal into the HPA. In a generic way, the output of the HPA can then be expressed as:

$$\widehat{x}(t) = g[r(t)]e^{j\theta(t)} \tag{2}$$

where $g[r(t)]$ is a complex nonlinear distortion function, which only depends on the envelope of the transmitted symbols. The nonlinear distortion function can be expressed as:

$$g[r(t)] = g_A[r(t)]e^{jg_P[r(t)]} \tag{3}$$

where $g_A[r(t)]$ is the amplitude-to-amplitude (AM-AM) and $g_P[r(t)]$ is the amplitude-to-phase (AM-PM) conversions. The (AM-AM) and (AM-PM) conversions for most popular memoryless HPA models used in communication systems are given below.

2.1 Saleh Model

The AM/AM and AM/PM functions for the Saleh model are given as [8]:

$$g_A[r] = \frac{\alpha_A r}{1+\beta_A r^2} \qquad g_P[r] = \frac{\alpha_P r^2}{1+\beta_P r^2} \tag{4}$$

where α_A is the small signal gain, $A_{is} = \frac{1}{\sqrt{\beta_A}}$ is the amplifier input saturation voltage, $A_{os} = \max_r = \{g[r(t)]\} = \frac{\alpha_A A_{is}}{2}$ is the output saturation voltage, and $\frac{\alpha_P}{\beta_P}$ is the maximum phase displacement that may be induced on the amplified signal.

2.2 Solid State Power Amplifier (SSPA) Model

The AM/AM and AM/PM conversions of the SSPA model can be expressed as [9]:

$$g_A[r] = \frac{r}{\left[1+\left(\frac{r}{A_{os}}\right)^{2\beta}\right]^{1/2\beta}} \quad g_p[r] = 0 \tag{5}$$

where β represents the smoothness of transition from linear region to the saturation region.

2.3 Soft Envelop Limiter (SEL) Model

The SEL model can be characterized by the following AM/AM and AM/PM conversions [10]:

$$g_A[r] = \begin{cases} \alpha_A r, & 0 \le r \le A_{is} \\ A_{is}, & r > A_{is} \end{cases} \quad g_p(r) = 0 \tag{6}$$

where $A_{is} = A_{os}$, and A_{os} is as defined above for the saleh model.

According to the Bussgang's theorem, the output of the HPA when the input is a Gaussian process is given as [11]:

$$\widehat{x}(t) = kx(t) + w(t) \tag{7}$$

where k (i.e. $0 \leq k \leq 1$) is an attenuation factor for the linear part and $w(t)$ is a nonlinear additive noise. $w(t)$ is a zero-mean complex Gaussian random variable (r.v,), with the in-phase and quadrature components mutually independent and identically distributed (iid), and with variance σ_w^2. The parameters k and σ_w^2 are referred to respectively as in-band and out-of-band distortion terms.

The closed-form expressions for the in-band and out-of-band distortion parameters for the SSPA HPA model are given by [12]:

$$k^{SSPA} = \begin{cases} \dfrac{\gamma^2}{A_{is}^2} + \sqrt{\dfrac{\pi\gamma^2}{A_{is}^2}}\left(\dfrac{1}{2} - \dfrac{\gamma^2}{A_{is}^2}\right)e^{\frac{\gamma^2}{A_{is}^2}} erfc\left(\dfrac{\gamma}{A_{is}}\right), \beta = 1 \\ \dfrac{\pi^{\frac{3}{2}}\gamma^{\frac{7}{2}}}{2^{\frac{1}{4}}A_{is}^{\frac{7}{2}}\Gamma\left(\dfrac{1}{4}\right)} Y_{\frac{5}{4}}\left(\dfrac{\gamma^2}{A_{is}^2}\right) - \dfrac{2\gamma^4}{3A_{is}^4}\,{}_1F_2\left(1; \left\{\dfrac{1}{2}, \dfrac{7}{4}\right\}; \dfrac{\gamma^4}{4A_{is}^4}\right), \beta = 2 \end{cases} \tag{8}$$

$$\left(\sigma^{SSPA}\right)_w^2 = \begin{cases} 1 - \dfrac{\gamma^2}{A_{is}^2} - \dfrac{\gamma^2}{A_{is}^2}e^{\frac{\gamma^2}{A_{is}^2}}\Gamma\left(0, \dfrac{\gamma^2}{A_{is}^2}\right) \\ + \sqrt{\dfrac{\pi A_{is}^2}{\gamma^2}}\left(\dfrac{2\gamma^4}{A_{is}^4}\right)e^{-\frac{\gamma^2}{A_{is}^2}} erfc\left(\dfrac{\gamma}{A_{is}}\right) \beta = 1 \\ -\pi\left(\dfrac{\gamma^4}{A_{is}^4} - \dfrac{\gamma^2}{A_{is}^2} + \dfrac{1}{4}\right)e^{\frac{2\gamma^2}{A_{is}^2}} erfc^2\left(\dfrac{\gamma}{A_{is}}\right), \\ -\dfrac{\pi\gamma^2}{2A_{is}^2}Y_1\left(\dfrac{\gamma^2}{A_{is}^2}\right) - \dfrac{\pi^3}{8P^5\Gamma^2\left(\dfrac{1}{4}\right)}Y_{\frac{5}{4}}\left(\dfrac{\gamma^2}{A_{is}^2}\right) \\ -\dfrac{\pi\gamma^2}{2A_{is}^2}H_{-1}\left(\dfrac{\gamma^2}{A_{is}^2}\right) - \dfrac{4\gamma^6}{9A_{is}^6}\,{}_1F_2^2\left(1; \left\{\dfrac{1}{2}, \dfrac{7}{4}\right\}; \dfrac{\gamma^4}{4A_{is}^4}\right) \beta = 2 \\ -\dfrac{2^{\frac{7}{4}}\pi^{\frac{3}{2}}\gamma^{\frac{11}{2}}}{3A_{is}^{\frac{11}{2}}\Gamma\left(\dfrac{1}{4}\right)} Y_{\frac{5}{4}}\left(\dfrac{\gamma^2}{A_{is}^2}\right)\,{}_1F_2\left(1; \left\{\dfrac{1}{2}, \dfrac{7}{4}\right\}; \dfrac{\gamma^4}{4A_{is}^4}\right) \end{cases} \tag{9}$$

$$\gamma = \sqrt{IBO} = \frac{A_{is}}{\sqrt{P}} \tag{10}$$

and the closed-form expressions for the in-band and out-of-band distortion parameters for the SEL HPA model are given as [13]:

$$k^{SEL} = 1 - e^{-\gamma^2} + \frac{1}{2}\sqrt{\pi}\gamma erfc(\gamma) \tag{11}$$

$$\left(\sigma^{SEL}\right)_w^2 = P\left(1 - e^{-\gamma^2} - k^2\right) \tag{12}$$

where, γ represents the clipping ratio (CR), A_{is} represents the input saturation voltage of the HPA for each transmission chain, and P represents the average input power into the HPA. $\Gamma(.)$ is the generalized hypergeometric function given by [14]:

$$\begin{aligned} &{}_pF_q\left(\{a_1, \ldots, a_p\}; \{b_1, \ldots, b_q\}; z\right) \\ &= \sum_{n=0}^{\infty} \frac{(a_1)_n \cdots (a_p)_n}{(b_1)_n \cdots (b_q)_n} \frac{z^n}{n!} \end{aligned} \tag{13}$$

and $\mathbf{Y}_v(z)$ is the Bessel function of the second kind, which is given by [15]:

$$Y_v(z) = \frac{J_v(z)\cos\pi v - J_{-v}(z)}{\sin\pi v} \tag{14}$$

where $\mathbf{J}_v(z)$ is the Bessel function of order v.

3 Nonlinear Multihop MIMO-OFDM AF Relay Network

The received signal on the m^{th} subcarrier at the first relay station after propagating through $N \times N$ MIMO channel, $\mathbf{H}_i[m]$, $i = 0, 1, \ldots, R$, in the frequency domain (FD) is given by,

$$y_1[m] = K_0 H_0[m]x[m] + H_0[m]w_0[m] + n_0[m],\ m = 1, \ldots n \tag{15}$$

where, $\mathbf{y}_1[m] = [y_1^1[m], y_1^2[m], \cdots, y_1^N[m]]^T$ and $y_i^j[m]$ is the received OFDM symbol at the j^{th} receiving antenna in the i^{th} hop transmission. The relay node demodulates the received signal and then normalize each subcarrier separately in the FD with a normalization parameter $\boldsymbol{\alpha} = diag\{\alpha_i^j\}_{j=1}^N$, $i = 1, \ldots, R$. The normalization parameter $\alpha_i^j[m] = \sqrt{\dfrac{P_R}{E\left[\left|y_i^j[m]\right|^2\right]}}$ is selected such that the total transmit power at the relay station is constrained to P_R (i.e. total power available at the relay station) for all subcarriers. Before transmitting to the next hop, the OFDM signal is passed through HPA at the RF stage of the relay station as shown in Fig 3. The received signal at the destination after propagating through R relays in the FD after the FFT, is given by

$$y_D = \prod_{i=0}^{R} K_i \prod_{i=1}^{R} \alpha_i (H_R H_{R-1} \ldots\ldots H_1 H_0)x + n_D \tag{16}$$

where $\mathbf{K}_i$, $i = 0 \ldots\ldots R$, is an $nN \times nN$ block diagonal matrix of the attenuation factors for the linear part of HPA,

$\boldsymbol{\alpha}_i$, $i = 1, \ldots, R$, is an $nN \times nN$ block diagonal matrix of the amplification parameter in the in i^{th} hop transmission, and $\mathbf{n}_D$ is the $nN \times 1$ complex additive noise vector that captures the over-all noise in the multi-hop channel. The over-all complex additive noise vector $\mathbf{n}_D$ is modeled as,

$$n_D = n_0 + n_1 + ... + n_R \tag{17}$$

where,

$$\begin{aligned}
n_R &= \widetilde{n}_R + H_R w_R, \\
n_{R-1} &= \alpha_R K_R H_R \widetilde{n}_{R-1} + \alpha_R K_R H_R H_{R-1} w_{R-1}, \\
n_{R-2} &= \alpha_R \alpha_{R-1} K_R K_{R-1} H_R H_{R-1} \widetilde{n}_{R-2} \\
&\quad + \alpha_R \alpha_{R-1} K_R K_{R-1} H_R H_{R-1} H_{R-2} w_{R-2}, \\
&\cdot \\
&\cdot \\
n_1 &= \alpha_R \cdots \alpha_2 K_R \cdots K_2 H_R \cdots H_2 \widetilde{n}_1 + \alpha_R \cdots \alpha_2 K_R \\
&\quad \cdots K_2 H_R \cdots H_1 w_1, \\
n_0 &= \alpha_R \cdots \alpha_1 K_R \cdots K_1 H_R \cdots H_1 \widetilde{n}_0 + \alpha_R \cdots \alpha_1 K_R \\
&\quad \cdots K_1 H_R \cdots H_0 w_0
\end{aligned} \tag{18}$$

$\tilde{n}_i$ is the $nN \times 1$ iid zero-mean complex AWGN vector, introduced at the i^{th} hop transmission on the n subcarriers. The covariance matrix of $\mathbf{n}_D$ can be computed as:

$$\begin{aligned}
E\left[n_D n_D^H\right] &= \sigma_{n_D}^2 I_{nN} = E\left[n_0 n_0^H\right] + E\left[n_1 n_1^H\right] + E\left[n_2 n_2^H\right] + \cdots \\
&\quad + E\left[n_{R-1} n_{R-1}^H\right] + E\left[n_R n_R^H\right] \\
&= \sigma_0^2 \left(1 + \sum_{i=1}^{R} \left[\prod_{j=1}^{R-i+1} \alpha_{R+1-j}^2 K_{R+1-j}^2 N^{R+1-i}\right]\right) I_{nN} \\
&\quad + \left(\sum_{i=0}^{R-1} \left[\prod_{j=1}^{R-i+1} \alpha_{R+1-j}^2 K_{R+1-j}^2 N^{R+1-i} \sigma_{w_i}^2\right]\right) I_{nN} + N\sigma_{w_R}^2 I_{nN}.
\end{aligned} \tag{19}$$

4 Ergodic Capacity of Nonlinear Multihop MIMO-OFDM AF Relay Network

The ergodic capacity of the multi-hop MIMO-OFDM AF relaying channel can be expressed as:

$$\begin{aligned}
C_{AF} &= \frac{1}{n} \sum_{m=0}^{n-1} E_{H[m]_R, \ldots, H[m]_0} \left[\log_2 \left\{ \det \left[I_N + \frac{P \prod_{i=0}^{R} K_i^2 \prod_{i=1}^{R} \alpha_i^2}{nN\sigma_{n_D}^2} H[m]_R H[m]_R^H H[m]_{R-1} H[m]_{R-1}^H \ldots H[m]_1 H[m]_1^H H[m]_0 H[m]_0^H \right] \right\}\right] \\
&= \frac{1}{n} \sum_{m=0}^{n-1} E_{H[m]_R, \ldots, H[m]_0} \left[\log_2 \left\{ \det \left[I_N + \frac{P \prod_{i=0}^{R} K_i^2 \prod_{i=1}^{R} \alpha_i^2}{nN\sigma_{n_D}^2} Q_{m,R} Q_{m,R-1} \cdots Q_{m,0} \right] \right\}\right]
\end{aligned} \tag{20}$$

where $\mathbf{Q}_{m,j} = \mathbf{H}[m]_j \mathbf{H}[m]_j^H$.

Using the eigenvalue decomposition of $\mathbf{Q}_{m,j}$, $j = 0, 1, \cdots, R$, we can express Eq. (20) in terms of the eigenvlaues $\lambda_1^{m,j}, \cdots, \lambda_N^{m,j}$ of $\mathbf{Q}_{m,j}$, $j = 0, 1, \cdots, R$ as:

$$C_{AF} = \frac{N}{n} \sum_{m=0}^{n-1} E_{\lambda^{m,R}, \ldots, \lambda^{m,0}} \left[\log 2 \left(1 + \frac{P \prod_{i=0}^{R} \overset{2}{K}_i \prod_{i=1}^{R} \overset{2}{\alpha}_i}{nN \overset{2}{\sigma}_{n_D}} \prod_{j=0}^{R} \lambda^{m,j}\right)\right] \tag{21}$$

where $\lambda^{m,j}$ is a randomly selected eigenvalue of $\mathbf{Q}_{m,j}$, $j = 0, 1, \cdots, R$, and the pdf of $\lambda^{m,j}$ in case of equal number of transmitting and receiving antennas can be expressed as [16]:

$$f_{\lambda^{m,j}}(\lambda) = \frac{1}{N} \sum_{p=0}^{N-1} \sum_{q=0}^{p} \sum_{r=0}^{2q} \frac{(-1)^r (2q)!}{2^{2p-l} q! r! (q)} \binom{2p-2q}{p-q} \binom{2q}{2q-r} \lambda^r e^{-\lambda} \tag{22}$$

The ergodic capacityof the nonlinear multihop MIMO-OFDM system in Eq. (21) can be expressed as:

$$C_{AF} = \frac{N}{n} \sum_{m=0}^{n-1} \int_0^{\infty} \cdots \int_0^{\infty} \left[\log_2 \left(1 + \frac{P \prod_{i=0}^{R} K_i^2 \prod_{i=1}^{R} \alpha_i^2}{nN\sigma_{n_D}^2} \prod_{j=0}^{R} \lambda^{m,j}\right)\right] \cdot f_{\lambda^{m,R}, \ldots, \lambda^{m,0}}\left(\lambda^{m,R}, \cdots, \lambda^{m,0}\right) d\lambda^{m,R}, \cdots, d\lambda^{m,0} \tag{23}$$

Putting Eq. (22) in Eq. (23), and solving the integrals in Eq. (23) we come up with the following closed form expression for the ergodic capacity of the nonlinear MIMO-OFDM AF relaying system [14, 15],[17, 18]:

$$
\begin{aligned}
C_{AF} \approx & \sum_{p=0}^{N-1}\sum_{q=0}^{p}\sum_{r=0}^{2q}\frac{(-1)^r(2q)!}{2^{2p-l}q!r!q}\binom{2p-2q}{p-q}\binom{2q}{2q-r}r! \\
& \cdot\Bigg\{\ln\left(\frac{P\prod_{i=0}^{R}K_i^2\prod_{i=1}^{R}\alpha_i^2}{nN\sigma_{n_D}^2}\right)\cdot\prod_{j=1}^{R}\left(\frac{1}{N}\sum_{p_j=0}^{N-1}\sum_{q_j=0}^{p_j}\sum_{r_j=0}^{2q_j}\frac{(-1)^{r_j}(2q_j)!}{2^{2p_j-r_j}q_j!r_j!q_j}\binom{2p_j-2q_j}{p_j-q_j}\binom{2q_j}{2q_j-r_j}r_j!\right) \\
& +\sum_{i=1}^{R}\Bigg[\left(\frac{1}{N}\sum_{p_i=0}^{N-1}\sum_{q_i=0}^{p_i}\sum_{r_i=0}^{2q_i}\frac{(-1)^{r_i}(2q_i)!}{2^{2p_i-r_i}q_i!r_i!q_i}\binom{2p_i-2q_i}{p_i-q_i}\binom{2q_i}{2q_i-r_i}\right. \\
& \left.\left(\sum_{m=0}^{(r_i-1)}\frac{r_i!}{(r_i-m)!}-r_i!.0.577\right)\right)\cdot \\
& \cdot\prod_{j=1,j\neq i}^{R}\left(\frac{1}{N}\sum_{p_j=0}^{N-1}\sum_{q_j=0}^{p_j}\sum_{r_j=0}^{2q_j}\frac{(-1)^{r_j}(2q_j)!}{2^{2p_j-r_j}q_j!r_j!q_j}\binom{2p_j-2q_j}{p_j-q_j}\binom{2q_j}{2q_j-r_j}r_j!\right)\Bigg] \\
& +(-0.577)\prod_{j=1}^{R}\left(\frac{1}{N}\sum_{p_j=0}^{N-1}\sum_{q_j=0}^{p_j}\sum_{r_j=0}^{2q_j}\frac{(-1)^{r_j}(2q_j)!}{2^{2p_j-r_j}q_j!r_j!q_j}\binom{2p_j-2q_j}{p_j-q_j}\binom{2q_j}{2q_j-r_j}r_j!\right) \\
& +\sum_{m=1}^{r}\frac{1}{m}\cdot\prod_{j=1}^{R}\left(\frac{1}{N}\sum_{p_j=0}^{N-1}\sum_{q_j=0}^{p_j}\sum_{r_j=0}^{2q_j}\frac{(-1)^{r_j}(2q_j)!}{2^{2p_j-r_j}q_j!r_j!q_j}\binom{2p_j-2q_j}{p_j-q_j}\binom{2q_j}{2q_j-r_j}r_j!\right)\Bigg\}
\end{aligned}
\tag{24}
$$

5 Simulation Results and Discussions

In our simulation setup, we consider multi-hop MIMO-OFDM system for different number of subcarriers and different number of transmit and receive antennas. Fig. 4 compares our simulation results with the approximate analysis, for the cases of linear and nonlinear AF relaying systems, for the multi-hop capacity of MIMO-multiplexing system using SSPA HPA model. It can be observed that the analysis and simulation results agree closely in high-SNR region for both the linear and nonlinear cases. Similar results for the case of the SEL model are presented in Fig. 5, and it is similarly observed from these results that our analysis and simulation results agree closely.

Comparing Fig. 4 and Fig. 5, it is observed that the capacity of MIMO-multiplexing relaying system degrades more for the SSPA HPA model as compared to the SEL HPA model as expected. For example, for the case of $R = 2$ and $\gamma = 0dB$, the capacity loss due to SSPA HPA model is $(58.36 - 53.21) \approx 6$ bits/sec/Hz while for the SEL HPA model it is $(58.36 - 55.57) \approx 3$ bits/sec/Hz per OFDM subcarrier in high SNR. Fig. 4 and Fig. 5 also indicate that more capacity loss due to HPA nonlinearity are experienced in multi-hop relaying systems as more relay hops are involved.

Next, the effect of HPA nonlinearity on the capacity of different MIMO configurations is illustrated in Fig. 6. Here we observed that, as the dimension of the MIMO system increases the gap between the linear and nonlinear capacity also increases. From this we conclude that the capacity of the high-order MIMO-multiplexing relaying systems degrade more in the presence of HPA nonlinearity.

6 Conclusions

In this paper, we present closed-form expressions for the ergodic capacity of the nonlinear multihop MIMO-OFDM AF relay networks. We also verify the derived expressions using extensive computer simulations. Our results show the capacity of MIMO-multiplexing relaying system degrades more for the SSPA HPA model as compared to the SEL

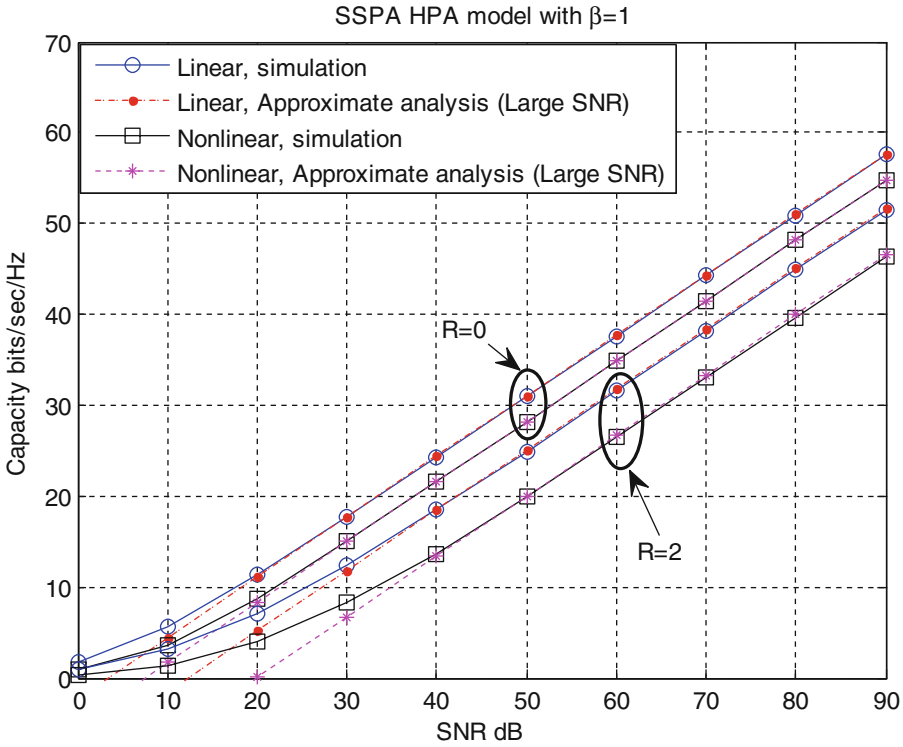


Fig. 4 Simulation and Analytical results for the ergodic capacity of the Linear and Nonlinear $2x2$ MIMO-OFDM AF relaying system with SSPA model, $\beta = 1$ and $\gamma = 0dB$

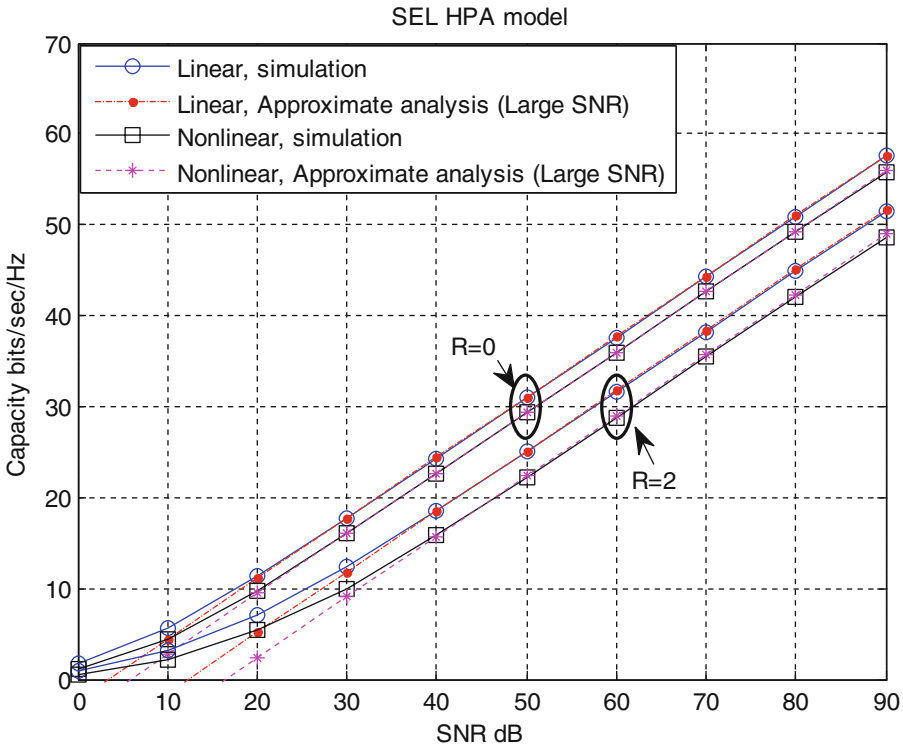


Fig. 5 Simulation and Analytical results for the ergodic capacity of the Linear and Nonlinear 2×2 MIMO-OFDM AF relaying system with SEL HPA model, $\gamma = 0dB$

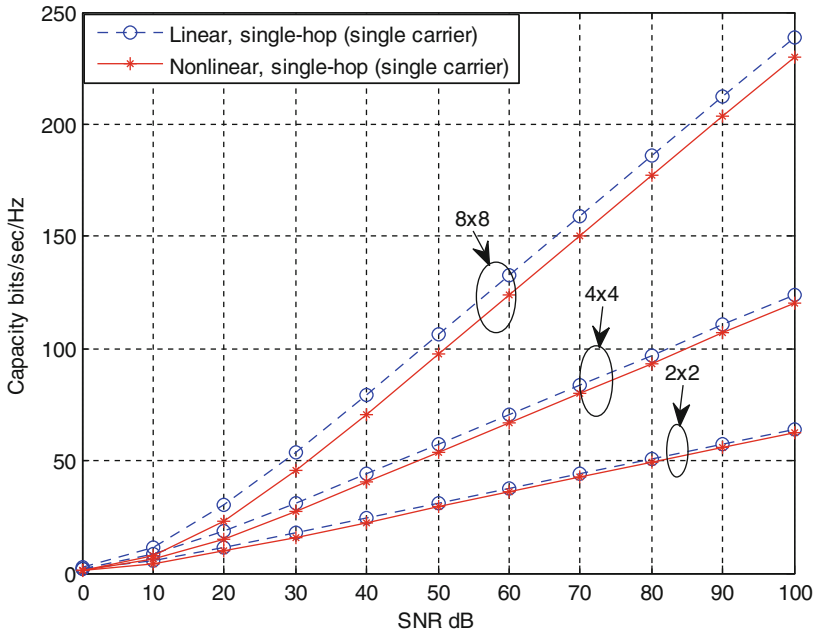


Fig. 6 Effect of HPA nonlinearity on the ergodic capacity of different MIMO for the SEL HPA model, 1-hop transmission

HPA. The results also indicate that high-order MIMO systems suffer more capacity loss due to HPA nonlinearity than low-order MIMO systems, and that more capacity loss are experienced in multihop relaying over nonlinear channels as more relay hops are involved.

References

1. Stüber, G. L., Barry, J. R., Mclaughlin, S. W., Li, Y., Ingram M. A., Pratt, T. G.: Broadband MIMO-OFDM wireless communications. Proceedings of the IEEE, vol. 92, No. 2, pp. 271-294 (Feb. 2004)
2. Le L., Hossain, E.: Multihop cellular networks: Potential gains, research challenges, and a resource allocation framework. IEEE Commun. Mag., vol. 45, no. 9, pp. 66-73 (Sept. 2007)
3. Banelli, P., Baruffa, G., Cacopardi, S.: Effects of HPA non linearity on frequency multiplexed OFDM signals. IEEE Trans. Broadcast., vol. 47, no. 2, pp.123 -136 (2001)
4. Ahmad, I., Sulyman, A.I., Alsanie, A., Alasmari, A., Alshebeili, Saleh.: Spectral Broadening Effects of High Power Amplifiers in MIMO-OFDM Relaying Channels. EURASIP Journal on Wireless Communications and Networking 2013, 2013:32 doi:10.1186/1687-1499-2013-32 (2013)
5. Sabbaghian, M., Sulyman, A.I., Tarokh, V.: Analysis of the Impact of Nonlinearity on the Capacity of Communication Channels. IEEE Trans. Information Theory, vol.59, no.11, pp.7671-7683 (Nov. 2013)
6. Ahmad, I., Sulyman, A.I., Alsanie, A., Alasmari, A.: On the effect of amplifier non-linearity on the capacity of MIMO systems. in Proc. IEEE GCC Conf. Exhib. (GCC), vol., no., pp.108-111, 19-22 (Feb. 2011)
7. Qi, J., Aissa, S.: Analysis and compensation of power amplifier nonlinearity in MIMO transmit diversity systems. IEEE Trans. Veh. Technol., vol.59, no.6, pp.2921-2931 (July 2010)
8. Saleh, A.: Frequency-independent and frequency-dependent nonlinear models of TWT amplifiers. IEEE Trans. Communications, Vol. 29 (1981)
9. Rapp, C., Effects of the HPA-nonlinearity on a 4-DPSK/OFDM signal for a digital sound broadcasting system. Conf. Rec. ECSC',91 (1991)
10. Rowe, H. E.: Memoryless nonlinearities with gaussian inputs: Elementary results. Bell Syst. Tech. J., vol. 61, pp.1519 -1525 (1982)
11. Bussgang, J. J.: Cross-correlation function of amplitude-distorted Gaussian signals, (1952)
12. Helaly, T., Dansereau, R., El-Tanany, M.: BER performance of OFDM signals in presence of nonlinear distortion due to SSPA. Wireless Personal Commun., pp. 1–12 (2011)
13. Gregorio, F. H.: Analysis and compensation of nonlinear power amplifier effects in multi-antenna OFDM systems. Ph.D. dissertation, Helsinki Univ. Technol., Espoo, Finland, (Nov. 2007)
14. Fritz, O.: Hypergeometric functions. In M. Abramowitz & I. A. Stegun (Eds.), Handbook of mathematical functions with formulas, graphs, and mathematical tables. USA: U.S. Department of Commerce, (1972)
15. Jeffrey, A., Hui-Hui D.: Bessel functions. In Handbook of mathematical formulas and integrals. USA: Academic Press, (2008)
16. Hyundong, S., Hong, L. J.: Capacity of multiple-antenna fading channels: spatial fading correlation, double scattering, and keyhole. IEEE Trans. on Inform. Theory, vol.49, no.10, pp. 2636- 2647, (Oct. 2003)
17. Sulyman, A. I., Takahara, G., Hassanein, H., Kousa, M.: Multi-hop capacity of MIMO-multiplexing relaying systems. IEEE Trans. Wireless Commun, vol. 8, pp. 3095-3103, (2009)
18. Gradshteyn, I. S., Ryshik, I. M.: Table o f Integrals, Series, and Products, 6th ed. Academia Press, (2000)

Stability Analysis of Continuous Time Sigma Delta Modulators

Kyung Kang and Peter Stubberud

1 Introduction

Many electronic systems, including instrumentation, signal processing systems and portable communication systems, digitize analog signals using analog to digital converters (ADCs). Both continuous time (CT) and discrete time (DT) sigma delta modulator (ΣΔM) ADCs are often used for such applications because they use relatively simple, low power, analog circuitry and a low order quantizer in a feedback loop to achieve high speed, high resolution and low power signal conversion. Although CT ΣΔMs and DT ΣΔMs have many similarities such as the ability to increase their resolution by increasing their quantizer's sampling rate, increasing the number of bits in their ΣΔM's quantizers and increasing the orders of their loop filters, CT ΣΔMs have some advantages which include having inherent antialiasing filtering in the CT ΣΔM's signal transfer function (STF) and operating at higher frequencies than DT ΣΔMs because CT ΣΔMs don't have settling time requirements in their loop filters [1]; however, because DT ΣΔMs are entirely implemented using discrete time, or clocked, components while CT ΣΔMs are implemented using both analog and discrete components, DT ΣΔMs are simpler to analyze and simulate than CT ΣΔMs. Stability analysis in particular is more difficult to perform for CT ΣΔMs than it is for DT ΣΔMs.

Because a ΣΔM's output is typically the output of the ΣΔM's quantizer, ΣΔMs cannot be unstable in the bounded input bounded output (BIBO) sense. Instead, a ΣΔM is considered to have become unstable when the amplitude of a ΣΔM's input is increased over a value which causes the ΣΔM's output signal to quantization noise ratio (SQNR) to decrease dramatically and the ΣΔM's output SQNR cannot be restored to its previous values even when the ΣΔM's input is decreased to its previous amplitudes. Other phenomenon, such as input overload, can also cause a ΣΔM's output SQNR to decreased dramatically when the ΣΔM's input is increased over a certain value; however, in these cases, the ΣΔM's output SQNR can be restored to its previous values when the ΣΔM's input is decreased to its previous amplitudes.

In a general, a CT ΣΔM can be modeled by the canonical feedback loop shown in Fig. 1 where $X(s)$ and $Y(s)$ are the Laplace transforms of the input signal and the output signal, respectively, and $F(s)$, $G(s)$ and $H(s)$ are the system functions of the pre-filter stage, the feedforward path and the feedback path, respectively. The quantizer block represents a clocked quantizer, and the DAC block represents a digital to analog converter (DAC). The quantizer delay and DAC delay are often represented by a single delay block as they are in Fig. 1, and the combination of these two delays is often referred to as the excess loop delay.

Fig. 2 shows a linear model of the CT ΣΔM shown in Fig. 1 where the quantizer has been modeled by a variable gain, K, for the ΣΔM's STF, $Y(s)/X(s)$, and by an additive quantization noise, $E(s)$ for the ΣΔM's noise transfer function (NTF), $Y(s)/E(s)$. The ΣΔM's STF can be written as

$$STF(s) = \frac{K \cdot F(s) \cdot G(s)}{1 + K \cdot e^{-sD} \cdot G(s) \cdot H(s) \cdot DAC(s)} \quad (1)$$

and the ΣΔM's NTF can be written as

$$NTF(s) = \frac{1}{1 + e^{-sD} \cdot G(s) \cdot H(s) \cdot DAC(s)} \quad (2)$$

where the exponential function, e^{-sD} is the Laplace transform of the excess loop delay, D. Although the DAC is not explicitly modeled, a typical zero order hold (ZOH) DAC would have the system function

K. Kang (✉) • P. Stubberud
Department of Electrical and Computer Engineering, University of Nevada, Las Vegas, USA
e-mail: kangk3@unlv.nevada.edu; peter.stubberud@unlv.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_71,

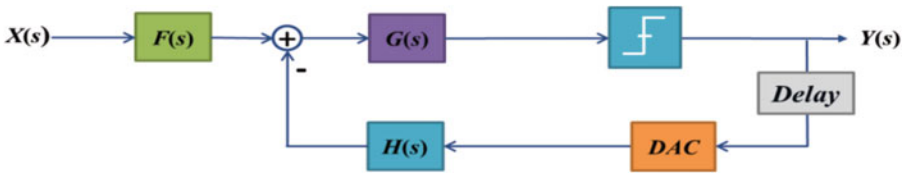


Fig. 1 The canonical form of the feedback system

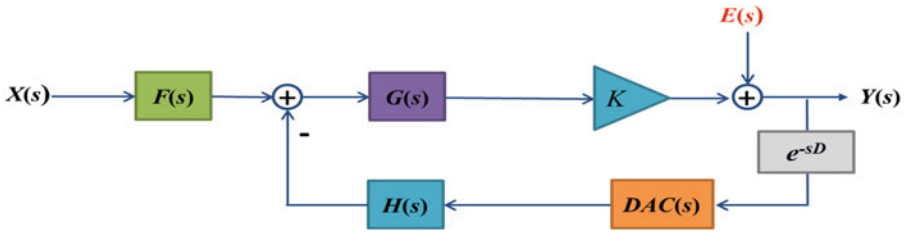


Fig. 2 A linear model of the CT ΣΔM

$$DAC(s) = \frac{1 - e^{-sT}}{sT} \tag{3}$$

where T is the ΣΔM's sampling period. Most other DACs are also typically modeled using exponential functions.

Root locus methods have been successfully used to determine the stability of DT ΣΔMs; however, because the denominator terms of both the STF and NTF contain exponential functions, traditional root locus methods cannot be used for determining the stability of CT ΣΔMs. Instead, several other methods have been developed for predicting the stability of CT ΣΔMs. One such method models the nonlinear quantizer using two linear gains, one for the signal gain and one for quantization noise gain [2]. This approach has not received much attention because of its complexity and because it cannot predict stability for several classes of ΣΔMs. Other approaches predict CT ΣΔM stability by assuming that the ΣΔM's has a DC input and then by performing a simple stability analysis. These methods are effective for predicting stability for lower order ΣΔMs but not for higher order ΣΔMs [3–9]. Another method attempts to determine ΣΔM's stability by using a one-norm of the ΣΔM's NTF to determine stability in a BIBO sense. It has been shown that the one-norm condition is available only for second order lowpass modulators [10]. Therefore, a mixture of one-norm, two-norm and infinity-norm constraints have been proposed to predict the stability of higher order modulators [11]. Lee's rule is another method used to determine the stability of ΣΔMs [12]. Lee's rule states that a ΣΔM will be stable if the gain of the ΣΔM's NTF is less than two for all frequencies. It has been shown that Lee's rule is neither a necessary nor a sufficient condition to ensure stability in ΣΔMs [13].

In this paper, an analytical root locus method is used to determine the stability criteria for CT ΣΔMs that include exponential functions in their characteristic equations. This root locus method determines the range of quantizer gains for which a CT ΣΔM is stable. These values can then be used to determine input signal and internal signal ranges that prevent ΣΔM from becoming unstable. A circuit designer can then take measures to prevent the ΣΔM from becoming unstable. Examples of 3rd order CT ΣΔMs illustrate this method.

2 ΣΔM Stability Analysis using an Analytical Root Locus Method

Root locus analysis is a method for examining how the poles of a system change as function of a certain system parameter. This method is commonly used to determine the stable region of feedback systems as a function of open loop gain by plotting the poles of the system's closed loop transfer function as a function of the system's open loop gain. As shown in (1), CT ΣΔMs typically have characteristic equations of the form

$$1 + K \cdot e^{-sD} \cdot G(s) \cdot H(s) \cdot DAC(s) = 0 \tag{4}$$

where D is the ΣΔM's excess loop delay and $DAC(s)$ contains at least one exponential function.

When $D = 0$ and $DAC(s) = 1$, root locus analysis of the characteristic equation in (4) can be performed using standard graphical analysis methods [14] or using an analytical method [15,16]. When $D \neq 0$ and $DAC(s)$ contains at least one exponential function, root locus analysis of the

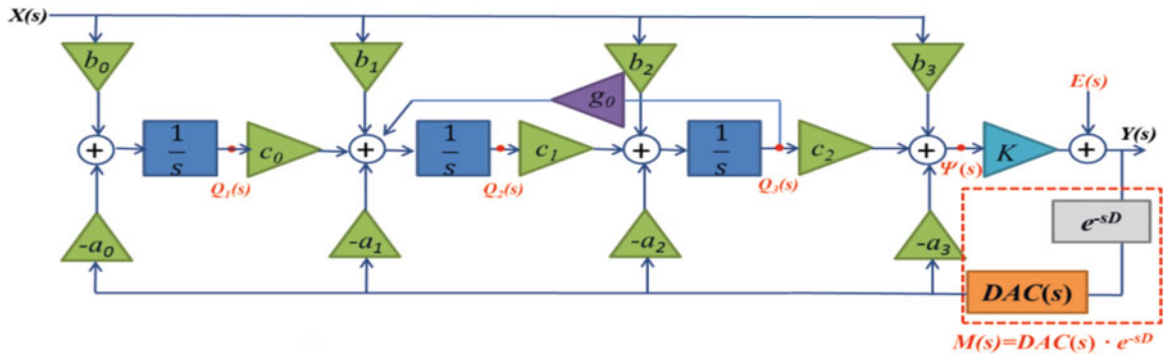


Fig. 3 3rd order low pass CT ΣΔ block diagram

characteristic equation in (4) can be performed using an extended graphical analysis method [17–19] or using the analytical method in [20]. In this paper, the analytical method in [20] is used to determine the quantizer gains that allow CT ΣΔM to remain stable.

To illustrate this method, the term $e^{-sD}\cdot G(s)\cdot H(s)\cdot DAC(s)$ in (4) is written as

$$e^{-sD}\cdot G(s)\cdot H(s)\cdot DAC(s)=\frac{N(s)}{D(s)} \tag{5}$$

which implies that (4) can be written as

$$D(s)+K\cdot N(s)=0\,. \tag{6}$$

Solving (6) for K,

$$K=-\frac{D(s)}{N(s)}=-\frac{\mathrm{Re}\{D(s)\}+j\mathrm{Im}\{D(s)\}}{\mathrm{Re}\{N(s)\}+j\mathrm{Im}\{N(s)\}}\,. \tag{7}$$

In standard form, (7) can be written as

$$K=\frac{-\mathrm{Re}\{D(s)\}\cdot\mathrm{Re}\{N(s)\}-\mathrm{Im}\{D(s)\}\cdot\mathrm{Im}\{N(s)\}}{(\mathrm{Re}\{N(s)\})^2+(\mathrm{Im}\{N(s)\})^2} +j\frac{\mathrm{Re}\{D(s)\}\cdot\mathrm{Im}\{N(s)\}-\mathrm{Im}\{D(s)\}\cdot\mathrm{Re}\{N(s)\}}{(\mathrm{Re}\{N(s)\})^2+(\mathrm{Im}\{N(s)\})^2} \tag{8}$$

Because the quantizer's variable gain, K, is real, (8) implies that

$$K=\frac{-\mathrm{Re}\{D(s)\}\cdot\mathrm{Re}\{N(s)\}-\mathrm{Im}\{D(s)\}\cdot\mathrm{Im}\{N(s)\}}{(\mathrm{Re}\{N(s)\})^2+(\mathrm{Im}\{N(s)\})^2} \tag{9}$$

and that

$$\mathrm{Re}\{D(s)\}\cdot\mathrm{Im}\{N(s)\}-\mathrm{Im}\{D(s)\}\cdot\mathrm{Re}\{N(s)\}=0\,. \tag{10}$$

Plotting (10) in the s-plane renders the root locus of (4) for $-\infty<K<\infty$.

3 Examples

ΣΔMs achieve high resolution by using a feedback loop to attenuate quantization noise in the frequency band of interest while passing the input signal to the output. Because of the importance of attenuating the quantization noise over the frequency band of interest, the ΣΔM's NTF is designed before the STF. After determining an NTF and STF, the NTF and STF coefficients need to be implemented in a hardware structure, such as a cascade of resonators feedback (CRFB), cascade of resonators feedforward (CRFF), cascade of integrator feedback (CIFB), and cascade of integrator feedforward (CIFF) implementations. In the following example, 3rd order Chebyshev Type 2 NTFs are implemented using a CIFB implementation.

Fig. 3 shows the linear model of a 3rd order lowpass CT ΣΔM using a CIFB implementation. In Fig. 3, the signals, $X(s)$, $E(s)$ and $Y(s)$, are the Laplace transforms of the input signal, the quantization noise signal and the output signal, respectively. The quantizer has been modeled by a variable gain, K, an additive quantization noise, $E(s)$, and a delay. The DAC block represents a digital to analog converter (DAC), and the excess loop delay is represented by the *Delay* block as it is in Fig. 2. The blocks with the symbols, a_0, a_1, a_2, a_3, b_0, b_1, b_2, b_3, c_0, c_1, c_2 and g_0 represent scalar multiplication with gains associated by the blocks' respective symbols.

The STF and NTF of the ΣΔM shown in Fig. 3 can be calculated from the block diagram by calculating the states and the output as

$$Q_1(s)=\{b_0X(s)-a_0\cdot M(s)\cdot Y(s)\}\frac{1}{s} \tag{11}$$

$$Q_2(s)=\{b_1X(s)-a_1\cdot M(s)\cdot Y(s)+c_0Q_1(s)+g_0Q_3(s)\}\frac{1}{s} \tag{12}$$

$$Q_3(s)=\{b_2X(s)-a_2\cdot M(s)\cdot Y(s)+c_1Q_2(s)\}\frac{1}{s} \tag{13}$$

$$\Psi(s) = b_3X(s) - a_3 \cdot M(s) \cdot Y(s) + c_2Q_3(s) \tag{14}$$

$$Y(s) = K \cdot \Psi(s) + E(s) \tag{15}$$

where $M(s)=DAC(s)\cdot e^{-sD}$. Substituting (11) into (12), (12) into (13), (13) into (1414) and (14) into (15), the STF and the NTF can be written as

$$STF(s) = \frac{\frac{K\cdot\left\{b_3s^3+b_2c_2s^2+(b_1c_1c_2-g_0b_3c_1)s+b_0c_0c_1c_2\right\}}{s(s^2-g_0c_1)}}{1 + \frac{K\cdot M(s)\cdot\left\{a_3s^3+a_2c_2s^2+(a_1c_1c_2-g_0a_3c_1)s+a_0c_0c_1c_2\right\}}{s(s^2-g_0c_1)}} \tag{16}$$

and

$$NTF(s) = \frac{1}{1 + \frac{M(s)\cdot\left\{a_3s^3+a_2c_2s^2+(a_1c_1c_2-g_0a_3c_1)s+a_0c_0c_1c_2\right\}}{s(s^2-g_0c_1)}} \cdot \tag{17}$$

Comparing (1) with (16) and (2) with (17), it can be seen that

$$F(s) = b_3s^3 + b_2c_2s + (b_1c_1c_2 - g_0b_3c_1)s + b_0c_0c_1c_2 \tag{18}$$

$$G(s) = \frac{1}{s(s^2 - g_0c_1)} \tag{19}$$

$$H(s) = a_3s^3 + a_2c_2s^2 + (a_1c_1c_2 - g_0a_3c_1)s + a_0c_0c_1c_2\ . \tag{20}$$

Assuming that the DAC is implemented using a ZOH, $DAC(s)$ can be written as shown in (3). The gains, a_0, a_1, a_2, a_3, b_0, b_1, b_2, b_3, c_0, c_1, c_2 and g_0 can be determined by equating the STF coefficients in (16) and the NTF coefficients in (17) with the desired STF and NTF coefficients, respectively. For example, if the NTF is a high pass Chebyshev Type 2 filter and the STF has a lowpass characteristic, then the rational function describing the Chebyshev Type 2 filter is set equal to the rational NTF in (17) and the rational function describing the lowpass STF is set equal to the rational STF in (16).

Using (5), (19), (20) and (3), $N(s)$ and $D(s)$ can be determined to be

$$N(s) = \left\{a_3s^3 + a_2c_2s^2 + (a_1c_1c_2 - g_0a_3c_1)s + a_0c_0c_1c_2\right\} \cdot \left(1 - e^{-sT}\right) \cdot e^{-sD} \tag{21}$$

and

$$D(s) = \left(s^2 - g_0c_1\right) \cdot s^2 \cdot T\ . \tag{22}$$

Substituting $\sigma + j\omega$ for s where $\sigma = \mathrm{Re}\{s\}$ and $\omega = \mathrm{Im}\{s\}$,

$$\mathrm{Re}\{D(s)\} = T \cdot \sigma\left(\sigma^3 + a\sigma - 3\sigma\omega^2\right) - T \cdot \omega\left(-\omega^3 + a\omega + 3\sigma^2\omega\right) \tag{23}$$

$$\mathrm{Im}\{D(s)\} = T \cdot \omega\left(\sigma^3 + a\sigma - 3\sigma\omega^2\right) + T \cdot \sigma\left(-\omega^3 + a\omega + 3\sigma^2\omega\right) \tag{24}$$

$$\mathrm{Re}\{N(s)\} = (b\sigma^2 - b\omega^2 + c\sigma + d)\left\{e^{-D\sigma}\cos(D\omega) - e^{-(T+D)\sigma}\cos((T+D)\omega)\right\} - (2b\sigma\omega + c\omega)\left\{-e^{-D\sigma}\sin(D\omega) + e^{-(T+D)\sigma}\sin((T+D)\omega)\right\} \tag{25}$$

and

$$\mathrm{Im}\{N(s)\} = (b\sigma^2 - b\omega^2 + c\sigma + d)\left\{-e^{-D\sigma}\sin(D\omega) + e^{-(T+D)\sigma}\sin((T+D)\omega)\right\} + (2b\sigma\omega + c\omega)\left\{e^{-D\sigma}\cos(D\omega) - e^{-(T+D)\sigma}\cos((T+D)\omega)\right\}\ . \tag{26}$$

Substituting (23), (24), (25) and (26) into (10), the root locus of the 3rd order CT ΣΔM shown in Fig. 3 can be plotted for $-\infty < K < \infty$.

Fig. 4 (*a*), (*b*) and (*c*) show the plot of (10), or the root locus, for 3rd order CT ΣΔMs with a sampling frequency, f_s where $f_s=1/T$, of 1GHz and Chebyshev Type 2 NTF with 47dB, 37dB, and 30dB attenuation in the stopband for $D = 0$, $D = T/2$, and $D = T$, respectively. The plots in Fig. 4 include both the positive gain ($K > 0$) root locus and the negative gain ($K < 0$) root locus. According to the plots, the CT ΣΔMs with $D = 0$, $D = T/2$, and $D = T$ are stable for $0.326 < K < 3.734$, $0.359 < K < 2.001$, and $0.376 < K < 1.648$, respectively. Although the root locus plots show that the CT ΣΔMs with $D = 0$, $D = T/2$, and $D = T$ are unstable for $K > 3.734$, $K > 2.001$, and $K > 1.648$, respectively, none of the modulators show a degradation in SQNR when K enters those ranges because when the modulator enters unstable regions for large values of quantizer gain, K, the feedback signal increases which reduces the quantizer gain, K, and moves the poles back into a stable region. However, when $K < 0.326$, $K < 0.359$, and $K < 0.376$ for the CT ΣΔMs with $D = 0$, $D = T/2$, and $D = T$, respectively, the modulator shows a degradation in SQNR because when the modulator enters those unstable regions the feedback signal increases which further reduces the quantizer gain, K, and consequently moves the poles further from the stable region. Therefore, a CT ΣΔM will remain stable if its quantizer gain, K, remains above its minimum value, $K_{\min}$, as determined from its CT ΣΔMs root locus plot.

Using this stability criterion, the maximum quantizer input, $\psi_{\max}$, that prevents the modulator from becoming

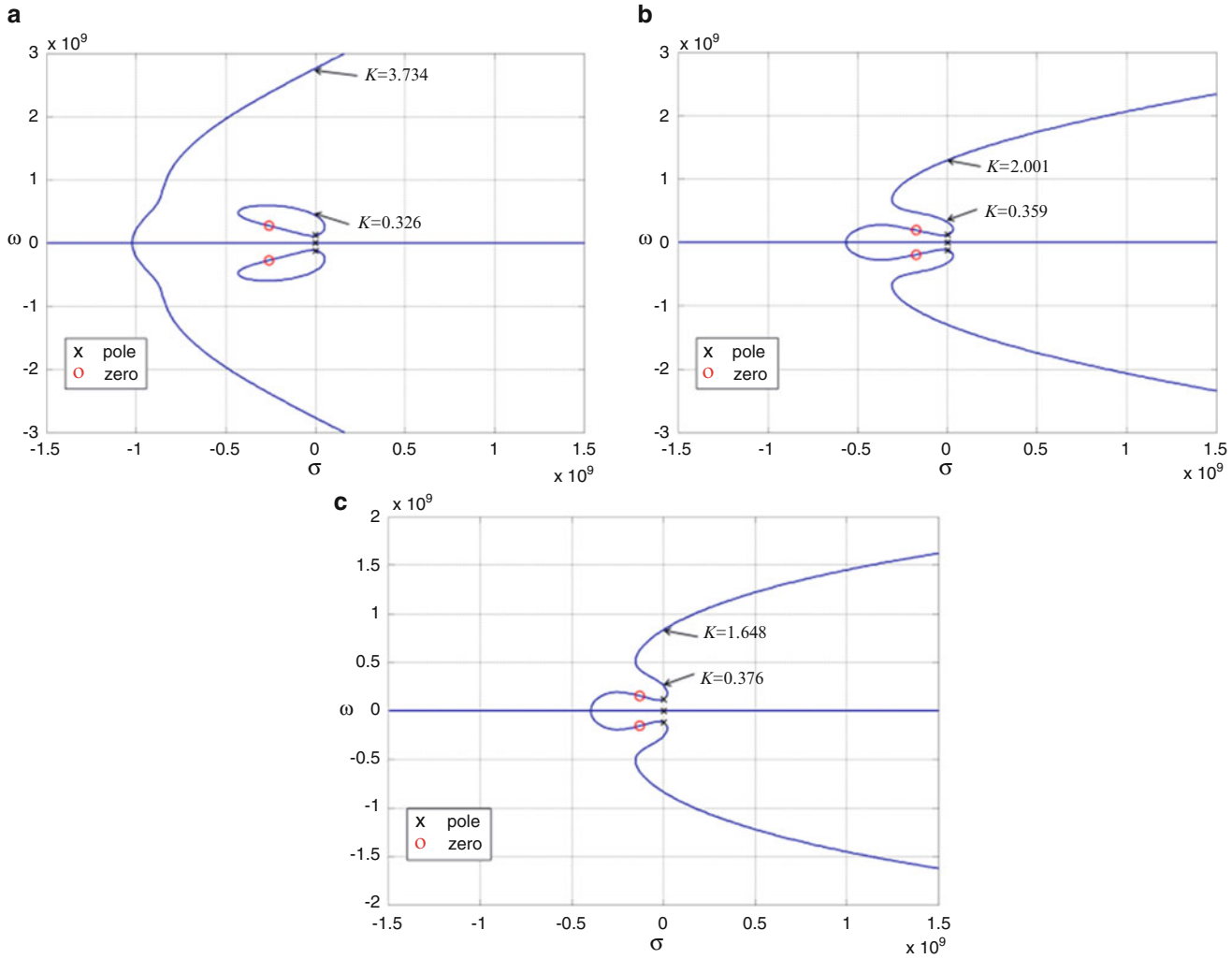


Fig. 4 The root locus of 3rd order CT ΣΔMs that uses Chebyshev Type 2 NTFs and a sampling frequency of 1GHz for (*a*) $D = 0$ (*b*) $D = T/2$ (*c*) $D = T$.

unstable is $1/K_{\min}$. Therefore, the CT ΣΔMs with $D = 0, D = T/2$, and $D = T$ are unstable when $\psi_{\max} > 3.07$, $\psi_{\max} > 2.78$, and $\psi_{\max} > 2.66$, respectively.

Assuming that the ΣΔM's input, output and quantization noise signals have means of zero, the maximum power, σ^2_x, of the ΣΔM's input signal can be calculated from

$$\sigma_y^2 = \sigma_x^2 \int_{-fs/2}^{fs/2} |STF(f)|^2 df + \frac{\sigma_e^2}{f_s} \int_{-fs/2}^{fs/2} |NTF(f)|^2 df \quad (27)$$

where σ^2_e is the quantization noise power when the maximum input signal power is applied to the input and σ^2_y is the output signal power which equals one. If the input signal is sinusoidal, then $\sigma^2_x = x_{\max}{}^2/2$ where $x_{\max}$ is the ΣΔM's maximum input amplitude. Assuming a sinusoidal input, the maximum input values are 0.432 (-7.2 dB), 0.447 (-7.0 dB), and 0.376 (-8.5 dB) for the CT ΣΔMs with $D = 0, D = T/2$, and $D = T$, respectively.

All three CT ΣΔMs were simulated using the method described in [21]. Fig. 5(*a*) shows SQNR and the minimum gain, $K_{\min}$, as a function of input signal amplitude for the ΣΔMs with $D = 0$. As shown in the Fig. 5(a), the ΣΔM's SQNR increases linearly until the input signal's amplitude is less than -8 dB, or $K_{\min} = 0.391$. As the input signal's amplitude is increased above -8 dB, the ΣΔM's SQNR no longer increases linearly. As the input signal's amplitude is increased above -6 dB, the ΣΔM's SQNR degrades dramatically and the ΣΔM's SQNR cannot be restored to its previous values even when the ΣΔM's input is decreased to its previous amplitudes. Fig. 5 (*b*) and (*c*) show the SQNR and the minimum gain, $K_{\min}$, as a function of input signal amplitude for the ΣΔMs with $D = T/2$ and $D = T$, respectively. From Fig. 5 (*b*), when the input signal's amplitude is increased above -7 dB, or $K_{\min}$ is less than 0.359, the ΣΔM's SQNR begins to degrade. Similarly, the ΣΔM's SQNR with $D =$

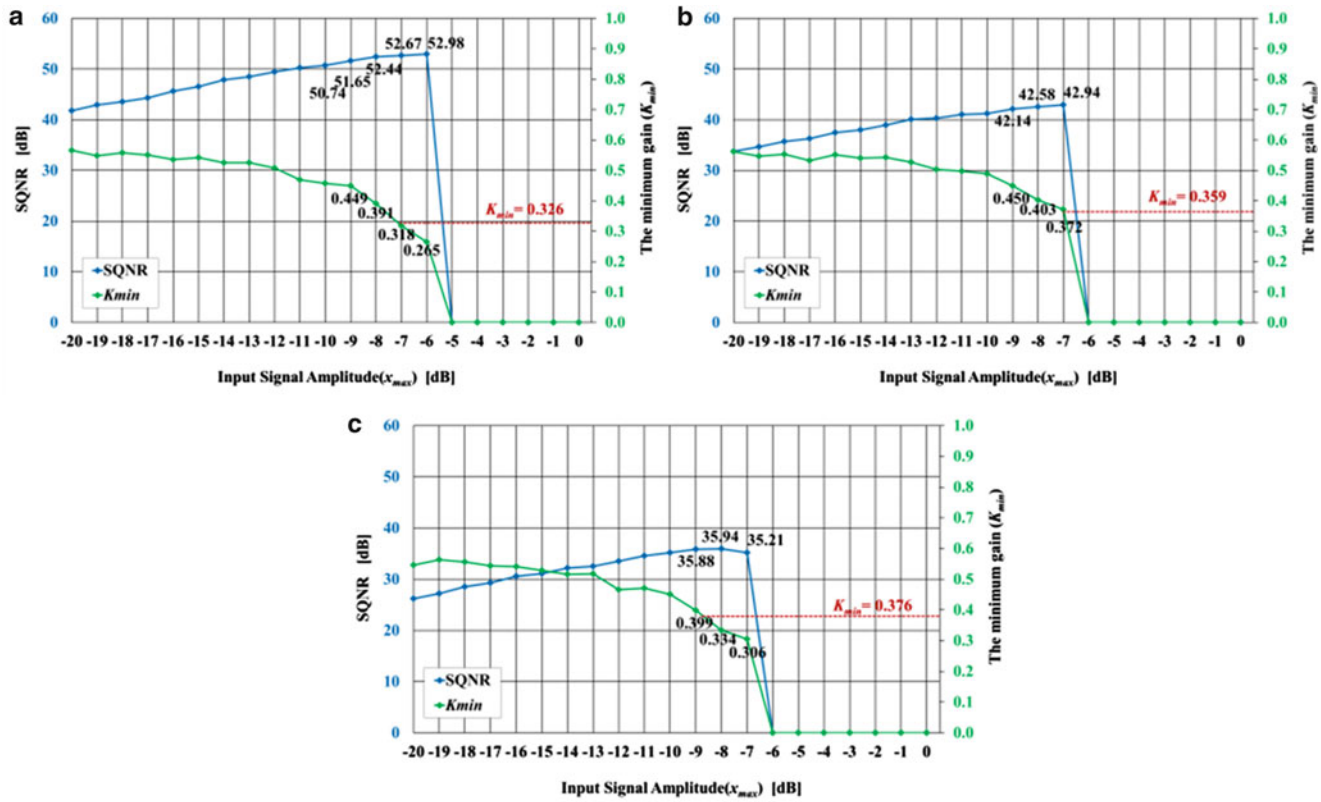


Fig. 5 Simulated SQNR and the minimum gain ($K_{\min}$) for the 3rd order CT ΣΔM for (*a*) $D = 0$ (*b*) $D = T/2$ (*c*) $D = T$

T is degraded when the ΣΔM's input is greater than -9 dB, or $K_{\min}$ is less than 0.376 as shown in Fig. 5 (*c*).

4 Conclusion

In this paper, an analytical root locus method was used to determine the minimum quantizer gains that keep a CT ΣΔM stable. It was then shown how to use the minimum quantizer gain to determine the maximum quantizer input amplitude and ΣΔM maximum input amplitude value that prevent ΣΔM from becoming unstable. Using these values, a circuit designer can then take measures to prevent the ΣΔM from becoming unstable. Examples of 3rd order CT ΣΔM illustrate this method.

References

1. J. A. Cherry and W. M. Snelgrove.: Continuous-Time Delta-Sigma Modulators for High Speed A/D Conversion Theory, Practice and Fundamental Performance Limits: Kluwer Academic, New York, NY, USA (2002)
2. S. H. Ardalan, and J.J. Paulos.: An analysis of non-linear behaviour in Σ-Δ modulators. In: IEEE Trans. on Circuits and Syst., vol. CAS-34, no. 6, pp.1157-1162 (1987)
3. N. Wong, and N.G. Tung-Sang.: DC stability analysis of higher-order, lowpass sigma-delta modulators with distinct unit circle NTF zeroes. In: IEEE Trans. on Circuits & Syst.-II, vol. 50, issue 1, pp. 12-30 (2003)
4. J. Zhang, P.V. Brennan, D. Juang, D, E. Vinogradova, and P.D. Smith.: Stable analysis of a sigma-delta modulator. In: Proc. IEEE Int. Symp. Circuits Syst., vol.1, pp.1-961-1-964 (2003)
5. P, Steiner, and W. Yang.: Stability analysis of the second-order sigma-delta modulator. In: Proc. IEEE Int. Symp. Circuits Syst., vol. 5, pp. 365-368 (1994)
6. J. Zhang, P.V. Brennan, D. Juang, E. Vinogradova, and P.D. Smith.: Stable boundaries of a 2nd-order sigma-delta modulator. In: Proc. South. Symp.Mixed Signal Design (2003)
7. N. A. Fraser, and B. Nowrouzian.: A novel technique to estimate the statistical properties of sigma-delta A/D converters for the investigation of DC stability. In: Proc. IEEE Int. Symp. Circuits Syst., vol.3, pp.111-289-111-292 (2002)
8. D. Reefman, J.D. Reiss, E. Janssen, and M.B. Sandler.: Description of limit cycles in sigma-delta modulators. In: IEEE Trans. on Circuits and Syst-I., vol. 52, issue 6, pp.1211 – 1223 (2005)
9. S. Hein, and A. Zakhor.: On the stability of sigma-delta modulators. In: IEEE Trans. on Signal Processing, vol. 41, no.7, pp. 2322-2348 (1993)
10. 10 R. Schreier and W. M. Snelgrove.: Bandpass Sigma-Delta modulator. In: Electronics letters, vol. 25, no. 23, pp. 1560-1561 (1989)
11. Lars Risbo.: FPGA Based 32 Times Oversampling 8th-order Sigma-Delta Audio DAC. In: Proc. 96th AES Convention, Preprint # 3808 (1994)
12. W. L. Lee and C. G. Sodini.: A topology for higher interpolative coders. In: Proceedings of ISCAS, pp. 459-462 (1987)

13. R. Schreier and G.C. Temes.: Understanding delta sigma data converters: Wiley-IEEE express (2004)
14. John, B.: Essentials of control techniques and theory, CRC pressInc (2009)
15. Bendrukov,G.A, Teodorchik,K,F: The analytic theory of constructing root loci. In: automation and remote control, vol. 20, pp. 340-344 (1959)
16. Cogan,B.:Use of the analytic method and computer algebra to plot root loci. In: International journal of electrical engineering education. vol. 35, pp. 350-356 (1998)
17. D'azzo, J.J, Houpis, C.H.: Linear control system analysis and design conventional and modern, Mc Graw Hill, New York (1988)
18. Ogata, K.: Modern control engineering, Prentice Hall (2002)
19. Palm,W.J.:Control system engineering, John Wiley & Sons (1986)
20. Cogan.B, Paor.A.M.: Analytic root locus and LAMBERT W function in control of a process with time delay. In: Journal of electrical engineering, vol. 62, pp. 327-334 (2011)
21. K.Kang, P.Stubberud: A comparison of continuous time sigma delta modulator simulation methods. In: IEEE International Midwest Symposium on Circuits and Systems (2014)

An Area Efficient Weighting Coefficient Generation Architecture for Polynomial Convolution Interpolation

D. Selvathi and C. John Moses

1 Introduction

Image interpolation technique is a commonly used scheme in image processing, medical imaging, and computer graphics [1] which is used to construct new data points within the range of a discrete set of known data points. [2]. Interpolation processes are transformations between two regularly sampled grids, one at the input resolution, and another at output resolution [3]. A variety of applications require image zooming, such as digital cameras, electronic publishing, third-generation mobile phones, medical imaging, and image processing [4]. Image interpolations based on estimates of the model sinc kernel (pixel replication, bilinear, bi-cubic, and higher-order splines) are normally used for their flexibility and speed, however, these methods may produce blurring, ringing artifacts, jagged edges, and abnormal depiction (curves of substance intensity) [5]. This kind of problems can be avoided by using the sinc-approximating kernel to the image being interpolated. Nowadays, different kinds of interpolation kernels are used for scaling the image by either down-sampling or up-sampling the image for improving the image quality [6]. The algorithms used for resampling digital images are broadly classified as non adaptive interpolation algorithms and adaptive interpolation algorithms. In non-adaptive interpolation scheme, linear and fixed pattern of computation is applied in every pixel. This technique is fixed irrespective of the input image features and has low computational complexity. In adaptive interpolation scheme, non linear type of computation is applied based on sharp edges and smooth texture [7, 8]. This technique is computationally inefficient and expensive.

Convolution based interpolation scheme is the most commonly used non adaptive interpolation scheme for digital image scaling. In Convolution based interpolation scheme the input image is multiplied with the convolution kernel to find the resized image. Generally, the quality of the interpolated image depends upon the kernel used for the interpolation [8]. Therefore, the numerical accuracy and computational cost of interpolation algorithms are directly tied to the interpolation kernel. As a result, interpolation kernels are the target of design and analysis. So a kernel of higher order and with simpler weighting coefficient circuit is preferred to generate high quality images with low computational complexity.

The two criteria which are essential to produce high quality images are the order of kernel and the complexity in generating weighting coefficients [9]. The order of kernel is increased, the quality of the scaled image is increased, at the same time it increases the computational burden of generating weighting coefficients. By decreasing the number of components in the weighting coefficient circuit the computational complexity such as area and cost can be reduced.

Keys image interpolation method is one of the widely used cubic convolution interpolations. The kernel of this interpolation is of third order. The weights are produced by means of this kernel. The effect achieved by this kernel is better than the nearest neighbor image interpolation and bi-linear image interpolations [10]. But this kernel function has more computational complexity and it is not able to enhance the scaling result [11].

The distinctive method of cubic convolution image interpolation group is bi-cubic convolution image interpolation [12]. To develop the effectiveness of bi-cubic interpolation, Lin et al. demonstrated a very-large-scale integration (VLSI) design of bi-cubic convolution interpolation for digital image processing [13]. Its architecture diminishes the computational complexity of producing weighting

D. Selvathi (✉)
ECE Department, Mepco Schlenk Engineering College, Sivakasi, India
e-mail: dselvathi@gmail.com

C.J. Moses
ECE Department, St. Xavier's Catholic College of Engineering, Nagercoil, India
e-mail: jofjef@yahoo.com

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_72,

coefficients and the amount of memory access times. Though, the kernel of cubic convolution image interpolations still needs complex computations.

A better quality interpolation scheme does not make interpolated image distortion, nor does it need complex computation. To defeat the above said shortcomings FFOPCI is presented [10]. FFOPCI is one of the non adaptive interpolation algorithms for digital image scaling. The kernel of FFOPCI method is developed with third-order approximation of normal interpolation kernel. This scheme has the advantages of low operation complexity, weighting coefficient generation effort, and hardware cost. Consequently, this architecture has solved the problem of blur and blocking effect and enhances the quality of image. Still there exists complexity in generating the weighting coefficients.

This work aims to optimize the WCG of FFOPCI by reducing number of arithmetic elements for generating weighting coefficient. By reducing number of arithmetic elements of WCG, the interpolator can perform fast with less chip area. This work can be evaluated by using field programmable gate array (FPGA).

FPGAs are widely used for rapid prototyping of digital signal processing (DSP) systems [14]. FPGA technology is used to improve performance, while providing programmability and dynamic reconfigurability. FPGAs have millions of gates, reasonably on-chip memory and fast input-output interface. Therefore, FPGAs can provide an easy and cost effective way to evaluate image processing algorithms from an implementation perspective. The proposed algorithm is synthesized using Xilinx ISE for Xilinx Virtex-6-xc6vsx315t-3ff1156 FPGA.

This paper is organized as follows. Chapter 2 describes the conventional architecture of image interpolation algorithm. Chapter 3 presents the fast first order polynomial convolution interpolation (FFOPCI). Chapter 4 introduces the proposed low complexity architecture of WCG. Chapter 5 provides experimental results and comparison. Finally, chapter 6 offers the conclusion.

2 Conventional Architecture of Image Interpolation

The block diagram of the hardware architecture for digital image scaling is shown in figure 1 which includes Coordinate Calculation Unit, Memory Bank, WCG, Vertical and Horizontal Interpolation units and Virtual Pixel Buffer [15].

The coordinate calculation unit includes interpolated coordinate accumulator, row/column address calculator and vertical and horizontal distance calculator. The coordinate of the interpolated point $Q(x_n, y_m)$ is obtained in the interpolation coordinate calculator.

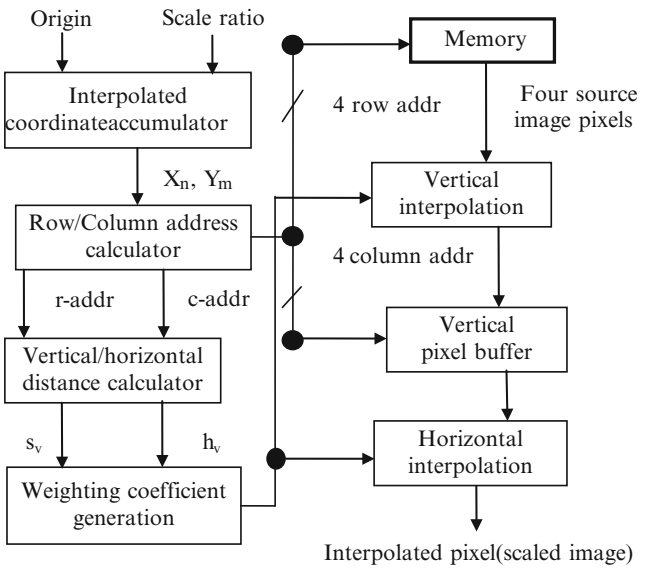


Fig. 1 The block diagram of the hardware architecture for digital image scaling

In the circuit of row/column address calculator, the operation of vertical or horizontal address orientation is controlled by the signal of vertical/horizontal. This signal determines the vertical (y_m) and the horizontal (x_n) coordinates. If the signal vertical/horizontal is vertical, then the row addresses and the vertical interval s_v can be obtained. Otherwise the column addresses and the horizontal interval h_v can be found.

To simplify the weighting coefficient generation, the distance in vertical direction and horizontal direction has to be found before the calculation of vertical and horizontal weighting coefficients. The vertical distance calculator calculates distance between the source pixel and the virtual interpolated pixel. Similarly, horizontal distance calculator calculates distance between the virtual interpolated pixel and the final interpolated pixel.

The image which has to be scaled is stored in memory bank. The pixels are retrieved from the memory bank based upon the generated row and column address.

The most important computation in convolution based scaling is the calculation of interpolation weighting coefficients. The Weighting Coefficient Generator, as depicted in figure 2 is designed for producing vertical and horizontal weighting coefficients. The vertical and horizontal interpolations have the same operation, but they have to execute in parallel to accelerate scaling speed. The vertical interpolation unit performs interpolation in column wise manner to produce virtual pixel. The horizontal interpolation unit performs interpolation for these virtual pixels to produce interpolated pixel. The virtual pixels created from vertical interpolation are stored in the virtual pixel buffer,

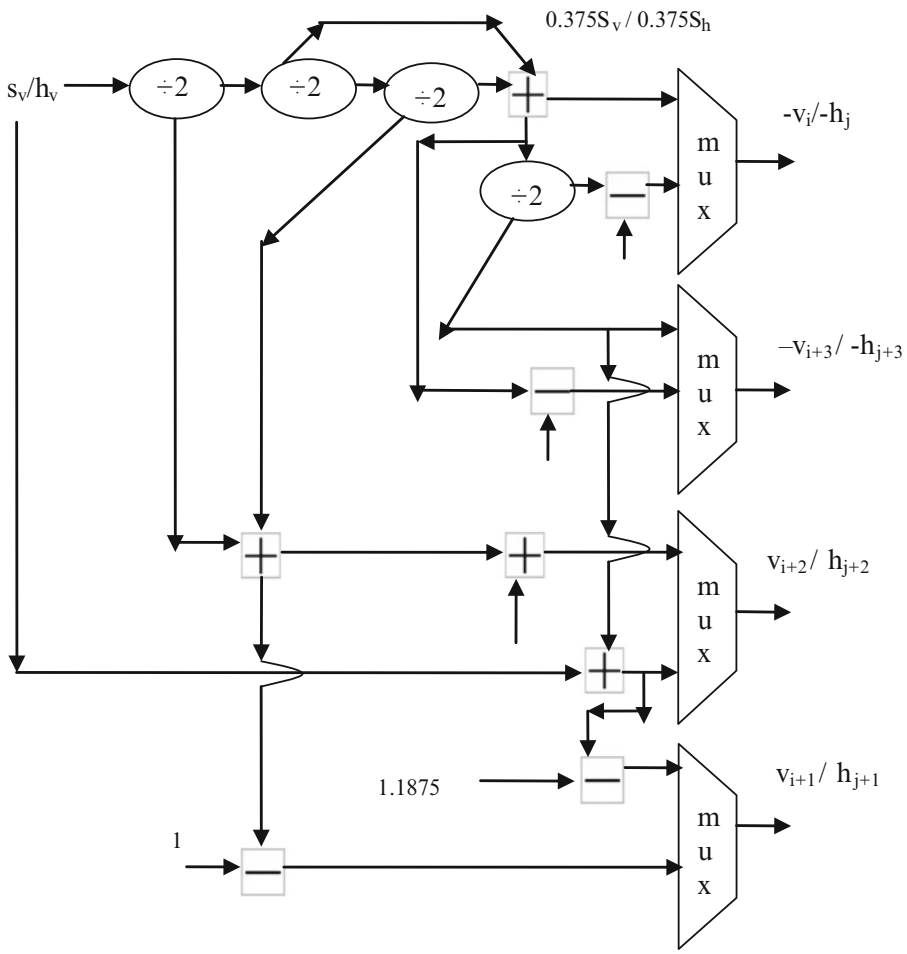


Fig. 2 Weighting coefficient generation circuit of FFOPCI

as shown figure 1 where to be accessed in the process of horizontal interpolation.

3 Fast First Order Polynomial Convolution Interpolation

FFOPCI is a high-performance image scaling method, with high excellence and diminished calculation complexity of generating weights. The kernel of fast FFOPCI was made on third order interpolation kernel [10]. The FFOPCI requires 16 source pixels that around an interpolated point of a source image.

The coordinate of the interpolated point is obtained in the interpolation coordinate calculator. The row addresses and column addresses of its 16 neighboring source pixels then can be determined. Afterward, the values of the 16 source pixels in off-chip memory can be obtained and FFOPCI can proceed.

The FFOPCI method estimates the ideal sinc-function in the distance [−2, 2]. Therefore, the novel kernel is gained. Also the polynomial of the novel kernel can be characterized by the equation (1).

$$k_p(s) = \begin{cases} 1 - (1 + \frac{4\alpha}{9})s, & 0 \le s < \frac{1}{3} \\ (1 - \frac{2\alpha}{9}) - (1 - \frac{2\alpha}{9})s, & \frac{1}{3} \le s < 1 \\ -\frac{4\alpha}{9} + \frac{4\alpha}{9}s, & 1 \le s < \frac{4}{3} \\ \frac{4\alpha}{9} + \frac{2\alpha}{9}s, & \frac{4}{3} \le s < 2, \end{cases} \tag{1}$$

Here α is sharpness parameter [10]. The excellence of the interpolated images is straightly attached to this sharpness parameter. By rightly tuning this parameter, elevated excellence images can be gained. The optimal value of α can be

derived by applying the sum of standard deviation between -1 and -0.5. This standard deviation can approximate an ideal sinc function. Generally, the minimum standard deviation is obtained at $\alpha = -0.853$. Therefore, the kernel of the FFOPCI for $\alpha = -0.853$ can be obtained by using equation (2).

$$k_{p_{-0.853}}(S) = \begin{cases} 1 - 0.6209s, & 0 \le s < \frac{1}{3} \\ 1.1896 - 1.1896s, & \frac{1}{3} \le s < 1 \\ 0.3791 - 0.3791s, & 1 \le s < \frac{4}{3} \\ -0.3791 + 0.1896s, & \frac{4}{3} \le s < 1 \\ 0, & otherwise \end{cases} \quad (2)$$

In 1-dimensional image interpolation, vertical as well as horizontal weighting coefficients are required to be determined. The vertical weighting coefficients are determined by using vertical distance (s_v) in equation (2) as:

$$v_i(1+s_v) = \begin{cases} -0.375s_{v'} & 0 \le s_v < \frac{1}{3} \\ -0.1875 + 0.1875s_{v'} & \frac{1}{3} \le s_v < 1 \end{cases}$$

$$v_{i+1}(s_v) = \begin{cases} 1 - 0.625s_{v'} & 0 \le s_v < \frac{1}{3} \\ 1.1875 - 1.1875s_{v'} & \frac{1}{3} \le s_v < 1 \end{cases}$$

$$v_{i+2}(1+s_v) = \begin{cases} 1.1875s_{v'} & 0 \le s_v < \frac{2}{3} \\ 0.375 + 0.625s_{v'} & \frac{2}{3} \le s_v < 1 \end{cases}$$

$$v_{i+3}(2+s_v) = \begin{cases} -0.1875s_{v'} & 0 \le s_v < \frac{2}{3} \\ -0.375 + 0.375s_{v'} & \frac{2}{3} \le s_v < 1 \end{cases} \quad (3)$$

where v_i, v_{i+1}, v_{i+2}, and v_{i+3} are the vertical weighting coefficients of the consequent source pixels $A_{i,\ j}$, $A_{i+1,\ j}$, $A_{i+2,\ j}$, and $A_{i+3,\ j}$. s_v is the space between the source pixel $A_{i+1,\ j}$ and the virtual interpolated pixel P_j. Likewise, all the horizontal weighting coefficients can be established by

$$h_j(1+s_h) = \begin{cases} -0.375s_{h'} & 0 \le s_h < \frac{1}{3} \\ -0.1875 + 0.1875s_{h'} & \frac{1}{3} \le s_h < 1 \end{cases}$$

$$h_{j+1}(s_h) = \begin{cases} 1 - 0.625s_{h'} & 0 \le s_h < \frac{1}{3} \\ 1.1875 - 1.1875s_{h'} & \frac{1}{3} \le s_h < 1 \end{cases}$$

$$h_{j+2}(1-s_h) = \begin{cases} 1.1875s_{h'} & 0 \le s_h < \frac{2}{3} \\ 0.375 + 0.625s_{h'} & \frac{2}{3} \le s_h < 1 \end{cases}$$

$$h_{j+3}(2-s_h) = \begin{cases} -0.1875s_{h'} & 0 \le s_h < \frac{2}{3} \\ -0.375 + 0.375s_{h'} & \frac{2}{3} \le s_h < 1 \end{cases} \quad (4)$$

The existing hardware structural design for weighting coefficient generation of FFOPCI is shown in figure 2 [10].

The existing WCG has 12 arithmetic elements. In the WCG circuit, two separate 1-D interpolation, the measurements of vertical weighting coefficients and horizontal weighting coefficient are not determined at the same time. All the coefficients are measured according to equations (3) and (4). This existing WCG includes that about four adders, four subtractors and four scalers [10].

As shown in figure 2, the circuit is designed for determining vertical and horizontal weighting coefficients. By using s_v, the first stage identifies the values of 0.375 s_v and 0.625 s_v. The final output values are $-v_i$, v_{i+1}, v_{i+2} and $-v_{i+3}$ if the select signal (vertical/horizontal) is vertical. If the select signal is horizontal then the outputs are $-h_j$, h_{j+1}, h_{j+2} and $-h_{j+3}$.

4 Proposed Architecture of Weighting Coefficient Generator

To reduce the area of the VLSI architecture of FFOPCI [10], an optimized design of WCG is proposed. The most computational effort in the proposed WCG is the reduction of number of arithmetic elements. The simplified WCG includes only two adders, three dividers and four subtractors as shown in figure 3.

Based on equation (3) and (4), the first scaling factor for generating weighting coefficient is 0.375. In the existing WCG circuit [10], the scaling factor 0.375 is determined by using three dividers/scalers and one adder. But, in the proposed WCG circuit, this scaling factor 0.375 is

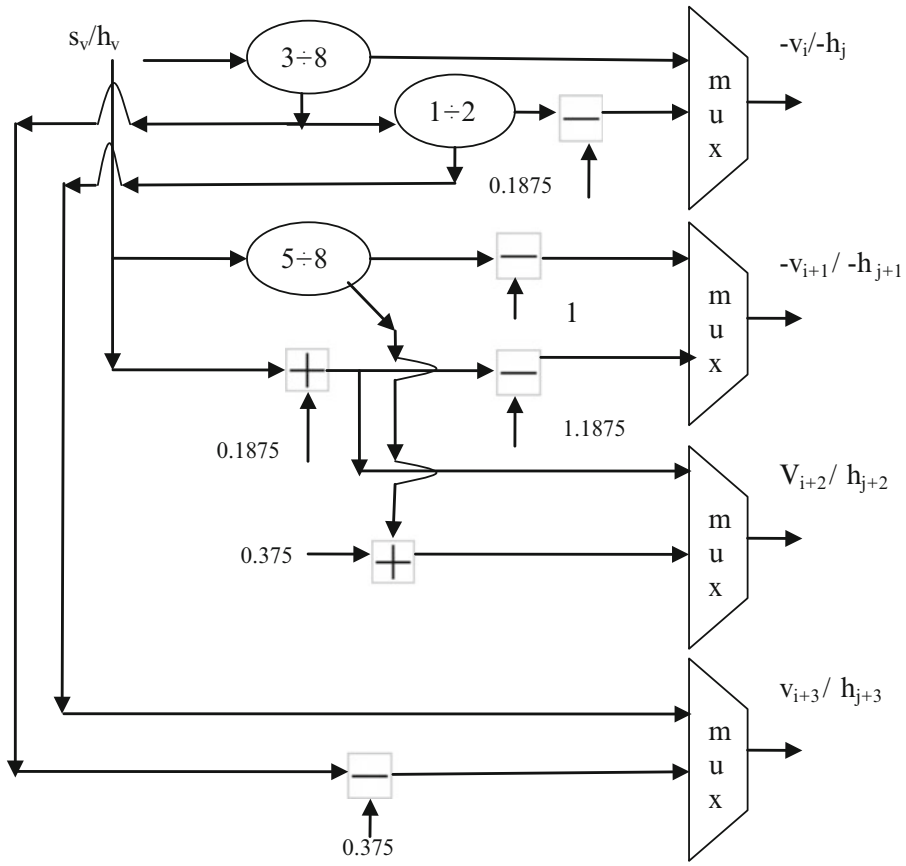


Fig. 3 Proposed weighting coefficient generation circuit of FFOPC

determined by using only one divider/scaler. Similarly, the next scaling factor 0.625 can be derived by using a single divider/scaler. Thus, the proposed WCG circuit reduces number of arithmetic operations. Totally, for generating all vertical and horizontal coefficients, the proposed WCG requires only two adders, three dividers/scalers and four subtractors as shown in table 1.

Based on table 1, the proposed VLSI architecture of WCG consists of only nine arithmetic elements, which is less than that for the WCG designed by Lin et al. [10].

5 Experimental Results of VLSI Architecture

To evaluate the performance of the VLSI architecture of WCG, the existing [10] and the proposed WCG circuits are simulated by using MATLAB Simulink and synthesized by using Xilinx ISE 12.3. The FPGA used for this evaluation is Virtex-6 xc65vs315t-3ff1156. The synthesis reports indicate that the proposed WCG architecture utilizes a less number of LUTs (Look up tables) and slices on Virtex-6 FPGA. Table 2 describes that the area dependent parameters of the VLSI architecture of both WCG of FFOPCI [10] and the proposed architecture of WCG. As shown in table 2, the proposed WCG used only 113 LUTs, which is much less than for the WCG suggested by Lin et al. [10].

Table 1 Number of arithmetic elements used in weighting coefficient generation circuits

Arithmetic Elements	WCG of FFOPCI [10]	Proposed WCG
Adders	4	2
Dividers	4	3
Subtractors	4	4
Total arithmetic elements	12	9

Table 2 Comparison of area dependent parameters for weighting coefficient generation

Parameters	WCG of FFOPCI [10]	Proposed WCG
Number of slice LUTs	132	113
Number of occupied slices	36	30
Memory usage (Kilo bytes)	164620	162800

6 Conclusion

In this paper, an area efficient VLSI architecture of weighting coefficient generation is proposed for fast first order polynomial convolution interpolation. This architecture has the advantage of low operation complexity for generating weighting coefficient. The computation burden is less, when using this architecture for generating weighting coefficient. It utilizes only nine arithmetic elements which is much less than the other methods. Therefore, the proposed architecture can simplify the circuit and reduce the chip area. Furthermore, the synthesis results demonstrated that proposed low complexity architecture provides a simple hardware than other weighting coefficient generation circuit. This architecture utilizes only 113 LUTs and 30 slices on the FPGA used for the evaluation. So the proposed weighting coefficient generator can be used to implement an area efficient fast first order polynomial convolution interpolation.

References

1. Juelin Leng, Guoliang Xu and Yongjie Zhang, "Medical Image Interpolation based on Multi-resolution Registration", *Computers and Mathematics with Applications 66, ELSEVIER,* 2013, pp. 1-18.
2. C. John Moses, Dr. D. Selvathi, J. Perpet Beena, and S. Sajitha Rani, "FPGA Accelerated Partial Volume Interpolation", *Proceedings of ICETECT, IEEE 2011*, pp. 816-819.
3. Francisco Cardells-Tormo and Jordi Arnabat-Benedicto, "Flexible Hardware_Friendly Digital Architecture for 2-D Separable Convolution_Based Scaling", *IEEE Transactions on Circuits and Systems*-II, vol. 53, no. 7,July 2006, pp 522-526.
4. Angelos Amanatiadis, Ioannis Andreadis and konstantinos Konstantinidis, "Design and Implementation of a Fuzzy Area-Based Image-Scaling Technique", *IEEE Transaction on Instrumentation and Measurements*, vol.57, no.8, August 2008, pp. 1504-1513
5. Christine M. Zwartakes, "Segment Adaptive Gradient Angle Interpolation" *IEEE Transaction on Image Processing*, vol. 22, no.8, August 2013, pp. 2960-2969.
6. Ramtin Madani, Ali Ayremlou, Arash Amini and Farrokh Marvasti, "Optimized Compact Support Interpolation Kernels," *IEEE transactions on signal processing*, Vol.60, no.2, February 2012, pp. 626-633,
7. Jianping Xiao, Xuecheng Zou, Zhenglin Liu, and Xu Guo, "Adaptive Interpolation Algorithm for Real-time Image Resizing," *Proceedings of the First International Conference on Innovative Computing, Information and Control*, 2006, pp. 221 – 224.
8. Nira Shezaf, Hagit Abramov, Iilan Suskover and Ran Bar Sella " Adaptive Low Complexity Algo rithm for Image Zooming at Fractional Scaling Ratio, *21st IEEE Convention of the Electrical and Electronic Engineers in Israel, proceedings*, 2000, pp. 253-256.
9. Chia-Sheng Tsai, Hsu-Huan Liu and Ming-Chieh Tsai, "Design of a Scan Converter Using The Cubic Convolution Interpolation with Canny Edge Detection," *International Conference on Electric Information and Control Engineering*, 2011, pp. 5813 - 5816.
10. Chung-chi Lin, Ming-hwa Sheu, Chishyan Liaw, and Huann-keng Chiang, " Fast First - Order Polynomial Convolution Interpolation for Real - Time Digital Image Reconstruction," *IEEE transactions on circuits and sys- tems for video technology*, vol. 20, no.9, September 2010, pp. 1260-1264.
11. Ran Feng, LIU Jiang and XU Meihua, "Interpolating Algorithm Optimization and FPGA Implemetation in Image Scaling Engine", *International Conference on Electronic Packaging Technology and High Desnsity Packaging*", 2007, pp. 1-4.
12. Marco Aurelio Nuno – Magand and Miguel O. Arias-Estrada, "Real -Time FPGA - Based Architecture for Bi-cubic Interpolation: An Application for Digital Image Scaling," *Proceedings of the 2005 International Conference on Reconfigurable Computing and FPGAs, IEEE*, 2005, page(s);8, pp. -1.
13. Chung-chi Lin, Ming-hwa Sheu, Huann-keng Chiang, Chishyan Liaw and Zeng-chuan Wu, "The Efficient VLSI Design of BI-CUBIC Convolution Interpolation For Digital Image Processing," *in Proc. IEEE Int. Symp. Circuits Syst.,* May 2008, pp. 480–483.
14. Erik H. W. Meijering, Karel J. Zuiderveld, and Max A. Viergever, "Image Reconstruction by Convolution with Symmetrical Piecewise n^{th}-Order Polynomial Kernels," *IEEE transactions on image processing*, vol. 8, no. 2, February 1999, pp. 192-201.
15. Chung-chi lin, Ming-hwa sheu, Huann-keng chiang, Chishyan liaw, Zeng-chuan wu and Wen-kai tsai, "An Efficient Architecture of Extended Linear Interpolation for Image Processing", *Journal of information science and engineering 26*, 2010, pp. 631-648.

Counting of water-in-oil droplets for targeted drug delivery systems using capacitive sensing technique

Cátia Barbosa and Tao Dong

1 Introduction

Two-phase microfluidic flows occur when two immiscible or in part immiscible fluids are in contact within microfluidic devices [1]. Gas-liquid [2] and liquid-liquid [3] flows are the most common and important. Several applications for two-phase flows have been documented; biomedical applications are one of the most important, including synthesis of biomolecules, drug delivery and medical diagnostic [4]. Among several flow patterns, droplet flow pattern is very popular because control and manipulation of droplets is easy and involved mechanisms in droplet formation can be directly observed [1].

Due to confinement of the reagents volume to one single droplet, droplet-based microfluidics offer several advantages. Reduction of reagents volume allows reducing costs and reaction times [5]. Besides, the use of droplets as containers for reactions avoids contact of reagents with solid walls, which is essential to prevent adsorption of reagents to channel walls [5]. For the stated reasons, several investigations have been performed in the last years to develop new systems to control, sort, mix and functionalize droplets [4].

The broad applicability of droplets makes droplet-based microfluidics especially interesting for biomedicine. Tang et al. [6] investigated the use of water droplets in targeted drug delivery by inhalation. Microemulsions of water-in-oil are also used for preparation of nanosized particles, as magnetic particles, used for medical diagnosis [7]. The water droplets are stabilized in the microemulsions with the use of surfactant agents [8]. For water-in-oil emulsions, hydrophilic drugs can be solubilized and slowly released; however, diffusion of hydrophobic drugs is less controlled and they will be quickly released [9].

C. Barbosa • T. Dong (✉)
Department of Micro and Nano Systems Technology, Høgskolen i Buskerud og Vestfold, P.O. Box 235, 3603 Kongsberg, Norway
e-mail: Tao.Dong@hbv.no

Droplets monitoring and detection usually comprises optical methods [10]–[13] however these systems provide big quantities of information, making analysis of collected data more difficult. Other systems for droplet and two-phase flows monitoring have been proposed, and make use of capacitive sensing technique [14]–[20]. Detection using capacitive sensing technique provides real time and accurate detection of droplets in microfluidic devices, providing only the needed and essential information. Besides, important information as droplets size and speed is provided by the acquired signals [16].

This article studies the use of a capacitive sensor using interdigital electrodes (IDEs) for water-in-oil droplets counting. Further sections provide an overview on sensor structure and working principle. Besides a system to read the signal from the sensor is proposed and experimental results are presented and discussed.

2 Sensor structure and working principle

2.1 Sensor structure

In this section we provide a description on the sensor structure and the main physical principles associated to its operation.

The sensor includes a three basic layer structure and patterned gold electrodes. A 200 µm thickness glass layer forms the substrate, on top of which the gold electrodes are patterned. A SU-8 passivation layer with 2 µm thickness covers the electrodes and is bonded to the glass substrate, avoiding cross-contamination of fluids. Besides, a 400 µm PDMS cover bonded to the structure, forms the 400 µm width x 150 µm high microchannel. The sensing area has 40 electrodes pairs, with 1 µm high, 5 µm width and 850 µm length; the distance between each finger is 5 µm. Fig. 1 shows the structure of the sensor, including the flow focusing junction, responsible for allowing formation of water droplets in the microchannel. The sensor has three inlets and one outlet. The continuous phase (oil) is injected

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_73,

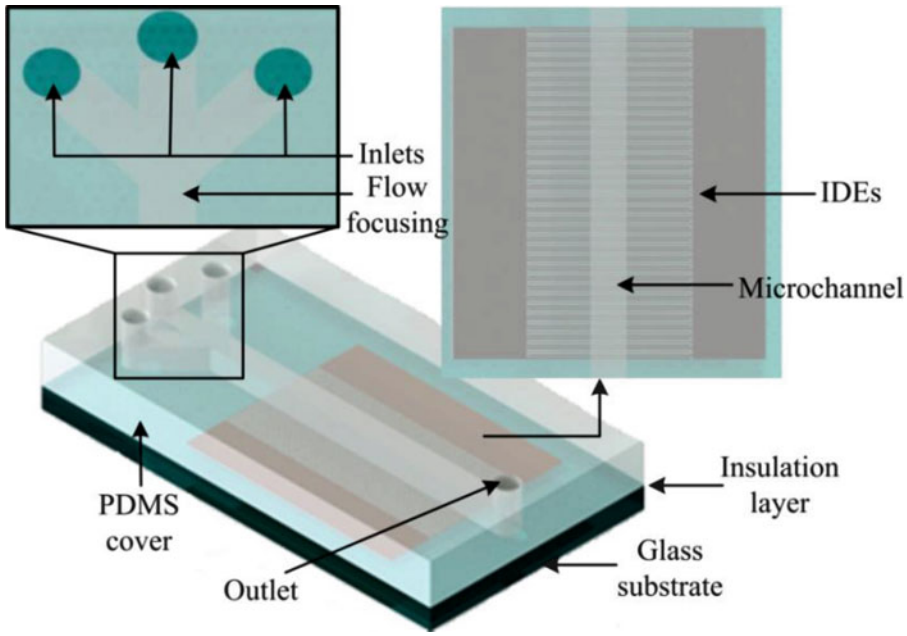


Fig. 1 Sensor structure with the glass, SU-8 and PDMS layers; besides IDEs geometry, inlets and flow focusing junction are enlarged.

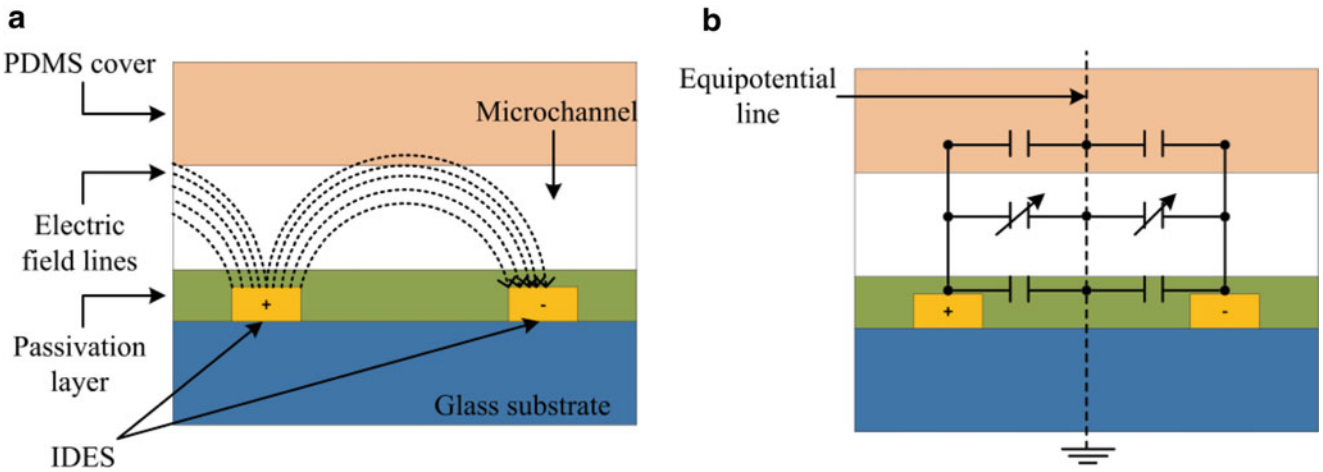


Fig. 2 Schematics of the cross-section view of the capacitive sensor. (a) Between electrodes with distinct polarities an electrical field is formed; inside the microchannel, where two-phase flows occurs, distinct materials will cross the electric field, providing different measurements in capacitance. (b) Between electrodes with different polarities there is an equipotential line; besides, all capacitances remain unchanged in both passivation and cover layers, however the capacitance in the microchannel changes due to changes in material.

through the edge inlets and the dispersed phase (water) is injected via middle inlet.

Main fabrication steps of the sensor followed the techniques proposed by Yang et al. [18]. In addition, the geometrical dimensions of the interdigital electrodes were studied in a previous work [21].

2.2 Working principle

Detection of droplets depends on the relative permittivity of continuous and dispersed phases. To make the droplets detectable, the relative permittivity of phases must be very different. This is verified for water and oil, being the values of relative permittivity (ε_r) 80 and 2,4, respectively.

Considering the configuration from Fig. 2(a), an electrical field is established between electrodes with different polarities. The electrical field lines cross the layers of the sensor, including the microchannel. Relative permittivity of layers remains approximately constant, except for the microchannel where two-phase flows occur (Fig. 2(b)) [16]. According to the material crossing the microchannel, capacitance will change, depending on the material properties. If the material has a low relative permittivity value, the sensor will

show a high impedance response; however, if a material with higher relative permittivity is present, impedance decreases and electric current will preferentially pass through this phase [22]. Based on these facts, detection of droplets is possible using capacitive sensing technique.

Besides the influence of the material in total measured capacitance, the geometrical properties of the IDEs will also influence capacitance, as Chen et al. demonstrated [14]. Working principle of IDEs is similar to parallel plate capacitors, changing the position of electrodes towards each other. The same parameters will have influence in measurements, namely the electrodes material and area, gap between electrodes and the material crossing the sensing region [23], as referred above.

3 Experimental system apparatus and procedure

Tests were performed using 0,001 ml/min water and oil flow rates. Due to very low flow rates, we waited more than one hour to obtain a stable droplet flow pattern. Because of the surface properties of PDMS, oil wets the walls of the microchannel and the observed flow pattern was very stable. To acquire the signal from the sensor we used a Wien bridge configuration, with one variable capacitor to balance the bridge. Afterwards, an instrumentation amplifier with high gain was used to calculate the difference between inputs coming from the bridge branches, as Fig. 3 shows. Fluids were driven to the droplet counter using two syringe pumps holding 10 ml volume syringes. Confirmation of results was performed using an electronic microscope (Fig. 3), by recording 20seconds duration films, at the same time as electrical data was acquired.

Final step in signal acquisition includes using a 16-bit DAQ from National Instruments, NI USB-6211 to read the signals from the electronic circuit to the computer. A LABVIEW interface is used to perform basic signal processing operations, as low pass filtration with existing virtual instruments and finally to display the signal in a graphic. Besides, the LABVIEW interface was also used to set a manual threshold for counting of water droplets. If the acquired signal is above a value determined by the threshold and at least 100 samples are above that value, the program identifies presence of one droplet and the counter will be increased in one unit.

4 Experimental Results and Discussion

4.1 Water droplets identification

Counting of water-in-oil droplets is achieved by using the integrated interdigital electrodes structure, as in Fig. 1. The observed water droplets were approximately 900 μm.

Due to higher relative permittivity of water when compared to oil, a capacitance increase is expected when a water droplet crosses the sensing area of the sensor. Since the voltage is inversely proportional to capacitance, an increase in capacitance will lead to a decrease in voltage drop across the sensor. For that, the value of *In1* (Fig. 3) will increase and the output voltage will increase. This is because *In2* (Fig. 3) remains constant and the instrumentation amplifier calculates the difference between *In1* and *In2*.

Fig. 4 shows the voltage variation verified for a droplet flow pattern. In 20 seconds time acquisition, three water

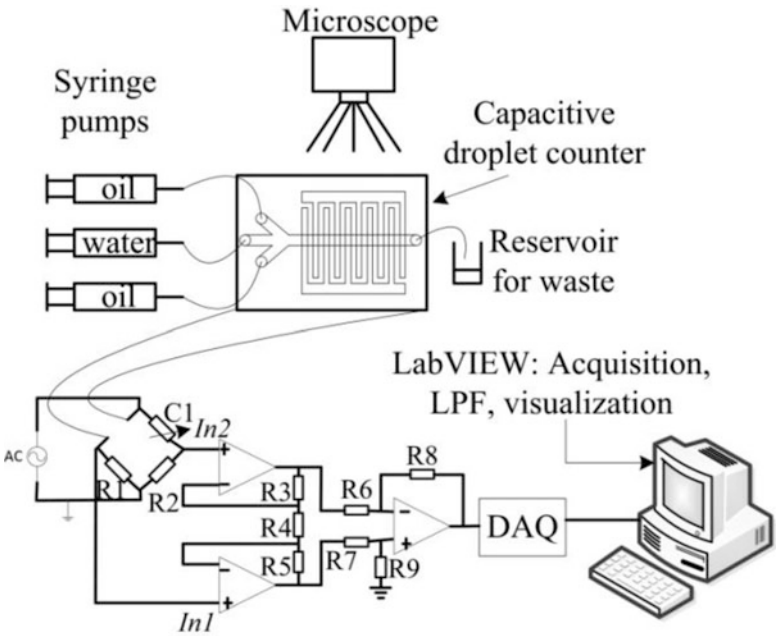


Fig. 3 Schematics on the acquisition and visualization apparatus used in the experiments.

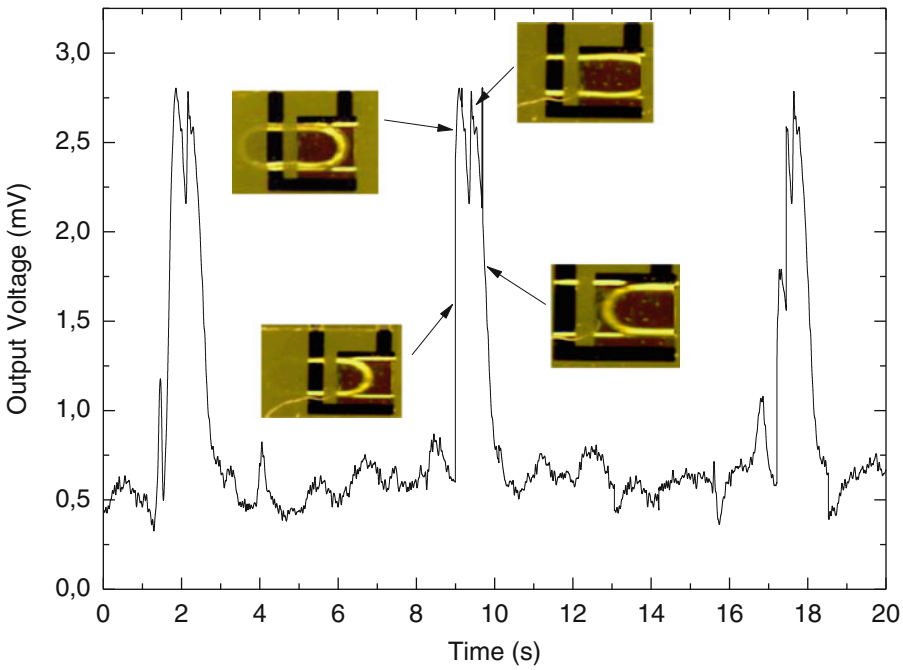


Fig. 4 Voltage variation associated to water-in-oil droplets crossing the sensing area of the sensor.

droplets cross the sensing area of the sensor. The verified voltage variation is approximately 2,2 mV, and the droplet remains in the sensing area 1,5seconds. The time interval between two followed droplets is approximately six seconds.

When the curved front of the droplet enters the sensing area, there is an increase in voltage, followed by signal saturation. Finally, when the curved back of the droplet leaves the sensing area, the voltage suddenly drops. After this stage, only oil occupies the sensing area. When the droplet is on top of the sensing area, the voltage signal is saturated (Fig. 4). This is due to the fact that the droplet length (approximately 900 µm) is higher than the length of the sensing area of the sensor (400 µm). When the droplet completely occupies the sensing area, there is no space for more water on top of the sensing area and, for that; the signal does not show variation. However, depending on the relative position of the droplet towards the electrodes, a wavy pattern appears in the saturated signal (Fig. 4). The insets in Fig. 4 show distinct positions occupied by the droplet inside the microchannel.

Moreover, some noise is present on the signal, which can be caused by bad shielding of the cables connecting the sensor to the electronic circuit. Improvements in the acquisition circuit must include a better isolation of the acquisition circuit.

4.2 Water droplets counting – interface

In this section we show the interface used for counting of the water droplets. LABVIEW was used to acquire the signals from the electronic circuit and plot the results in a user interface. Fig. 5 shows the interface used for signal acquisition, and includes four main components. The trigger to start acquisition allows the user to choose when data should start to be acquired; considering the acquired signal, the user must manually select a threshold to adjust for which voltage levels a droplet is considered; the number of droplets is shown in a text box. To be considered a droplet, at least 100 points of acquired signal have to be above the threshold value set by the user.

The previous interface is useful to monitor biomedical processes, including monitoring of water droplets used for target drug delivery. Due to the operational principle of the sensor, other materials can be detected in the microchannel, as long as the continuous and dispersed phases have very distinct values of relative permittivity.

5 Conclusions and Future Work

In this work we present a capacitive sensor structure capable of identifying water-in-oil droplets, used for targeted drug delivery. The sensor is versatile and, thus it can be used to identify other materials in the microchannel, as long as the continuous and dispersed phases have very different values of relative permittivity. The sensor is a low-cost option for monitoring of two-phase flows since the involved fabrication steps include very common processes. Besides, the sensor can be integrated in existing lab-on-chip devices.

In the future, we intend to better isolate the acquisition circuit, so that less noise affects the acquired signal. In

Fig. 5 Overview on the LABVIEW interface used for signal visualization.

addition we intend to improve the LABVIEW system, automatizing the calculation of the threshold, i.e., we intend to create an algorithm to automatically decide if a droplet is crossing the sensing area, instead of the threshold value being introduced by the user. The system will also be tested using other fluids.

Acknowledgements This article is supported by Norsk regional kvalifiseringsstøtte fra Oslofjorfondet (Høy-gjennomstrømnings mikrofluidisk plattform integrert med lavkostnads fotodetektorer for sensitiv kvantifisering av vannbårne patogener, Prosjekt nr: 229857). The Research Council of Norway is acknowledged for the support through the Norwegian Micro- and Nano-Fabrication Facility, NorFab (197411/V30). NCE Micro- and Nanotechnology and Innovation Norway are acknowledged for the support of this work.

The authors would like to acknowledge PhD Zhaochu Yang for the helpful suggestions and support throughout this work.

References

1. C.-X. Zhao and A. P. J. Middelberg, "Two-phase microfluidic flows," *Chem. Eng. Sci.*, vol. 66, pp. 1394–1411, 2011.
2. D. Huh, C.-H. Kuo, J. B. Grotberg, and S. Takayama, "Gas-liquid two-phase flow patterns in rectangular polymeric microchannels: effect of surface wetting properties," *New J. Phys.*, vol. 11, p. 75034, Jan. 2009.
3. A. Salim, M. Fourar, J. Pironon, and J. Sausse, "Oil-water two-phase flow in microchannels: Flow patterns and pressure drop measurements," *Can. J. Chem. Eng.*, vol. 86, pp. 978–988, 2008.
4. S.-Y. Teh, R. Lin, L.-H. Hung, and A. P. Lee, "Droplet microfluidics," *Lab Chip*, vol. 8, pp. 198–220, 2008.
5. H. Gu, M. H. G. Duits, and F. Mugele, "Droplets Formation and Merging in Two-Phase Flow Microfluidics," *Int. J. Mol. Sci.*, vol. 12, pp. 2572–2597, 2011.
6. K. Tang and A. Gomez, "Generation by electrospray of monodisperse water droplets for targeted drug delivery by inhalation," *J. Aerosol Sci.*, vol. 25, no. 6, pp. 1237–1249, 1994.
7. P. Tartaj, P. Morales, S. Veintemillas-verdaguer, and T. Gonz, "The preparation of magnetic nanoparticles for applications in biomedicine," *J. Phys. D. Appl. Phys.*, vol. 36, pp. 182–197, 2003.
8. S. a Agnihotri, N. N. Mallikarjuna, and T. M. Aminabhavi, "Recent advances on chitosan-based micro- and nanoparticles in drug delivery.," *J. Control. Release*, vol. 100, no. 1, pp. 5–28, Nov. 2004.
9. O. M. Koo, I. Rubinstein, and H. Onyuksel, "Role of nanotechnology in targeted drug delivery and imaging: a concise review.," *Nanomedicine*, vol. 1, no. 3, pp. 193–212, Sep. 2005.
10. X. Zhao, T. Dong, and Z. Yang, "Compatible immuno-NASBA LOC device for quantitative detection of waterborne pathogens: design and validation," *Lab Chip*, vol. 12, pp. 602–612, 2012.
11. B. Kuswandi, Nuriman, J. Huskens, and W. Verboom, "Optical sensing systems for microfluidic devices: A review," *Anal. Chim.*, vol. 601, pp. 141–155, 2007.
12. N.-T. Nguyen, S. Lassemono, and F. A. Chollet, "Optical detection for droplet size control in microfluidic droplet-based analysis systems," *Sensors Actuators B*, vol. 117, pp. 431–436, 2006.
13. X. Zhao and T. Dong, "A Microfluidic Device for Continuous Sensing of Systemic Acute Toxicants in Drinking Water," *Int. J. Environ. Res. Public Health*, vol. 10, 2013.
14. J. Z. Chen, A. A. Darhuber, S. M. Troian, and S. Wagner, "Capacitive sensing of droplets for microfluidic devices based on thermocapillary actuation," *Lab Chip*, vol. 4, pp. 473–480, 2004.
15. M. Demori, V. Ferrari, D. Strazza, and P. Poesio, "A capacitive sensor system for the analysis of two-phase flows of oil and conductive water," *Sensors Actuators A Phys.*, vol. 163, pp. 172–179, 2010.
16. C. Elbuken, T. Glawdel, D. Chan, and C. L. Ren, "Detection of microdroplet size and speed using capacitive sensors," *Sensors Actuators A Phys.*, vol. 171, pp. 55–62, 2011.

17. X. Niu, M. Zhang, S. Peng, W. Wen, and P. Sheng, "Real-time detection, control, and sorting of microfluidic droplets," *Biomicrofluidics*, vol. 1, no. 4, p. 44101, Jan. 2007.
18. Z. Yang, T. Dong, and E. Halvorsen, "Identification of microfluidic two-phase flow patterns in lab-on-chip devices," *Biomed. Mater. Eng.*, vol. 23, pp. 77–83, 2013.
19. Z. Yang, E. Halvorsen, and T. Dong, "Electrostatic energy harvester employing conductive droplet and thin-film electret," *J. Microelectromechanical Syst.*, vol. 23, no. 2, pp. 315–323, 2014.
20. Z. Yang, E. Halvorsen, and T. Dong, "Power generation from conductive droplet sliding on electret film," *Appl. Phys. Lett.*, vol. 100, p. 213905, 2012.
21. C. Barbosa, C. Silva, and T. Dong, "Integratable capacitive bubble counter for lab-on-chip devices," in *International Symposium on Medical, Measurements and Applications, June 11–12*, 2014, p. 4.
22. J. Janouš, J. Čech, P. Beránek, M. Přibyl, and D. Šnita, "AC electric sensing of slug-flow properties with exposed gold microelectrodes," *J. Micromechanics Microengineering*, vol. 24, no. 1, p. 015002, Jan. 2014.
23. A. V. Mamishev, K. Sundara-Rajan, F. Yang, Y. Du, and M. Zahn, "Interdigital Sensors and Transducers," *Proc. IEEE*, vol. 92, no. 5, pp. 808–845, 2004.

Privacy Preserving Biometric Voice Authentication System – SIPPA-based Approach

Bon K. Sy

1 Introduction

This research paper presentsa privacy preserving biometric voice authentication system based on SIPPA. A traditional biometric authentication system stores the biometric enrollment of a user as a credential, and uses it to compare against a biometric sample provided by a user during the authentication process. Although biometric authentication provides convenience through its natural mechanism of using *what one is* as a security token, it also raises concern on privacy leak if the enrolled biometric data is stolen. SIPPA offers a solution to address this concern.

SIPPA [1,2] is a two-party secure computation method for comparing the private data of two parties without each party disclosing their private data to each other. If their data are "sufficiently" similar, one party can reconstruct the private data of the other party. In the reconstruction process, the server party provides helper data for the client party to reconstruct server data that preserve perfect accuracy, or an accuracy proportional to the similarity of the private data of both parties. These unique characteristics of SIPPA allow us to realize a novel authentication mechanism, which can be described as below:

- Sensitive credential information for authentication is encoded by the personal private information of an individual.
- Only the encoded information is stored. Sensitive credential information and personal private information are never stored. But the credential information can be reconstructed during the execution of SIPPA protocol — when the personal private information presented by an individual is sufficiently similar to that used for encoding the sensitive information.

B.K. Sy (✉)
Computer Science Department, Queens College and University Graduate Center/CUNY, Flushing, NY 11367, USA
e-mail: bon@cs.qc.cuny.edu

In our approach, private sensitive information will be derived on demand. This eliminates the risk on information leak since no private sensitive information is stored in the first place. Therefore, information privacy is assured. Furthermore, the SIPPA protocol has been analyzed under different security models and situations [1,2], the behavior and the security of the authentication protocol can be derived from that of the SIPPA protocol, and formally analyzed and assessed accordingly.In this research, a particular embodiment of the proposed approach utilizing biometric voice signature will be described — although the embodiment could be based on any modalities and devices.

2 Formulation and system architecture

In this research, the identity of an individual is characterized by three facets [3]: (i) *what one knows*, referred to as a UID (Universal ID) — a unique ID generated by the system based on personal information PID; e.g., phone number or birth date, (ii) *what one has*, referred to as a DID (Device ID) such as a personal mobile phone, and (iii) *what one is*, referred to as BID (biometric ID) such as the biometric voice, face or fingerprint. More specifically, the identity of an individual is a 3-tuple composed of DID, a biometrically encoded encryption key — BID + K, and the decryption of the encrypted hash on UID; where K is a secret key. Formally, an identity is then represented by a 3-tuple: <DID, BID + K, Dec(K, Enc(K, Hash(UID))> .

The architecture of our system consists of three components; namely, a voice gateway (VG), an Enrollment Module (EM) comprised of SIPPA server and a local database, and an Identity Storage and Verification Module (ISVM) comprised of SIPPA client and a centralized database. The delineation of the system architecture is shown in Figure 1.

In our design, the local database of the enrollment module will be the storage for the encryption/decryption secret K. The centralized database will be the storage for the identity

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_74,

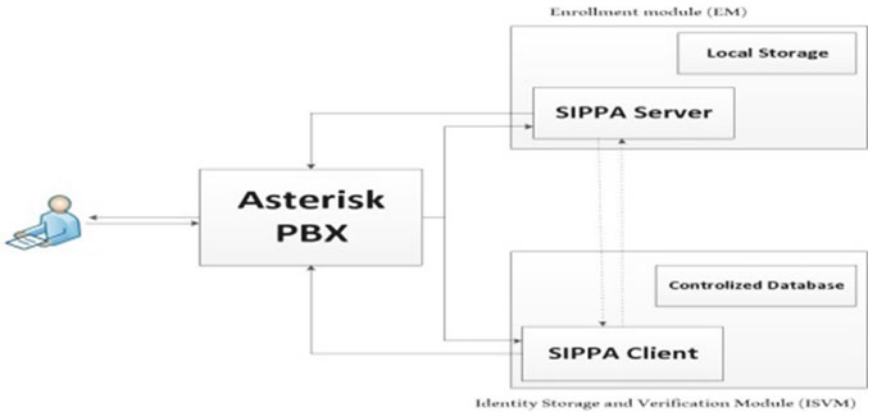


Fig. 1 Voice-based authentication system

information. For privacy assurance, the EM and the ISVM do not directly share with each other the information in their databases. Furthermore, the two modules do not even have to treat each other as a trust worthy party. This is different from the traditional systems where the trustworthiness [4] between the system components similar to that of EM and ISVM is assumed.

3 Literature review

Privacy preserving identity management is an active research topic in many different domains. In general, the goal is to minimize disclosure on the identity information of an individual, certain content information about an identity such as phone number or birth date, the linkability of the identity information and its usage, the issuer of the identity, and the data matching.

The research in this area can broadly be classified into cryptographic based and non-cryptographic based approach. In cryptographic based approach, Attribute Based Credential for Trust (ABC4Trust) [5] is an exemplary state-of-the-art that allows an individual to use not one public key, but possibly multiple public keys. In addition, certificate is based on the individual's secret key and attributes that may be hidden from the Certificate Authority. Furthermore, ABC4Trust also requires proof of knowledge of certificate about identical secret key used in different certificates. This is different from the conventional Public Key Infrastructure in that a certificate is based on an individual's public key, and the certificate (thus the information in the certificate) is revealed. Although ABC4Trust is an improvement over the traditional approach, the implicit deployment assumption of ABC4Trust is that a secure and trustworthy issuer (typically a Certificate Authority) exists and is always available.

An interesting aspect of non-cryptographic based approach is the idea of Privacy Preserving Data Matching (PPDM) as exemplified by Scannapieco et al [6]. The key idea behind PPDM is the use of an embedding space SparseMap [7] that preserves the similarity distance between two data objects in the metric space. The embedding space is constructed by using a subset of data objects serving as a reference set, and the distance between two data objects is mapped to two distance measures in the metric space; i.e., between a data object and the reference set, and the other data object and the reference set. Through triangular inequality, a lower bound distance measure between the two data objects can be obtained; thus realizing the privacy preserving approximate matching.

Our proposed SIPPA approach towards a privacy preserving voice-based authentication shares similar characteristics to the research just mentioned. Yet it distinguishes itself with characteristics that are unique and attractive for privacy preserving authentication. In both our approach and ABC4Trust, Public Key Infrastructure (PKI) is required. The main difference lies on the extent that the PKI is used. In ABC4Trust, a key characteristic is to issue every user multiple keys so that privacy protection can be achieved. In our proposed approach, Certificate Authority is only required for the key infrastructure; i.e., the Voice Gateway (VG), Enrollment Module (EM), and Identity Storage and Verification Module (ISVM). In our specific applications of SIPPA approach in a real world situation with a large population of unknown users, it will be extremely challenging — if practical at all — to assume the existence of a trustworthy environment for every user to securely receive the private and public keys needed as in ABC4Trust.

In reference to PPDM, our approach also tackles the problem of privacy preserving data comparison through an alternative metric space. However, our approach is completely different from that of PPDM. While PPDM relies on SparseMap for the construction of the embedding space, SIPPA maps the data objects to their Eigen space through the symmetric matrices derived from the data objects. More importantly, PPDM aims at privacy preserving approximate

matching. SIPPA, on the other hand, aims at utilizing the mathematical properties implicit in the Eigen space mapping that allows precise reconstruction of the private data based on sufficiently similar data objects.

4 Theory, System Realization, and Analysis

A significant contribution of this research is an authentication system that incorporates privacy assurance with the following properties:

- The identity of an individual is multi-facet and is based on *what one knows* (UID), *what one has* such as a mobile phone (DID), and *what one is* such as biometric voice signature (BID).
- A system that is fail-safe; i.e., it preserves the privacy of personal information — even if the system is compromised.

Our approach towards the development of a fail-safe system is to employ cryptographic key to protect the confidentiality of the UID/DID. The cryptographic key is generated, used and discarded. It is never stored. Only the biometrically encoded encryption key K + BID is stored; where BID is a biometric ID as discussed in section 2. The key is regenerated based on the biometrics of an individual whenever it is needed. Given a biometric sample S, the pre-processing step of the regeneration is a simple cancellation operation; i.e., (K + BID) – S.

By trivial inspection, the cryptographic key K can be perfectly regenerated in the pre-processing step if BID = S. However, personal biometrics can seldom be reproduced identically. Therefore, in general BID and S are different. When BID and S are from the same individual, the error incurred by BID-S is typically small. Otherwise BID-S is expected to be relatively large.

SIPPA Theory

SIPPA [1,2] is a 2-party secure computation protocol [8] where a client party can reconstruct source data of a server party under the following conditions:

1. The client party must possess some client data that is a "sufficiently good approximation" of the source data, in order to initiate the SIPPA process.
2. Rather than revealing the source data of the server party to the client party, only some helper data related to the Eigen components of the source data is provided (by the server party) to the client party for reconstructing the source data.

In our case, the SIPPA client retrieves K + BID from the centralized database, and performs the cancellation K + BID-S. K is stored in the local database of the SIPPA server. Through the execution of the SIPPA protocol, the SIPPA client will be able to reconstruct K if (K + BID-S) and K are sufficiently similar. The formulation, the key results of SIPPA summarized as two theorems, and the algorithmic steps are already reported elsewhere [1,2].

Realization of SIPPA Protocol for User Enrollment and Authentication

In our application of SIPPA, the server data **de** is a vector of 20x1 of real numbers in the range [0,1]. The secret K stored in the local database of SIPPA server is a 20x1 vector of normalized integer values that are a fixed point representation of the real numbers. During an encryption/decryption, an AES key is generated from the MD5 hash of K. The client data **dv** is also a vector of 20x1 of real numbers derived from (K + BID)-S; where BID and S each is a normalized 20x1 vector representing a biometric voice template of cepstrum coefficient [9] in the frequency range of 0-4KHZ based on Mel scale using triangular filters.

Protocol for identity enrollment:

1. An individual established connection through a secure authenticated channel [10] to download client-side software such as an applet capable of biometric voice signature extraction and cryptographic key generation.
2. The individual submits – through the downloaded client-side software – to the voice gateway his phone number that can be recognized as a caller ID for a call back.
3. If the caller ID is valid and unique, the voice gateway signs the caller ID and calls the individual back. It returns the signed version of the caller ID as the device ID – DID, as well as a token T (e.g. a random number or a timestamp). In addition, the voice gateway also sends T to ISVM.
4. The individual records his/her voice sample and uses the downloaded client-side software to extract the individual's voice signature as his biometric ID – BID. The client-side software also generates a cryptographic secret key K.
5. The cryptographic secret key K and DID — <DID, K> — are encrypted (using the public key of EM Enc (K^{+}_{EM}, <DID, K>)) and sent to the Enrollment Module (EM) through a secure authenticated channel; and decrypted by EM (i.e., Dec(K^{-}_{EM}, Enc(K^{+}_{EM}, <DID, K>))) and stored upon receiving.
6. Three pieces of information are derived by the individual using the client-side software: Generate a UID using some personally known information and the token T, and then hash UID — Hash(UID); Encrypts the hash using K—Enc(K, Hash(UID)); Computes K + BID + N where N is some noise generated by the individual.
7. 3-tuple <DID, K + BID + N, Hash(UID)> is encrypted and sent to the Identity Storage and Verification Module (ISVM) through a secure authenticated channel; and decrypted by ISVM upon receiving.

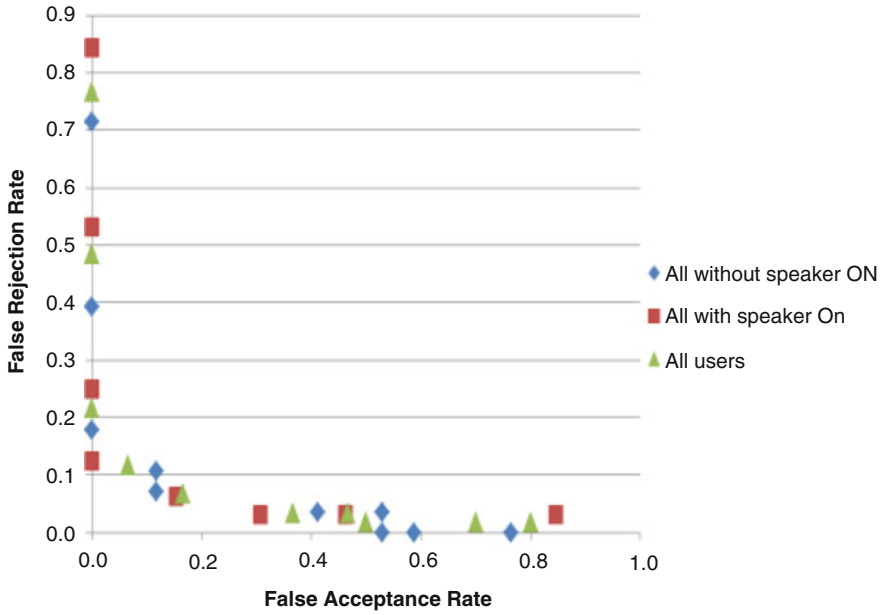


Fig. 2 ROC with(out) speaker ON

8. The downloaded client-side software is terminated and discarded. K, UID, BID, and hash(UID) are also discarded. The individual retains only DID, T, and Enc (K, Hash(UID)).

It is noteworthy that the enrollment process described above does not rely on Certificate Authority to verify the identity of an individual. Instead, the enrollment process above allows an individual to create and self-sign an identity, whereas the process to bind an individual to a unique identity is based on what an individual has (e.g., mobile phone). It does not care the individual information that an individual may specify, since it is not relevant to the identity verification process. As such, two individuals could have, for example, the same name but different DID and BID. They will be identified as two different entities as distinguished by different 3-tuples.

Protocol for identity verification:

1. An individual presents to voice gateway (VG) his DID and a noise-added biometric sample S + N.
2. Voice gateway relays DID to SIPPA server, and voice gateway relays S + N and DID to SIPPA client.
3. Based on DID, SIPPA client retrieves Hash(UID) from the centralized database. SIPPA client also retrieves (K + BID + N), and computes (K + BID + N)-(S + N).
4. Execute SIPPA protocol for the SIPPA client to construct a secret K'; i.e., SIPPA(client: (K + BID + N)-(S + N), server: k) -> (client-result: K', server-result: similarity between K and K + BID-S); where K' = K if (K + BID + N)-(S + N) is sufficiently similar to K.
5. SIPPA client returns K' through the voice gateway to the individual for the individual to derive Dec(K', Enc(K, Hash(UID))).
6. Compute Dec(K', Enc(K, Hash(UID))) (by individual/SIPPA client).
7. Present Dec(K', Enc(k, Hash(UID))) and the token T to ISVM for comparing against that stored in ISVM; ISVM accepts the claimed identity if Dec(K', Enc(K, Hash (UID))) is found identical to Hash(UID) of ISVM with a matching T during the authentication.

Security analysis:

Claim: As long as there is no collusion between an individual user/impostor and the identity management system, the privacy of the sensitive identity information UID and BID is guaranteed even if all the entities are compromised; i.e., the system components are compromised by some adversary independent of the impostors.We now provide a sketch for the security analysis:

@Claim: If all entities are compromised, EM and ISVM can collude to derive hash(UID); i.e., EM provides the secret key K to ISVM to decrypt Enc(K, hash(UID)). However, UID cannot be reverse engineered from hash(UID) because adversaries have only polynomial bounded computing power. In addition, since N is not known to any entity, BID cannot be uncovered from K + BID + N even K is made know through collusion. Therefore, UID and BID are protected even if all parties are compromised.

Furthermore, if an impersonator is able to steal an individual's mobile phone and derive the DID, and if the personally known information of the individual (e.g., age or work address zip code) for generating UID is also made available for the public by a third party source, the impersonator still would not be able to reproduce UID (thus hash (UID)) without the knowledge of the token T generated through the commitment scheme during (the step 3 of) the enrollment process. Therefore, the impersonator will not be able to synthesize hash(UID). In other words, an attempt to fool the system into false acceptance by bypassing steps 1 - 6

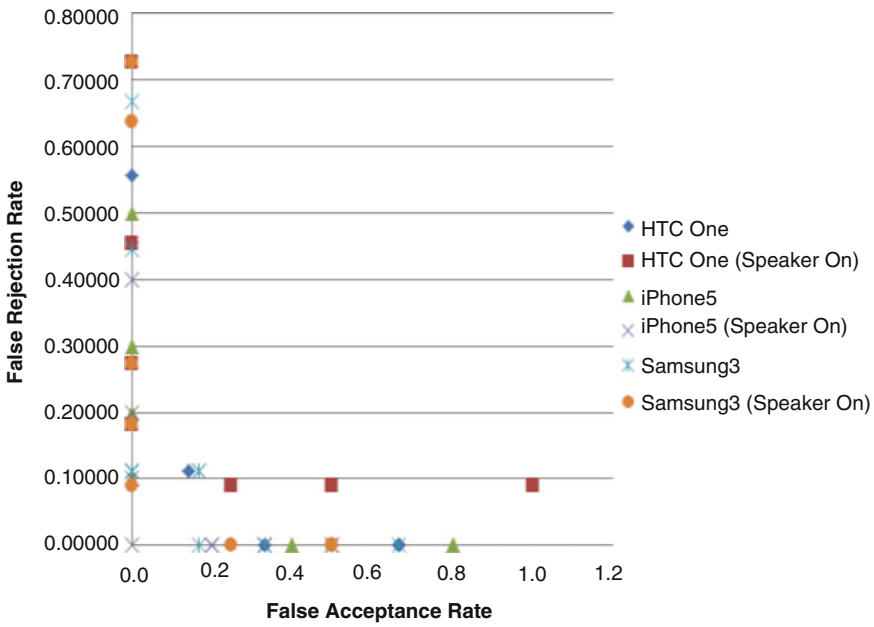


Fig. 3 ROC grouped by phone model

in the verification protocol through synthesizing hash(UID) will not succeed. The only one scenario stated as an exclusion of the claim that can leak UID is the disclosure of the token T by the compromised system to the impersonator (or vice versa) for the reconstruction of UID, thus hash(UID), using the personally known information found from a third-party source and the token T.

5 System implementation and experimentation

For proof-of-concept, we conduct an experimental study on a prototype of the proposed voice-based authentication system. In this experimental study, the test set consists of 90 calls from a pool of a dozen of individuals using three different phone models — with and without enabling the speaker phone mode.

The results of this study are shown in Figures 2 and 3 as ROC plots on false acceptance (FA) vs false rejection (FR). The plots in Fig. 1 are the changes in the ROC with the speaker phone mode enabled/disabled. The Equal Error Rate (EER) in all cases is about 0.1. The plots in Fig. 2 show different ROCsgrouped by phone models; whereas the EER ranges from 0 (iPhone 5) to about 0.14 (HTC One).

6 Conclusion

In this paper we present a privacy preserving biometric voice authentication system based on our previous work on SIPPA. Enrollment and verification protocol were analyzed under the security model with secure authenticated channel. The security analysis showed that the privacy of the sensitive information is preserved even if all the system components were compromised. For proof-of-concept, we conducted experimentations to investigate the effectiveness of the prototype system in regard to its performance summarized in the ROC. Our future work will investigate the effect of noise in the telephony channel on the performance and the extensibility of its applications.

Acknowledgement This work is supported in part from a grant by PSC CUNY Research Award.

Reference

1. Bon K. Sy and Arun P. Kumara Krishnan, “Generation of Cryptographic Keys from Personal Biometrics: An Illustration based on Fingerprints,” *New Trends and Developments in Biometrics*, ISBN 980-953-307-576-6, InTech, 2012.
2. Arun P. Kumara Krishnan and Bon K. Sy, “SIPPA-2.0 - Secure Information Processing with Privacy Assurance (version 2.0),” *Proc. of the 9th Annual Conference on Privacy, Security, and Trust*, Paris, France, July 2012.
3. William E. Burr, Donna F. Dodson, Elaine M. Newton, Ray A. Perlner, W. Timothy Polk, Sarbari Gupta, Emad A. Nabbus, *Electronic Authentication Guideline*; Special Publication 800-63-1; Dec 2011.
4. Peter G. Neumann, “System and Network Trustworthiness in Perspective,” *CCS 06*, October 30–November 3, 2006, Alexandria, Virginia, USA.
5. Jan Camenisch, Maria Dubovitskaya, Anja Lehmann, Gregory Neven, Christian Paquin, Franz-Stefan Preiss, "Concepts and Languages for Privacy-Preserving Attribute-Based Authentication," *Policies and Research in Identity Management: IFIP Advances in Information and Communication Technology*, Volume 396, 2013, pp 34-52.

6. Monica Scannapieco, Ilya Figotin, Elisa Bertino, Ahmed K. Elmagarmid, "Privacy Pre-serving Schema and Data Matching," Proceedings of the 2007 ACM SIGMOD international conference on Management of data, Jun 2007, Beijing China.
7. J. Bourgain, “On Lipschitz Embedding of Finite Metric Spaces in Hilbert Space,” Israel Journal of Mathematics, 52 (1985), no. 1-2, 46.
8. R. Cramer, I. Damgard, Jesper Buus Nielsen, Multiparty Computation: An Introduction; http://www.daimi.au.dk/~ivan/mpc.pdf
9. Thrasyvoulou T., Benton S.: Speech Parameterization Using the Mel Scale (Part II), (2003).
10. M. Fitzi, D. Gottesman, M. Hirt, T. Holenstein, A. Smith, “Detectable Byzantine Agreement Secure Against Faulty Majorities,” Proc. of the 21st ACM Symposium on Principles of Distibuted Computing (PODC), July 2002.

Monitoring Urban and Land Use Changes in Al-Kharj Saudi Arabia using Remote Sensing Techniques

Osama S. Algahtani, Ahmed S. Salama, Abdullah M. Iliyasu, Belal A. Selim, and K. Kheder

1 Introduction

Urbanisation is, among many factors, influenced by the economic and industrial development of a city factors that are known to accelerate the decline in agricultural (green) areas of the city. This green land cover, used mainly for agricultural and recreational purposes, vanishes leaving fewer land areas for the cultivation of crops, rearing of livestock, and other sporting and recreational activities. Such decline in these land use types is, however, known to have huge short and long term impacts on the health and well-being of the populace: food sourced from further distances, less exercise, etc.

Within the last two decades, Al-Kharj, a city located 70Km to the south of Riyadh, the capital city of KSA, has witnessed immense development and growth mainly as result of the Kingdom‘s efforts in developing cities surrounding the capital city. In tandem with this, capital projects worth Billions of Saudi riyals (a university, hospitals, industrial cities, etc.) have been cited in and around Al-Kharj. While these are desirable development strides, as mentioned a while ago, they also have dire consequences on the environmental well-being of the citizenry.

The main objective of this study is focused on assessing how much of Al-Kharj city‘s green areas (agricultural land cover) has depleted over the last decade-and-a-half, and in particular, the consequences of such depletion in the economic and environmental well-being of the city and its inhabitants.

Beyond Al-Kharj and the KSA, land use and land cover changes are major factors in global change because of its interactions with climate, ecosystem processes, biodiversity and human activates among many others [1–3].

In this study, we use the spatial analysis function of GIS to specify the urban expansion characteristics found in Al-Kharj City (Saudi Arabia) during the past 14 years, and then, using three scenes of Landsat Thematic Mapper (TM) images we detect and evaluate the land use and land cover changes that occurred in the city over last 14-year period covered by the study. Therefore, other objectives of the study include:

1. To explore and assess the temporal characteristics of urban expansion witnessed in the city over last decade-and-a-half;
2. To detect and evaluate the land use and land cover change due to urbanisation (between 2000 and 2013) and the output land cover maps; and, finally,
3. To analyse the depletion in green (agricultural) land cover of Al-Kharj city and assess its consequences on the economic, industrial, social and geo-political life of the citizenry.

2 Description of the study area

Al-Kharj, our study area, is located about 70Km south of Riyadh, the capital of the Kingdom Saudi Arabia on the geographical coordinates 23°59′N 47°09′E-24°22′N 47°06′′′E (Fig. 1 and Fig. 2).

Although always considered as an agrarian city, its status as one was further cemented by government’s determination, during the early 80’s, to boost agricultural production and domestic self-sufficiency in diary and livestock production. Based on that policy, farmers were offered many incentives including free land, interest-free loans for fertilizer, seeds and machinery in addition to guaranteed purchase

O.S. Algahtani • A.S. Salama • B.A. Selim • K. Kheder
College of Engineering, Salman Bin Abdulaziz University, P.O. Box 173, Al-Kharj 11942, Kingdom of Saudi Arabia

A.M. Iliyasu (✉)
College of Engineering, Salman Bin Abdulaziz University, P.O. Box 173, Al-Kharj 11942, Kingdom of Saudi Arabia

Department of Computational Intelligence & Systems Science, Tokyo Institute of Technology, Tokyo, Japan
e-mail: a.iliyasu@sau.edu.sa

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_75,

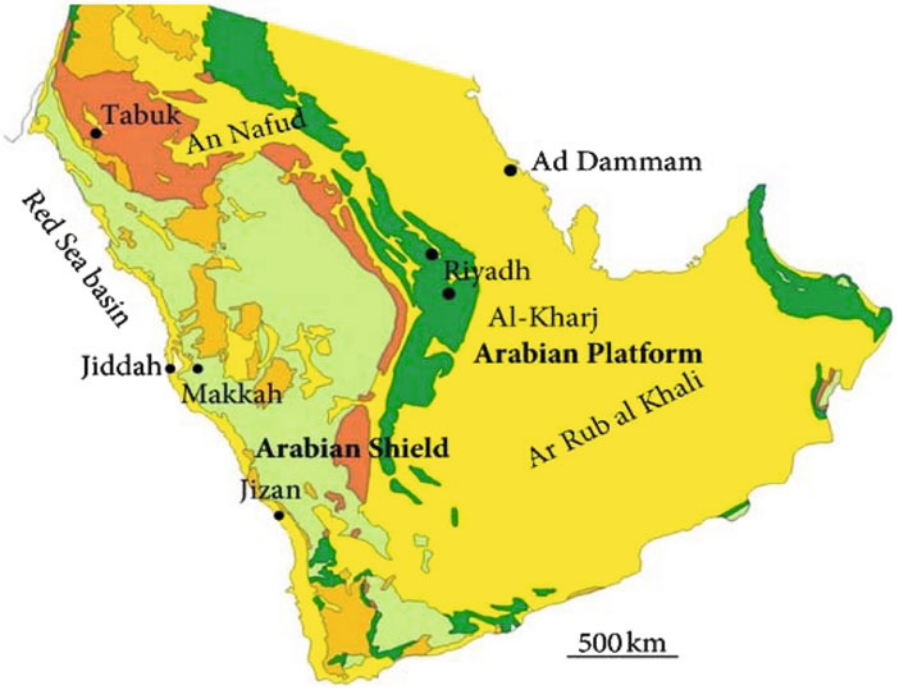


Fig. 1 Map showing geographical layout of sedimentary rock formation of Al-Kharj Governorate in Riyadh Province of KSA (Source: Saudi Geological Survey, 2008).

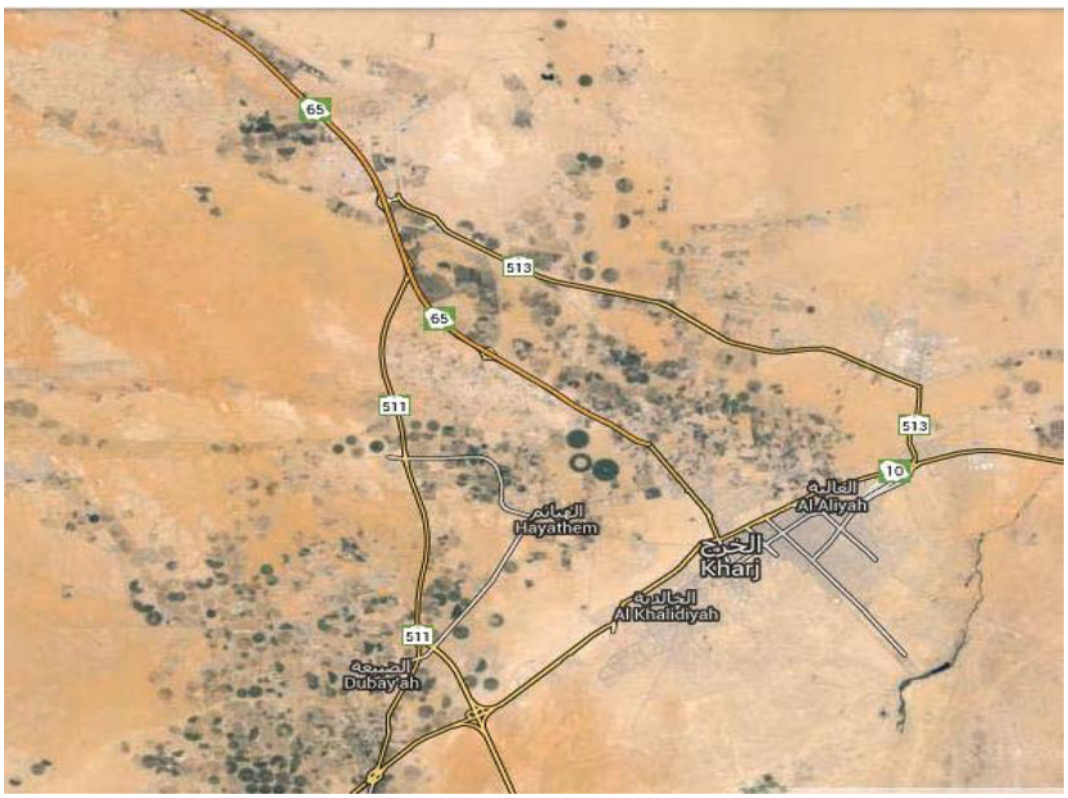


Fig. 2 GPS location indicating the sampling location in the Al-Kharj (in Riyadh province) in central Saudi Arabia (Source: Google earth).

of their produce. Although most of the farming was irrigation-driven, the farmers responded enthusiastically to the government's kind gestures leading to enormous increases in agricultural output and creation of jobs.

The advantages of government's policy to boost agricultural production are also manifest in the growth in the city's population. With a population of just around 180,000 inhabitants back in the year 2000, the population grew to near a quarter of a million by 2010 [4].

Economically, Al-Kharj, home to one of the world's largest dairy farms (owned by Almarai), produces 80 % of the Kingdom's diary needs. Because of the increase in population and growth of Riyadh, the capital city of the Kingdom, and the overstretching of some of its facilities, government has, of recent, initiated a deliberate policy to develop and expand a few of the surrounding cities.

These reasons, combined with Al-Kharj's proximity to the capital city have made it a logical choice among locations where most of these expansions are sited. Within the last decade projects worth billions of Saudi Riyals including a university, hospitals, many government offices, industrial and residential layouts have been initiated and executed, often from a scratch, in the city.

Together with the smaller cities surrounding it: Wadi Dawasir in the southwest; Hofuf in the northeast; Gassim in the northwest; and Taif in the west, Al-Kharj Governorate is considered one of the five major sources of agriculture crops in the Kingdom of Saudi Arabia [5].

The climate in this area is typically arid with an annual precipitation of 132 mm. Al-Kharj is considered to be a lowland with alluvium deposition. The soil type is entisol and aridisol with saline and calcareous profiles [6]. The aquifers in Al-Kharj are contained within the Eocene Dammam Formation and Miocene-Pleistocene sedimentary rocks present within the stable shelf tectonic unit between the Arabian shelf and western Precambrian shield [7, 8]. Major recharge of the aquifer systems in this region has been estimated to have occurred during pluvial periods, some 25,000 to 30,000 years ago, based on isotopic methods (stable carbon, oxygen, and hydrogen) [9]. The recharge of these aquifers has been estimated to occur at a rate of 15 % of the total annual rainfall that is 100 mm approximately.

3 Data and methods

3.1 Data

In this study, multi-temporal maps, and two scenes of satellite multi-spectral images of the study area (Al-Kharj) are used to evaluate the temporal and spatial characteristics of urban expansion from the years 2000, 2007 and 2013.

3.2 Maps processing

Three maps from different historical periods (within the 14 years covered by the study) were digitalised and processed using GIS (MapInfo5.0) and ERDAS software. The maps were geometrically inter-matched and converted to Universal Transverse Mercator map projection. The topographical map of the year 2000 was employed as the base map. The urban area borders in the different periods were determined for the purpose of calculating the extension rate. Use of historical maps to extract the boundaries of the city to investigate patterns of urban expansion is relatively simple. Even though it might have some potential errors in accuracy, because the maps have been drawn up with various degrees of accuracy, as argued in [10], the influence of this variability will be minimal due to the relatively coarse time scales used in the present study.

3.3 Annual urban growth rate

In order to evaluate the spatial distribution of urban expansion and its intensity, we adapted an indicator called annual urban growth rate (AGR) [10] for evaluating the 'urbanisation' speed of a unit area. AGR is defined as follows [1, 10]:

$$\mathrm{AGR} = \frac{\mathrm{UAn} + \mathrm{i} - \mathrm{UAi}}{\mathrm{nTAn} + \mathrm{i}} * 100\% \qquad (1)$$

Where $TAn + i$ is the total land area of the target unit to be calculated at the time point of $i + n$; $UAn + i$ and UAi the urban area or built-up area in the target unit at time $i + n$ and i, respectively, and n is the interval of the calculating period (in years). In this study, we opted for the geographical gridding unit. The annual urban growth rates of each unit were then calculated. Lastly, the grid-based annual urban growth rates were clustered and mapped to evaluate the spatial features of the 'expansion'.

3.4 Satellite image pre-processing

Cloud free images from three scenes were collected for the purpose of analysing land use and land cover changes between the three periods: 2000, 2007 and 2013. First, the image from the year 2000 (the base image) was geometrically corrected to Universal Transverse Mercator map projection system. Then the remaining images from 2007 and 2013 were also geo-encoded and matched to the TM image with the total root mean square (RMS) error of less than half-pixel.

3.5 Land use and land cover classification

Prior to land cover classification, a 9-class classification system was designed with consideration of the land use properties of the study area as urban/built-up, residential, crop field, vegetable field, forest/trees, orchard, grass, water body, and barren/sandy lands. The widely used supervised classification method, Maximum Likelihood [11], was employed to detect the land cover types.

4 Results

4.1 Landscape change in the last-decade-and-a-half (14 years)

In order to evaluate the landscape change due to high speed urbanisation in the recent 14 years, images from three scenes were used for land use/land cover classification as described earlier. From this, the land use and land cover change was detected and analysed. Figures 3 to 5 show the three land use/land cover maps prior to classification, while Figures 6 to 8 show the classified versions corresponding land use/land cover and a condensed chart showing land cover changes for the three different periods.

From the above analysis (Figures 3 through to 8 and Table 1), we see that the spatial patterns of urban expansion in Al-Kharj can be categorised in three different intervals 2000, 2007 and 2013. Referring, specifically, to Table 1, it is evident that there is progressive urban expansion caused by the economic development and population growth that corresponds to the period of massive injection of funds that came with government's policy to expand the city over the period. Logically, this has led to a steady decrease in the land area used up by vegetation and trees. For similar reasons, the percentage of land cover used up by crops, which had increased to 12.78 % in 2007, nosedived to 1.24 % in 2013. This change can be linked to the drought-like shortage

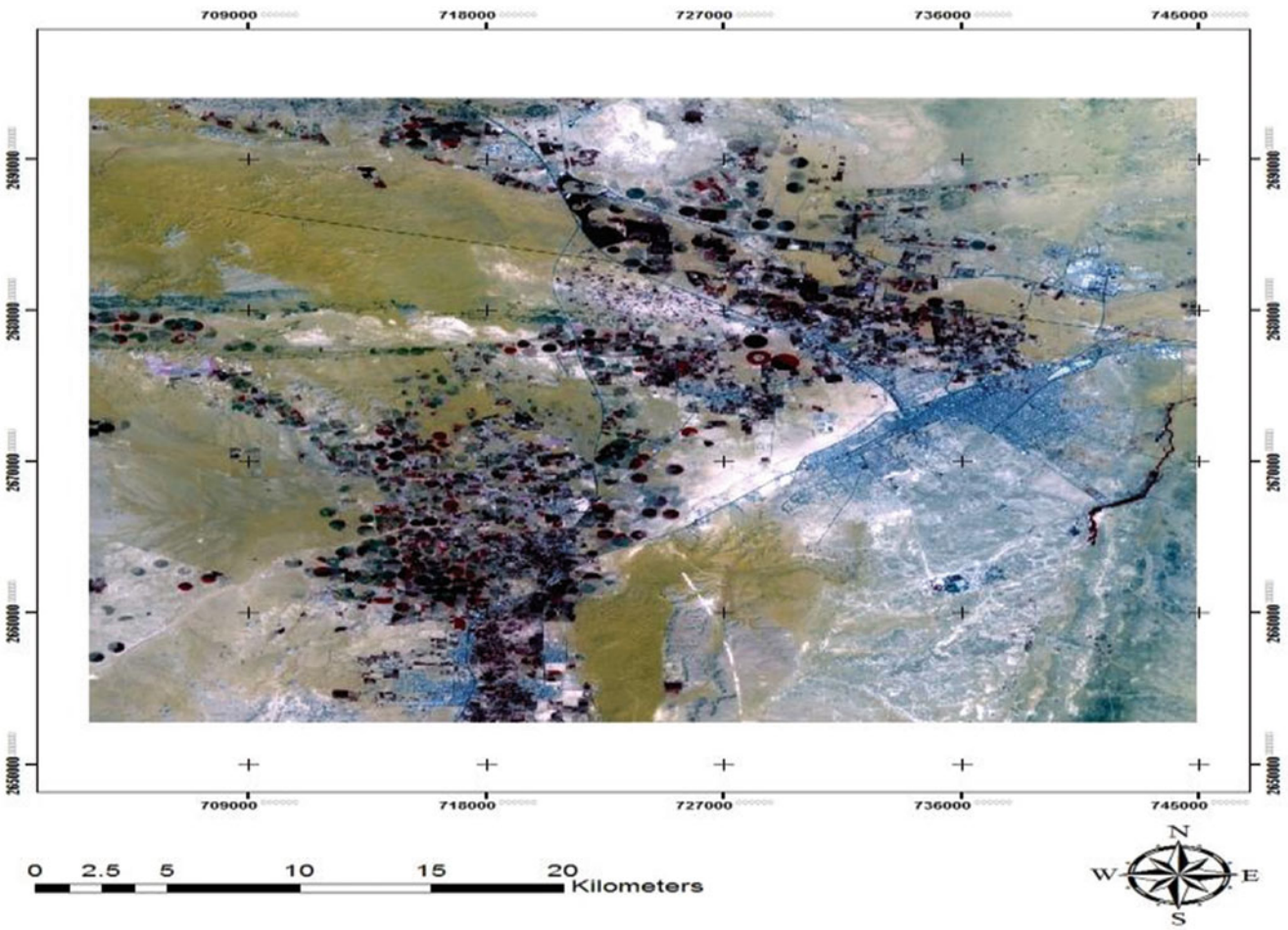


Fig. 3 The land use and land cover map of Al-Kharj in the year 2000.

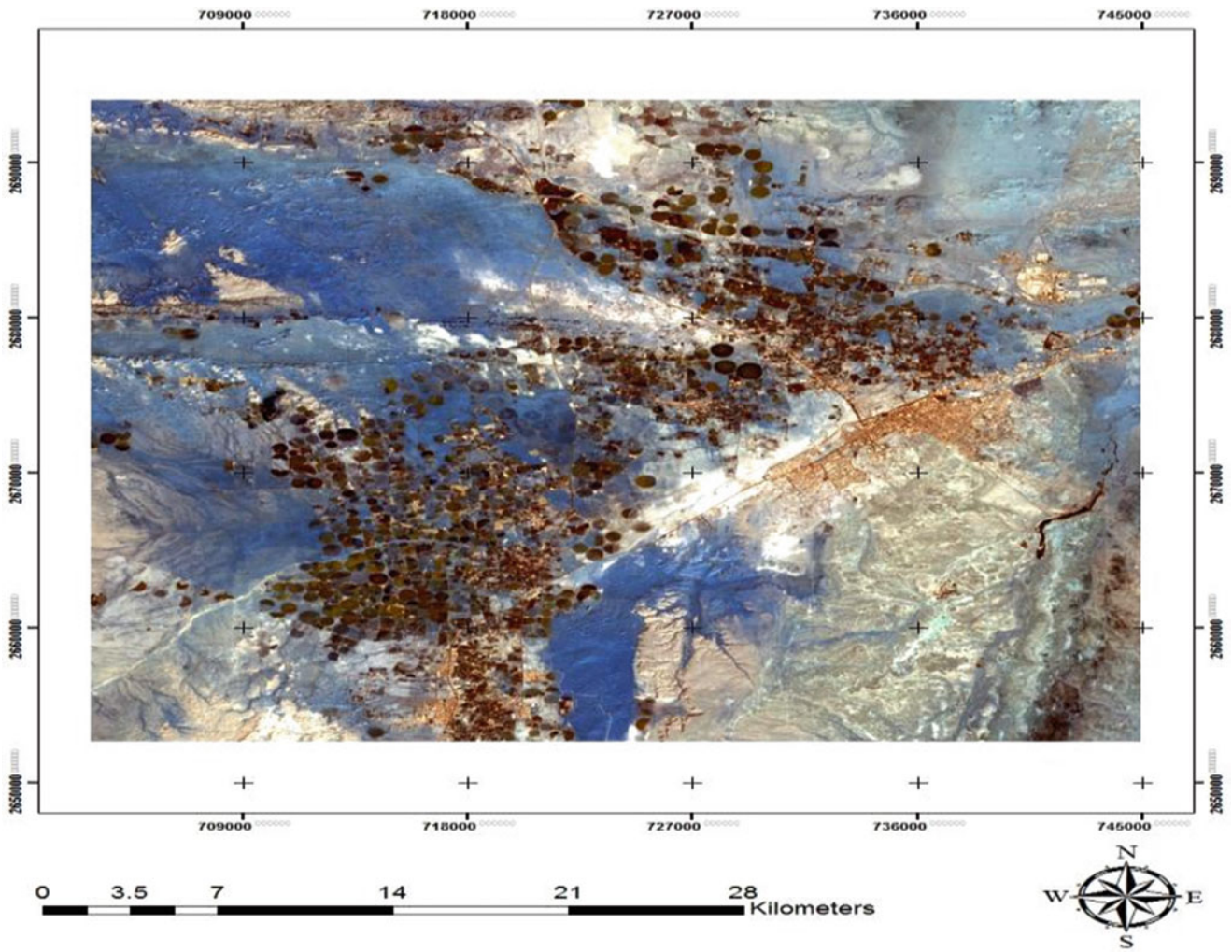


Fig. 4 The land use and land cover map of Al-Kharj in the year 2007.

of irrigation water needed to grow the crops during the period.

Although, some crops, such as palm trees, are known to be resistant to water shortages. However, the decrease in palm trees over the period is not directly attributed to shortage of water [12]. Rather, this was caused by the outbreak of the Red Palm Weevil pandemic witnessed during the same period. Areas of the Kingdom worst affected by this pandemic include Al-Kharj, Al Hasa, Qatif, Tabuk, Najran, Mecca, and Qasim. For instance, in 2010 alone, an estimated 60,000 palm trees were affected by the Red Palm Weevil, of which 30,000 trees had to be destroyed in Al-Kharj alone. In the meantime, as seen from the results, for each the remaining area land cover is taken up by soil and rocks.

Finally, the foregoing results are enough for us to infer that urban expansion in Al-Kharj region is following a spreading trend along radial corridors, such as major traffic lines, from the centre of the city. This is a very common phenomenon in the urban development of most cities in Saudi Arabia. One explanation often advanced for this pattern is that the traffic/industrial status determines this kind of urban growth pattern.

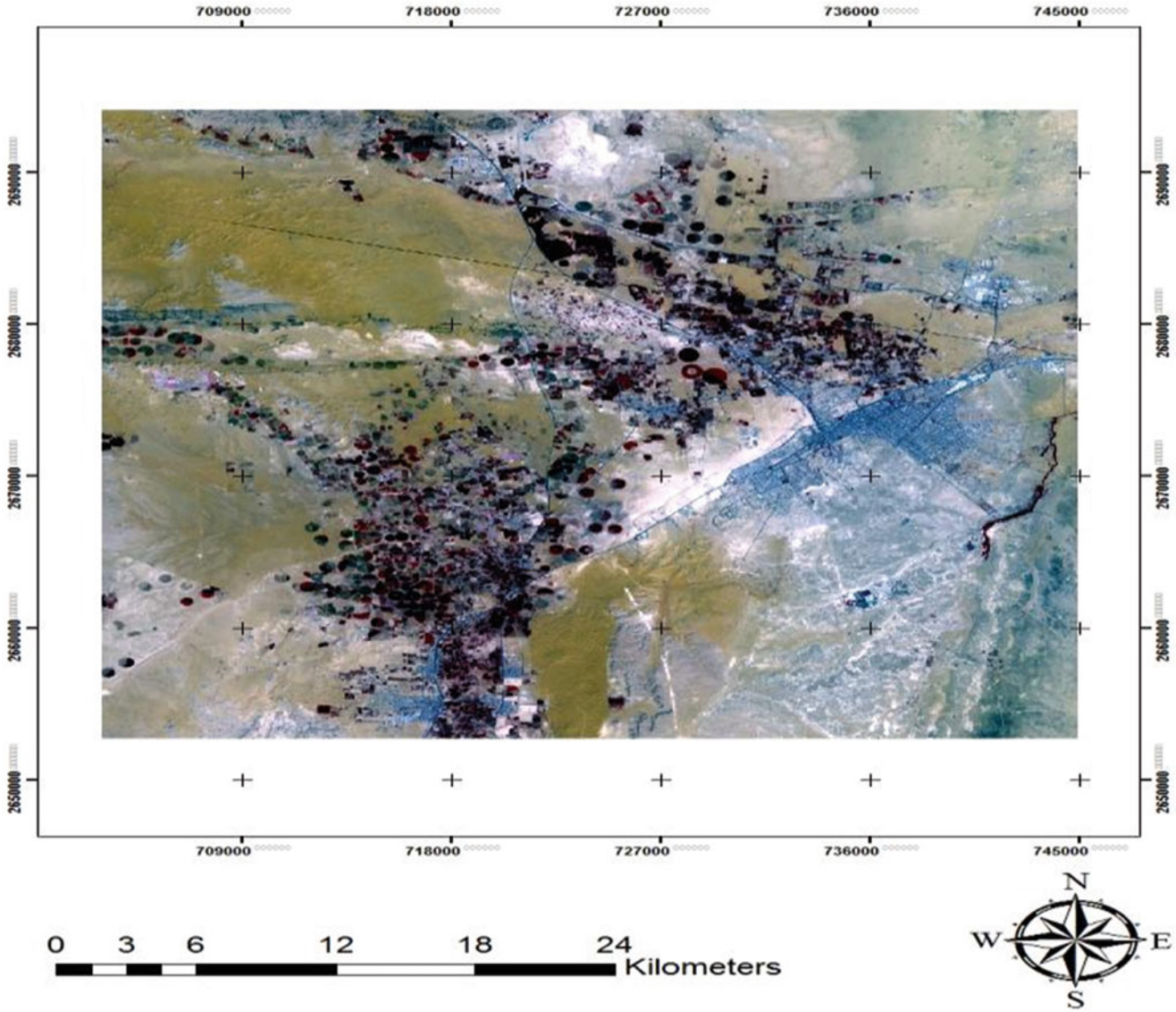


Fig. 5 The land use and land cover map of Al-Kharj in the year 2013

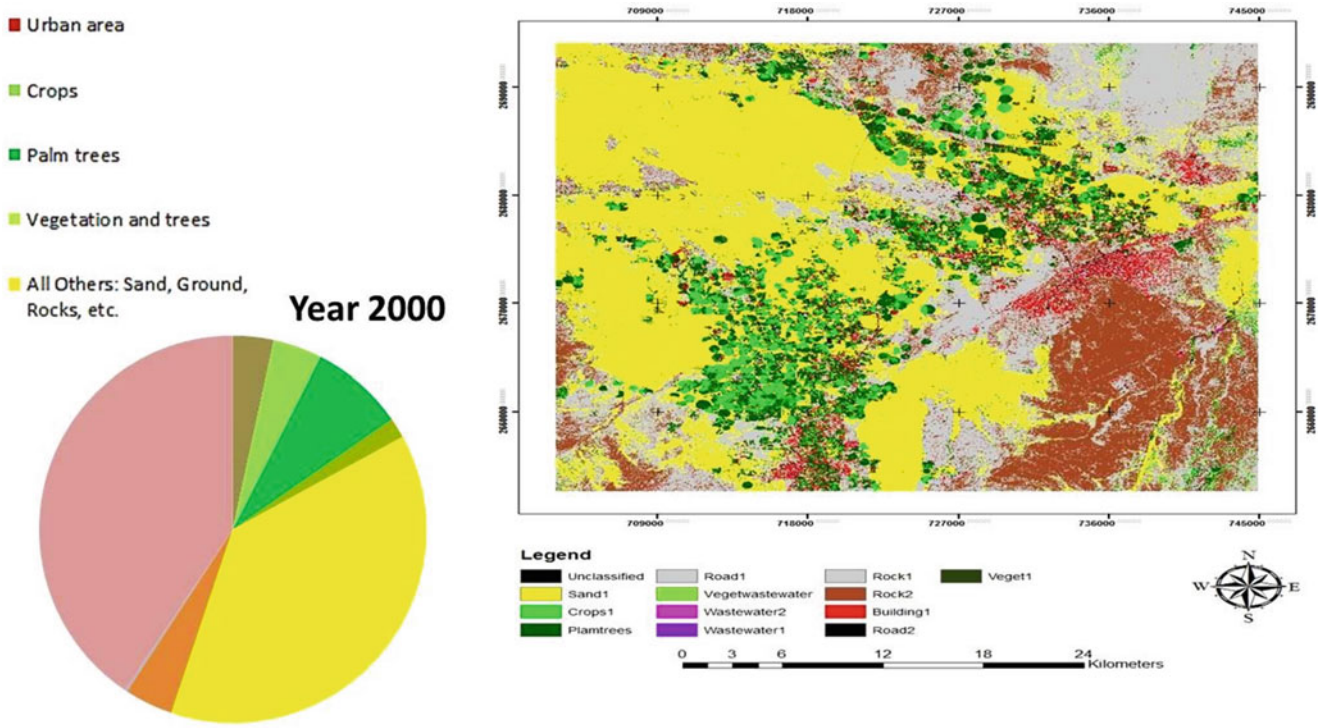


Fig. 6 The land use and land cover classification map of Al-Kharj in the year 2000.

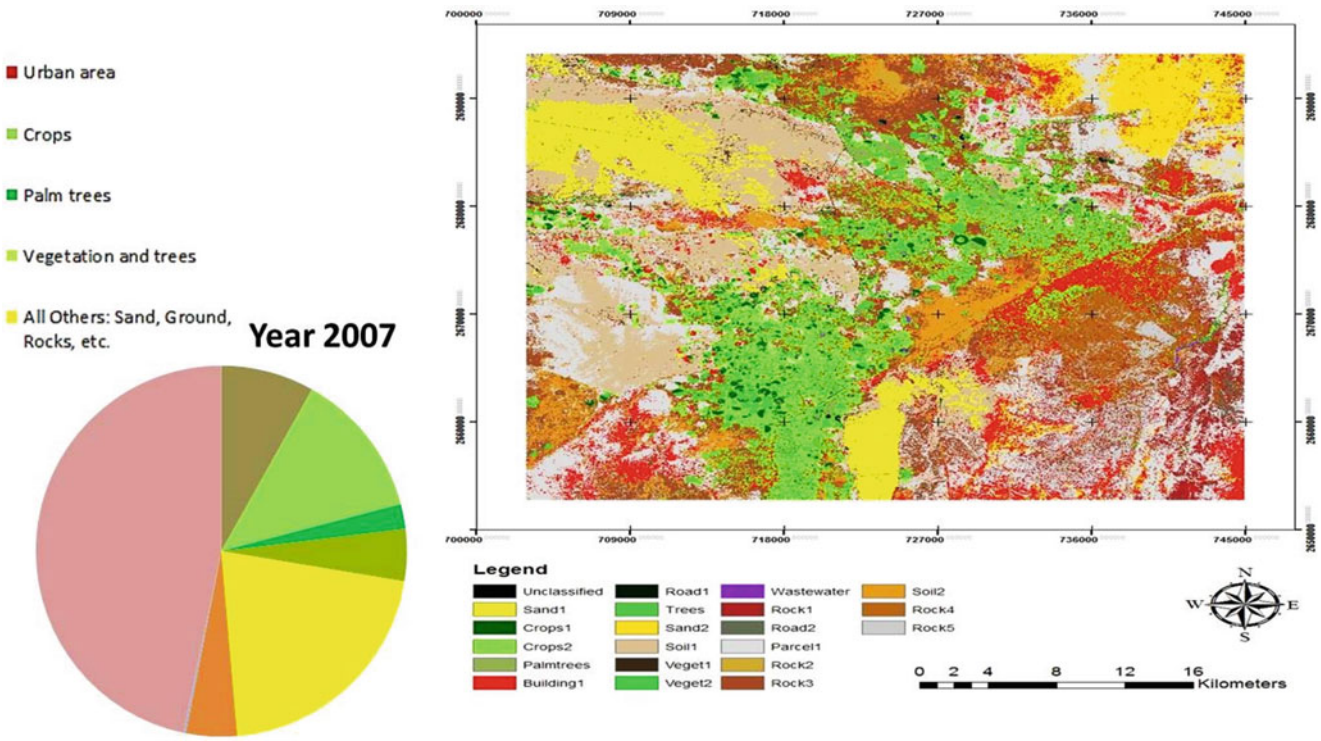


Fig. 7 The land use and land cover classification map of Al-Kharj in the year 2007.

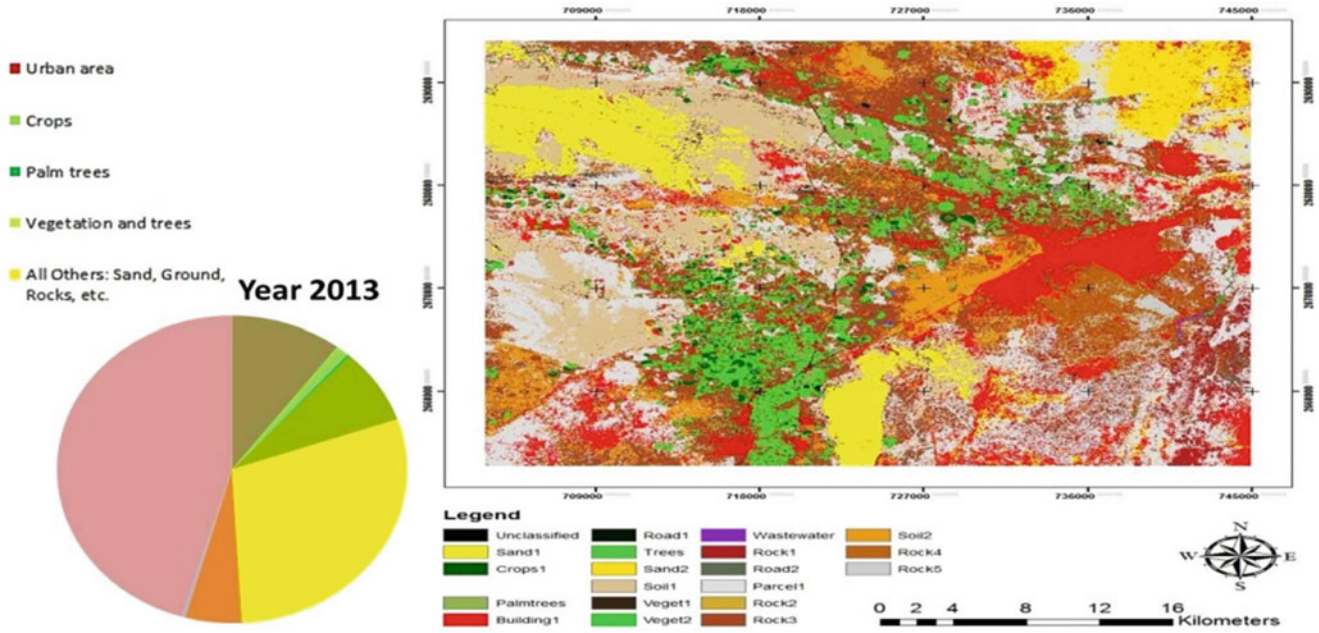


Fig. 8 The land use and land cover classification map of Al-Kharj in the year 2013.

Table 1 Percentage classification of land use/cover changes over the three periods (2000, 2007 and 2013)

	Year		
Land cover/Land use	2000	2007	2013
Urban area	3.41 %	8.14 %	10.29 %
Crops	4.09 %	12.78 %	1.24 %
Palm trees	7.84 %	2.13 %	0.20 %
Vegetation and trees	1.71 %	4.56 %	7.98 %
Ground Soils /Rocks	82.95 %	72.39 %	80.29 %

5 Conclusion

The study presented in this work analyses the land use and land cover changes in Al-Kharj city in Riyadh (central) Province of the KSA between the year 2000 to 2013 using GIS remote sensing techniques with particular emphasis on assessing the depletion in the agricultural land cover. The results show that during the period under review close to a quarter (25 %) of the total land area underwent one form of land use or land cover change or the other. Of this amount close to 15 % of the changes were directly related to changes that depleted the total green area of the city. This figure increases to a depletion of approximately 23 % in the total agricultural area of the city when the effects of Red Palm Weevil pandemic that occurred during the same period is taken into consideration.

While the aforementioned results indicate a reduction in agricultural activities in the city, its population has increased by 30 % over the same period. This increase is mainly attributed to the influx of people as a result of the rapid urbanisation witnessed in the city [4].

The foregoing results, combined with other environmental (desertification, pollution, ozone layer depletion, etc.) and natural factors, especially because of the city's proximity to Riyadh, the Kingdom's capital city, and government's planned development of Al-Kharj as one of the satellite cities (to reduce the burden on the capital) make the need to extend the results of the study along the directions of urban, regional and environmental planning of paramount importance. As an example, it would be crucial for the planned development of the city (along the directions mentioned earlier) if the correlation between the depletion in the water and other hydro-geological resources of the city and the land use and land cover changes reported in this study are assessed and evaluated to ascertain if they are in tandem with planned expansions envisaged by the Kingdom's policy makers.

From the image processing viewpoint, an algorithm to better classify the land use and land cover changes than those obtained using the GIS and ERDAS tools reported in this study is already being pursued.

Acknowledgements This study reported in this work was sponsored by the Salman Bin Abdulaziz University via the Deanship for Scientific Research.

References

1. I. I. Bashour, A. S. Al-Mashhady, J. D. Prasad, T. Miller, and M. Mazroa, "Morphology and composition of some soils under cultivation in Saudi Arabia," Geoderma, Vol. 29, No. 4, pp. 327–340, 1983.
2. A. Mukhopadhyay, J. Al-Sulaimi, E. Al-Awadi, and F. Al-Ruwaih, "An overview of the Tertiary geology and hydrogeology of the northern part of the Arabian Gulf region with special reference to Kuwait," Earth-Science Reviews, Vol. 40, No. 3-4, pp. 259–295, 1996.
3. M. M. Lababidi and A. N. Hamdan, "Preliminary Lithostratigraphic Correlation Study in OAPEC Member Countries, Organisation of Arab Petroleum Exporting Countries"GeoArabia, J. of the Middle East Petroleum Geosciences, Vol. 168, p.6 1985.
4. http://www.quandl.com/citypop/city_Al-Kharjrydsaudiarabia-Population-of-Al-Kharj-ryd-Saudi-Arabia. Accessed on 31 March 2014
5. A. S. Al-Farraj, M. I. Al-Wabel, M. H. El-Saeid, A. H. El-Naggar, and Z. Ahmed, "Evaluation of Groundwater for Arsenic Contamination Using Hydro-geochemical Properties and Multivariate Statistical Methods in Saudi Arabia",J. of Chemistry, Vol. 2013, http://dx.doi.org/10.1155/2013/812365.
6. J. G. Pike, "Groundwater resources and development in the central region of the Arabian Gulf," in Memoirs of the 18th Congress IAH Hydrogeology in the Services of Man, pp. 46–55, Cambridge, UK, 1985.
7. O. Quinn, "Regional hydrogeological evaluation of the Najd," Open-file Report CCEWR 48-86, Public Authority for Water Resources, Sultanate of Oman, 1986.
8. P. Beaument, "Water resources and their management in the Middle East," in Change and Development in the Middle East, J. I. Clarke and H. Bowen-Jones, Eds., pp. 41–72, Methuen, London, UK, 1981.
9. J. G. Pike, "Groundwater resources development and the environment in the Central Region of the Arabian Gulf," Water Resources Development, pp. 115–132, 1983.
10. J.Xiaoa, Y.Shenb, J.Gec, R.Tateishia, C.Tanga,Y.Liangd, Z. Huange, " Evaluating urban expansion and land use change in Shijiazhuang, China, by using GIS and remote sensing", Landscape and Urban Planning, Vol. 75, pp. 69–80, 2006.
11. Murai, S., "Remote Sensing Note", 2nd ed. Nihon Printing Co. Ltd., Tokyo, 1996.
12. http://www.redpalmweevil.com/arabicpage/arabic.htm. Accessed on 31 March 2014.

System Engineering Standards, Paradigms, Metrics, Testing, etc

Expert Systems Based Response Surface Models for Multidisciplinary Design Optimization

Ramesh Gabbur* and K Ramchand**

1 Introduction

Today aerospace design and development is not only multidisciplinary but also global in nature with design and engineering teams deployed around the world [1]. It requires high level of technical and techno-managerial expertise across various engineering disciplines to cater for stringent reliability, safety and performance requirement. This results in a high cost of development especially for combat aircraft with uncompromising performance requirement.

Present generation multi-role combat aircraft with fly by wire and state of the art weapons systems are complex systems in nature, which need specialists. Complexity of combat aircraft mandates the need for design teams to have multidisciplinary experience in the entire aircraft design with core expertise in their respective domains. This would enable design and development of optimal product/system in collaborative and cohesive integrated environment of various engineering domains.

Combat aircraft design and development is intensely model driven. Towards this systems engineering provides holistic approach for integrated design and development of aircraft and its associated systems [2]. One of the key challenges in collaborative design is integration of design and analysis methods of various systems in system engineering framework. Multidisciplinary system design is computationally intensive process that combines discipline analysis with design-space search and decision making. With the advances in Computer-Aided Design and Engineering (CAD/CAE) techniques, complex computer models and computation-intensive analyses/simulations (discipline analysis) are often used to accurately study the system behaviour towards design improvements. This design optimization process normally requires a large number of iterations before the optimal solution is identified. Design optimization, with high fidelity design tools, is computationally very expensive and time consuming. Typically approximation models or surrogates of high fidelity design tools are used to reduce this computational effort and time during multidisciplinary design optimization process. This paper brings out a expert systems based "*Smart RSM*" based methodology to generate approximation models in the design space around the point of interest with a limited number of computer experiments and *use of legacy data* for MultiDisciplinary Optimization (MDO).

2 Expert System for Surrogate Modelling

Complex aircraft engineering design problems are solved using high fidelity analysis/simulation software tools. The high computational cost associated with these analyses and simulations prohibits them from being used as performance measurement tools in the optimization of design for combat aircraft. Other Major drawback in using high fidelity analysis is numerical noise, which occurs as a result of the incomplete convergence of iterative processes, the use of adaptive numerical algorithms, round-off errors, and the discrete representation of continuous physical objects (fluids or solids)[3]. The use of approximation models or surrogates to replace the expensive high fidelity computer analysis, in MDO, is a natural approach to avoid the computation barrier and to take care of artificial minima

*Scientist, Aeronautical Development Agency and Doctoral Student at Jain University Bangalore, India
**President IDST and Research Guide at International Institute for Aerospace Engineering &Management, Jain University, Bangalore India

R. Gabbur (✉)
Aeronautical Development Agency, Bangalore, India
e-mail: gabbur@jetmail.ada.gov.in

K. Ramchand
Institute of Defence Scientists and Technologists (IDST), Bangalore, India

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_76,

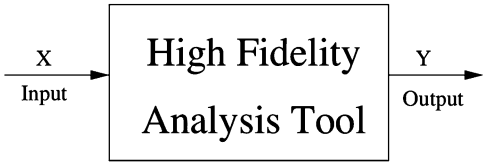


Fig. 1 High Fidelity Analysis

due to numerical noise. Renaud and Gabriele developed Response Surface Modelling (RSM) of multidisciplinary systems during concurrent subspace optimizations (CSSOs) [4] [5]. Korngold and Gabriele addressed discrete multidisciplinary problems using the RSM [6].

Expensive high fidelity computer analysis can be represented as black box function. In a simplest form the high fidelity analysis tools takes vector X as input and gives Y as the output as shown in Figure 1.

Representing it mathematically with limits on the design space

$$Y = f(\overline{x}) \qquad where \quad \overline{x} \in R^n$$
$$\overline{x}_{min} < \overline{x} < \overline{x}_{max} \quad Define\ the\ design\ space \tag{1}$$

This function would be replaced by polynomial based surrogate model. A typical second order model is shown below

$$y = \beta_0 + \sum_{i=1}^{n}\beta_i x_i + \sum_{i=1}^{n}\beta_{ii}x_i^2 + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\beta_{ij}x_i x_j$$
$$\overline{x}'_{min} < \overline{x}' < \overline{x}'_{max} \quad Define \quad the \quad model \quad subspace \tag{2}$$

Where β_i, β_{ii} and β_{ij} are regression coefficients. $\overline{x}$ is input vector and y is the output. The subspace of surrogate model is defined by the side constraints $\overline{x}'_{min} \quad and \quad \overline{x}'_{max}$.

2.1 Expert System

The flow chart for expert systems is shown in figure 2. It is a knowledge based algorithm which creates and maintains database for function calls and surrogates models. When the optimizer calls the analysis tool at the design point $\overline{x}$ through the expert system, it scans the I/O database for previous execution of the analysis tool at the point $\overline{x}$. If the points $\overline{x}$ exists then the corresponding $f(\overline{x})$ is returned to the optimizer. If the design point $\overline{x}$ does not exists in I/O database then it checks for availability surrogate models applicable for $\overline{x}$. The RSM database is scanned and surrogate models coefficient are returned for which $\overline{x}$ lies within the surrogate model subspace. If more then one models exist, then model for which $\overline{x}$ is closest to the center of the model subspace is used for returning $f(\overline{x})$.

If surrogate model is not available for $\overline{x}$, smart RSM is called to generate surrogate model. If the surrogate model is created by smart RSM, then the model is used to return $f(\overline{x})$. The RSM database is updated for values of model coefficient and its subspace.

There could be instances when surrogate model is not created if the subspace is smaller than a defined useful limit. In such instance analysis tool is called to return $f(\overline{x})$ to optimizer and RSM database is updated appropriately to prevent Smart RSM being called again for points lying within this subspace.

2.2 Smart RSM

The flow chart for smart RSM is shown in figure 3. Smart RSM algorithm aims at creating a response surface model and identifying the subspace for which model is valid. It comprises of six steps to generate a validated surrogate models and its subspace. The design space of the optimization problem is defined as the initial subspace for the model. The surrogate model is validated against analysis tool to a predefined accuracy. If the error values are above acceptable limits, then subspace for the surrogate model is reduced. This process is repeated iteratively until a satisfactory model is generated for the subspace or the subspace is smaller than the defined useful limit.

3 Design of Experiments - Latin Hypercube Design

A quadratic model of the form given in equation 2 is chosen to fit the response surface. The number of experimental design points (k) needed to fit a response surface should be equal or greater than the total number (p) of unknown coefficients $\beta's$ in equation 2.

$$p = (n+1)*(n+2)/2$$
$$k \geq p \qquad where \quad n \quad is \quad the \quad dimension \quad of \quad \overline{x}$$

The experimental design techniques developed for physical experiments using classical techniques of experimental blocking, replication, and randomization are irrelevant when it comes to deterministic computer experiments [7, 8]. Experimental design techniques for computer experiments should be space filling for obtaining information from the entire design space [9].

Widely used space-filling sampling methods are Orthogonal Array (OA) and Latin Hyper cube Design (LHD) [10]. OA can generate a sample with better space-filling property than LHD. However, the generation of an OA sample

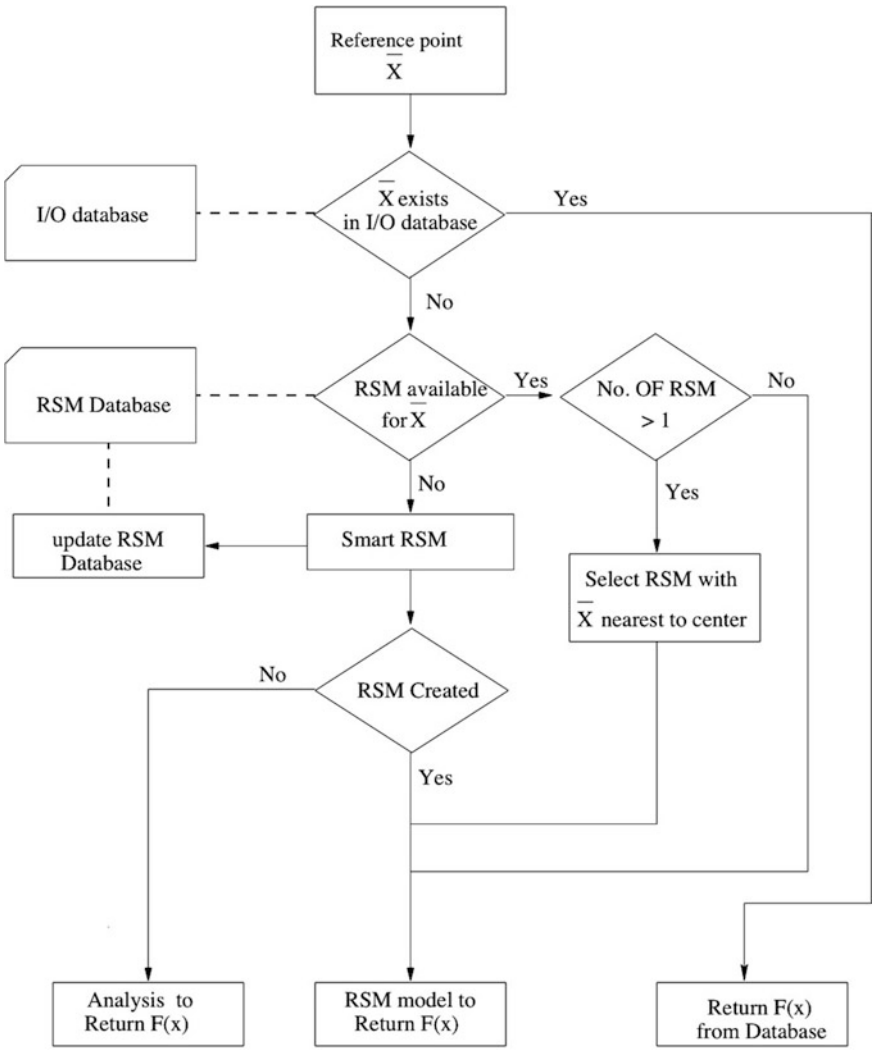


Fig. 2 Flow Chart for Expert Systems

is more complicated than LHD [11] [12]. In addition, OA demands strict level classification for each variable, which might bring difficulty in execution of experimental design. In real design, not all combinations of variable level lead to realistic design solutions, and some may cause the crash of the analysis or simulation. In that case, the engineers must manually adjust variables to an appropriate number, deviating from one of the defined levels. Thus the property of OA might be undermined [13]. LHD offer flexible sample sizes while ensuring stratified sampling, wherein each of the input variables is sampled at k levels. Due to this Latin Hypercube design is selected as a experimental design for generating RSM.

4 Model Validation and Reduced Model Subspace

Surrogate model fitted to the experimental points should capture the low frequency and high frequency behaviour of the analysis tool. Hence the validation of the model is done in two phases around a point in design space called check point. The check point is generated in such a way that if the model fails in the validation process then checkpoint should lie in the new reduced subspace.

When surrogate model does not accurately represent the analysis tool for the given design space then the design space

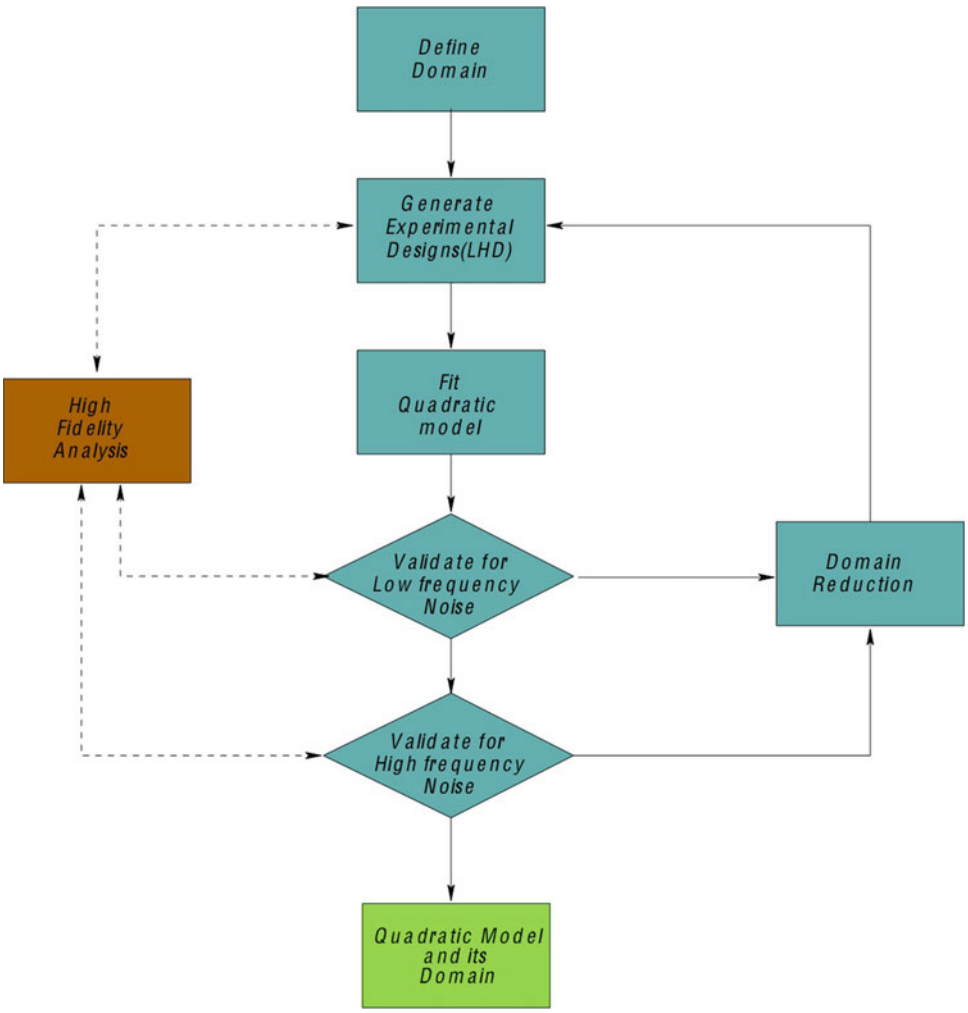


Fig. 3 Flow Chart for Smart RSM

needs to be reduced. A selective reduction of design space is employed. The strategy is to halve the design space for that variable for which the error residual is more than 1% (or user specified model accuracy) and design space is reduced (zoomed in) around the reference point. Mathematically x_{ri} is the reference point for i^{th} input and x_{li} and x_{ui} are its lower and upper limits respectively then the new design space lower limit$x_{li}^{'}$ and upper limit $x_{ui}^{'}$ are

$$x_{li}^{'} = x_{ri} - \frac{x_{ui} - x_{li}}{4} \tag{3}$$

$$x_{ui}^{'} = x_{ri} + \frac{x_{ui} - x_{li}}{4} \qquad for \quad i \quad = \quad (1, 2....n) \tag{4}$$

$$\begin{aligned} &if \qquad x_{li}^{'} < x_{li} \qquad then \\ &x_{li}^{'} = x_{li} \quad and \quad x_{ui}^{'} = x_{li} + \frac{x_{ui} - x_{li}}{2} \end{aligned} \tag{5}$$

$$\begin{aligned} &if \qquad x_{ui}^{'} > x_{ui} \qquad then \\ &x_{ui}^{'} = x_{ui} \qquad and \qquad x_{li}^{'} = x_{ui} - \frac{x_{ui} - x_{li}}{2} \end{aligned} \tag{6}$$

5 Test Cases

There are no standard set of test problems for testing of surrogate models. Problems used for testing global optimization algorithms are used here [14]. These sets of test cases are of varying complexity and range from simple smooth two dimension tests functions to complex test function in ten dimensions and are listed below

1. Branin function
2. McCormic function
3. Levy function
4. Box and Betts exponential quadratic sum
5. Rosenbrock function

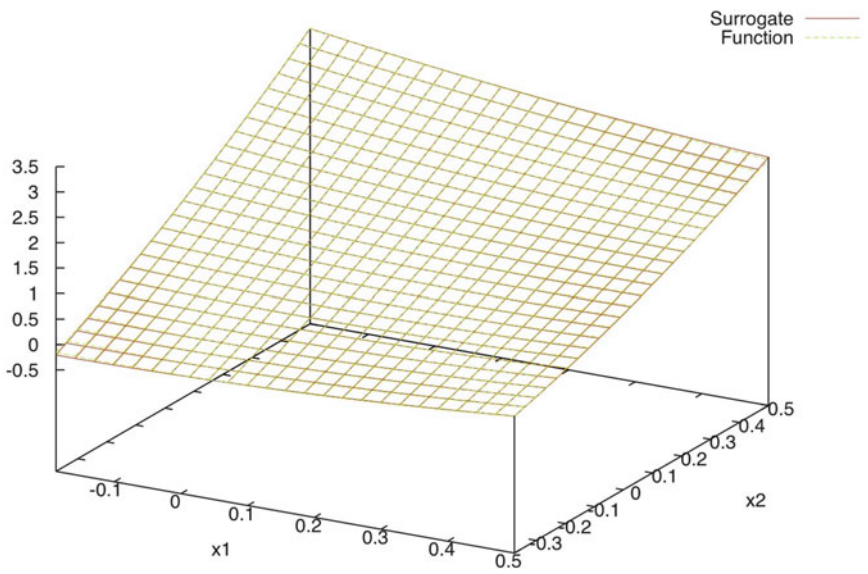


Fig. 4 Surface plot for McCormic Function and Model after 5 iteration

The algorithm was evaluated on the five test function. Validated models with its reduced subspace were generated. At first iteration the model is fitted to the full subspace. This model is validated and if the model is not satisfactory, the domain is reduced by the algorithm. This constitutes one iteration. This process is repeated for iteration number 2, 3 and so on until a validated model is achieved. These models were again checked for their accuracy by using 100 points randomly generated in the design space. The result along with the error plots, model subspace, number of iteration and number of function calls were analysed.

5.1 Branin Function

The Branin function given below, is smooth function in two dimension design space($\mathcal{R}^2$).

$$f(\overline{x}) = (1 - 2x_2 + \tfrac{1}{20}\sin 4\pi x_2 - x_1)^2 + (x_2 - \tfrac{1}{2}\sin 2\pi x_1)^2 \tag{7}$$

Surrogate model was generated after one iteration. The model was generated using 8 experimental points in the design space. The residuals of 100 randomly generated points in the design space was less than 10^{-10}.

5.2 McCormic Function

The function with its initial subspace is given below

$$\begin{gathered} f(\overline{x}) = \sin(x_1 + x_2) + (x_1 - x_2)^2 - 1.5x_1 + 2.5x_2 + 1 \\ \overline{x} \in \mathcal{R}^2 \; -1.5 \le x_1 \le 4 \quad and \quad -3 \le x_2 \le 4 \end{gathered} \tag{8}$$

For a reference design point [0,0], a validated model for McCormic function is generated after five iterations. Figure 4 show combined surface plots for function and the validated surrogate model after the five iteration. Figure 5 shows residuals for 100 points randomly generated in the design space

5.3 Levy Function

The function with its initial design space is given below,

$$\begin{gathered} f(\overline{x}) = \sin^2(3\pi x_1) + \sum_{i=1}^{n-1}(x_i^2 - 1)^2(1 + \sin^2(3\pi x_{i+1})) \\ +(x_n - 1)(1 + \sin^2(3\pi x_n)) \end{gathered} \tag{9}$$

$$\begin{gathered} \overline{x} \in \mathcal{R}^n \quad where \quad n = 4, 5, 6, 7 \\ -10 \le \overline{x} \le 10 \quad for \quad n = 4 \quad and \\ -5 \le \overline{x} \le 5 \quad for \quad n = 5, 6, 7 \end{gathered}$$

For ($n = 4$) and reference design point as [0,0,0,0] a validated quadratic model has been generated after 9 iteration. The residuals for 100 randomly generated points is shown in figure 6.

5.4 Box and Betts Exponential Quadratic Sum

The function with its initial design space is given below

$$\begin{gathered} f(\overline{x}) = \sum_{i=1}^{10} g_i(\overline{x})^2 \quad where \\ g_i(\overline{x}) = \exp(-0.1ix_1) - \exp(-0.1ix_2) - (\exp(-0.1i) - \exp(-i))x_3 \\ \overline{x} \in \mathcal{R}^3 \quad 0.9 \le x_i \le 1.2 \quad for \quad i = (1,3) \quad and \quad 9 \quad \le x_2 \le 11.2 \end{gathered} \tag{10}$$

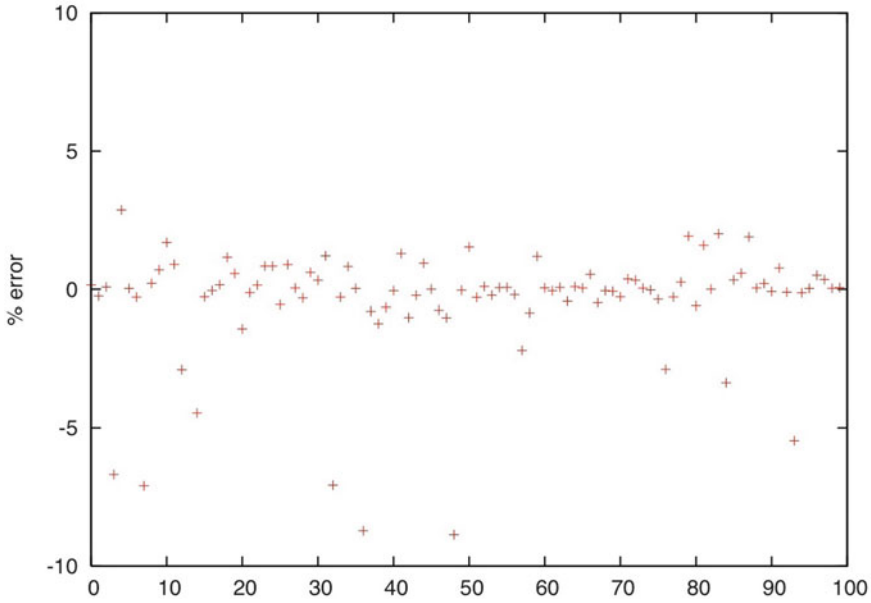


Fig. 5 Residuals for McCormic model at 100 random points

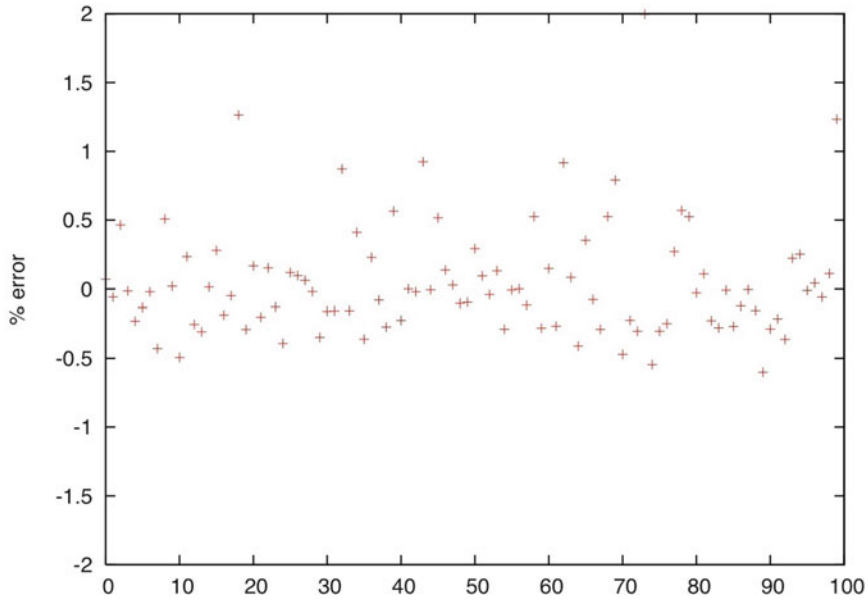


Fig. 6 Residuals Levy Function Model for 100 random points

At reference design point [1,11,1] a validated model was generated after 5 iteration. Figure 7 shows Residuals for 100 random points.

5.5 Rosenbrock Function

This is a very complex function in two dimensions and with numerous local minima. The function and its design space are given below

$$f(\overline{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \quad where \quad \overline{x} \in \mathcal{R}^2 \quad -2 \leq \overline{x} \leq 2 \tag{11}$$

Figure 8 shows Surface plot for Rosenbrock function for the full design space.

With a reference design point of [0,0], a validated model has been generated in 7 iteration. Figure 9 Shows the residuals for 100 random points.

6 Result

Table 1 is summary of the results for test functions indicating the number of iterations, initial design space and model subspace of validated surrogate model.

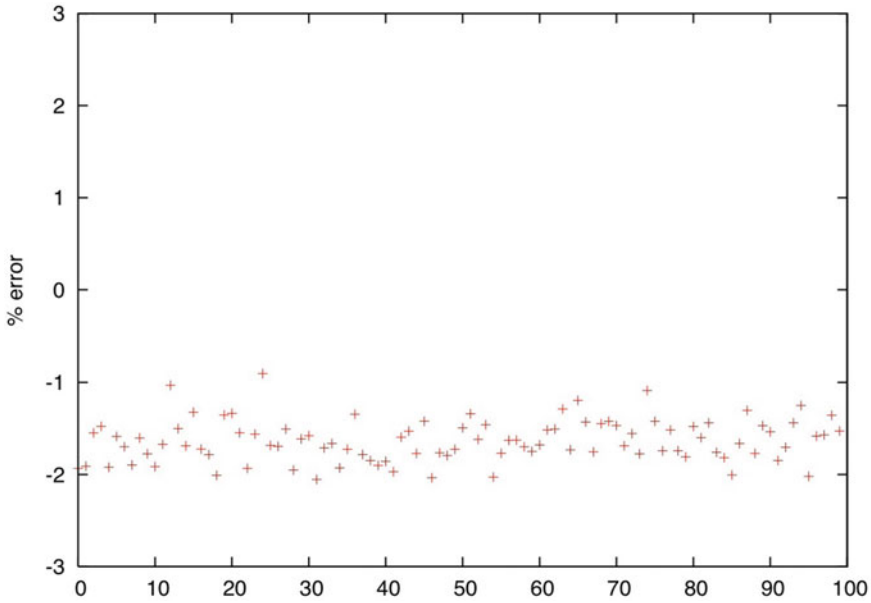


Fig. 7 Residual plots for Model of Box Betts function

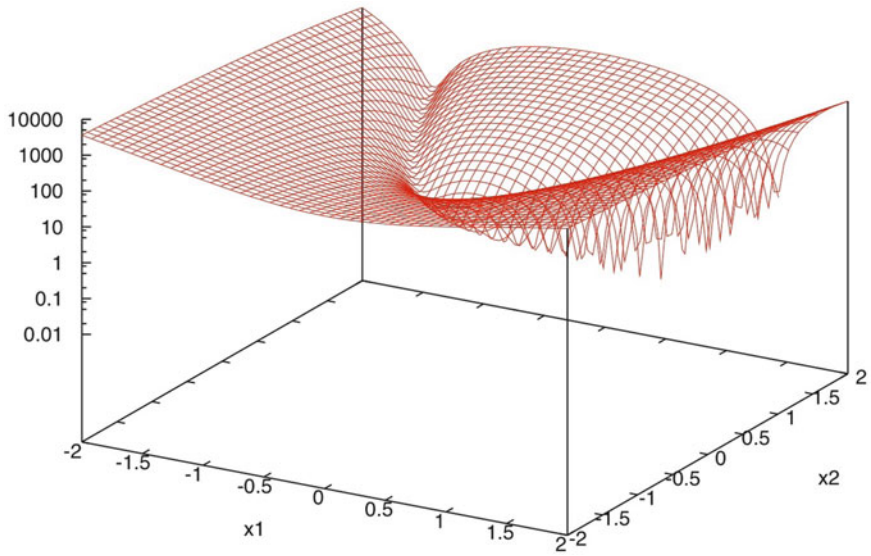


Fig. 8 Surface plot for Rosenbrock function

Table 2 shows the details regarding the number of function evaluations for model fitting and validations for each iteration for Levy Test Function (n=4). Column 2 indicates the number of experimental points required for fitting a surrogate model. Column 3 shows the legacy data used and column 4 indicates additional function calls. The last column indicates the reduction in number of function call with use of the expert system.

7 Conclusion

In conventional methods the model subspace is defined prior to generating the model and the accuracy of the model is not defined. Use of expert systems based algorithm generates surrogate models, for both simple and complex function, to any user defined accuracy levels. The difference would be in the reduction of model subspace with increase in number of iterations for complex functions.

The expert systems creates database for functions calls and surrogates models know as legacy data i.e. knowledge base. The number of function calls needed for model generation is reduced by the use of this legacy data. For a typical complex function, like Levy function (n=4), the number of additional function calls required with the use of knowledge base/legacy design data is reduced by about 30%.

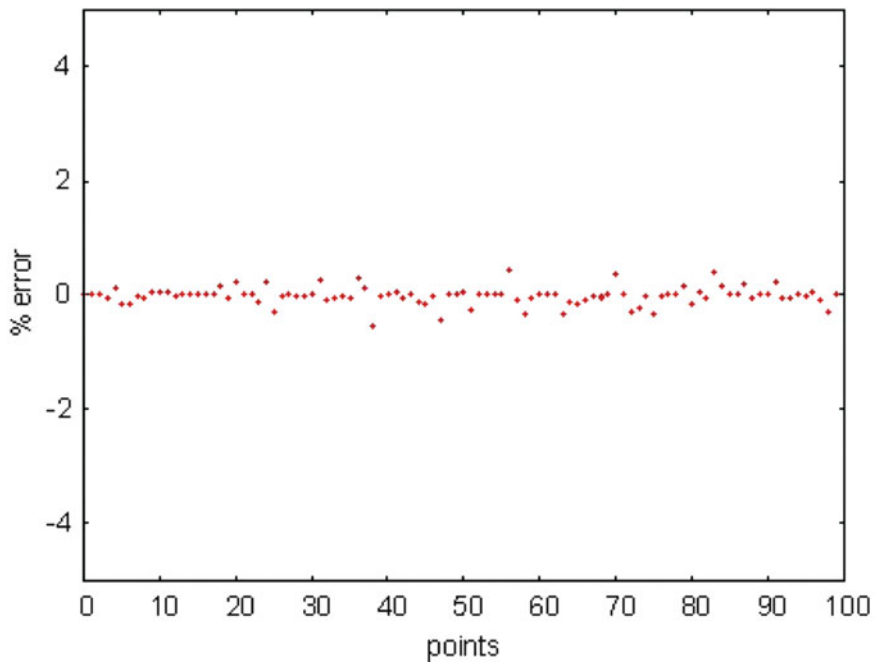


Fig. 9 Residual plot of Rosenbrock function Model

Table 1 Test Function and its Model Subspace

Test Function	No of Iteration for model generation	Initial Design space	Model subspace
Branin function	1	$\bar{x} \in \mathcal{R}^2$ $-10 \leq \bar{x} \leq 10$	$\bar{x} \in \mathcal{R}^2$ $-10 \leq \bar{x} \leq 10$
McCormic function	5	$\bar{x} \in \mathcal{R}^2$ $-1.5 \leq x_1 \leq 4$ *and* $-3 \leq x_2 \leq 4$	$x_1 = [-0.18750$ *to* $0.50000]$ $x_2 = [-0.37500$ *to* $0.50000]$
Levy function	9	$\bar{x} \in \mathcal{R}^n$ *where* $n = 4, 5, 6, 7$ $-10 \leq \bar{x} \leq 10$ *for* $n = 4$ $-5 \leq \bar{x} \leq 5$ *for* $n = 5, 6, 7$	$x_1 = [-0.03906$ *to* $0.03906]$ $x_2 = [-0.03906$ *to* $0.03906]$ $x_3 = [-0.03906$ *to* $0.03906]$ $x_4 = [-0.03906$ *to* $0.03906]$
Box and Betts exponential quadratic sum	5	$\bar{x} \in \mathcal{R}^3$ $0.9 \leq x_i \leq 1.2$ *for* $i = (1, 3)$ *and* $9 \leq x_2 \leq 11.2$	$x_1 = [0.98750$ *to* $1.02500]$ $x_2 = [10.8750$ *to* $11.01250]$ $x_3 = [0.99375$ *to* $1.01250]$
Rosenbrock function	7	$\bar{x} \in \mathcal{R}^2$ $-2 \leq \bar{x} \leq 2$	$x_1 = [-0.03125$ *to* $0.03125]$ $x_2 = [-0.03125$ *to* $0.03125]$

Table 2 Function calls for Levy Function

Iteration Number	Total points for LHD	Legacy data (previous iteration)	Additional Function calls	Reduction in function calls
1	19	0+0	19	0 %
2	19	0+5	14	26.3 %
3	19	0+5	14	26.3 %
4	19	1+5	13	31.6 %
5	19	3+5	11	42.1 %
6	19	1+5	13	31.6 %
7	19	2+5	12	36.8 %
8	19	1+5	13	31.6 %
9	19	1+5	13	31,6 %
Total	171		122	28.7%

Legacy data = earlier data + validation points of previous iteration

References

1. D R Towill. Man-machine interaction in aerospace control systems. *The Radio and Electronic Engineer*, 50(9):447–458, September 1980.
2. M Price, S Raughunathan, and R Curran. An integrated systems engineering approach to aircraft design. *Progress In Aerospace Science*, 42:331–376, 2006.
3. A Giunta A, Dudley J M, Narducci R, Grossman B, Haftka R. T, Mason W H, and Watson L. T. Noisy Aerodynamic Response and Smooth Approximations in HSCT design. *Proceedings of the 5th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, (94-4376):pp 1117–1128, Sep 1994.
4. Renaud J. E and Gabriele G. A. Improved coordination in non-hierarchical system optimization. *AIAA Journal*, 31:pp 2367–2373, 1993.
5. G. A. Renaud J. Eand Gabriele. Approximation in non-hierarchical system optimization. *AIAA Journal*, 32:pp 198–205, 1994.
6. J. C. Korngold and G. A. Gabriele. Multidisciplinary analysis and optimization of discrete problems using response surface methods. *Journal of Mechanical Design*, 119:pp 427–433, 1997.
7. J Sacks, W.J Welch, T.J Mitchel, and H.P Wynn. Design and analysis of computer experimant. *Statistical Science*, 4:pp 409–435, 1989.
8. Soma Roy and William I. Notz. Estimating percentiles in computer experiments: A comparison of sequential-adaptive designs and fixed designs. *Journal of Statistical Theory and Practice*, 8(1):12–29, 2014.
9. BartG.M. Husslage, Gijs Rennen, EdwinR. Dam, and Dick Hertog. Space-filling latin hypercube designs for computer experiments. *Optimization and Engineering*, 12(4):611–630, 2011.
10. J. R. Koehler and A. B Owen. '*Computer Experiments*' *Handbook of Statistics*, volume pp 261–308. Elsevier Science, New York, 1996.
11. G Taguchi, Y Yokoyama, and Wu. Y. Taguchi methods: Design of experiments. *American Supplier Institute, Allen Park, Michigan.*
12. A Owen. Orthogonal arrays for computer experiments, integration, and visualization. *StatisticaSinica*, 2:pp 439–452, 1992.
13. G. Gary Wang. Adaptive response surface method using inherited latin hypercube design points. *ASME, Journal of Mechanical Design*, 125:pp 210–220, June 2003.
14. Test problems for global optimization. url= http://www.imm.dtu.dk/~km/GlobOpt/testex/.

A Survey of Approaches used in Parallel Architectures and Multi-core Processors, For Performance Improvement

Surendra Kumar Shukla, CNS Murthy, and P. K. Chande

1 Introduction

Multi-core processors are consist with more than one CPUs [1]. CPUs exists in a single chip. These type of processors are also called chip multiprocessors. Multi-core processors are different from multiprocessors, in the context that-multiprocessors exits in septate chip[1]. Multi-core processors are developed to utilise the parallelism, in exist in program. Multi-core processors are now growing phenomenally in the market. The reason behind the popularity of multi-core processors is their ability to exploit the parallel processing concepts. Now a days multi-core processors are used every where like mobile phones, security systems. In future, there are enormous opportunities that are waiting for the multi-core processors.

Multi-core processors, that are coming in market, have two types. Homogeneous and heterogeneous multi-core processors. Heterogeneous multi-core processors are more powerful in terms of speed as compare to homogeneous multi-core processors. It is established fact, that to exploit the processing power available in multi-core processors, programs must have to be designed in such a way that, programs can be executed parallel. Algorithm designers must have to design the algorithms in such a way, so that programmer will not have to worry about the parallelism property of the program. There are so many parallel programming languages available in the market, but application designer are use-to make the programs in sequential programming languages. It is also necessary that researchers must have to know the clear difference or similarity on the concepts that are used in parallel architectures and multi-core architectures. If all the concepts which are used in parallel architectures can also be used in the multi-core architectures, it will be easy for the growth of the multi-core architecture. With the help of the parallel architectures techniques which has already been used, it will be helpful in the evolution of multi-core architecture. In this review paper, we will find-out the concepts, approaches that are used in parallel architectures and also can be used in multi-core architecture. We are also intent to investigate new techniques.

S.K. Shukla (✉)
Department of CSE, CDGI, Indore, M.P, India
e-mail: surendr.shukla@gmail.com

C. Murthy
Department of CSE, CDGI, Indore, M.P., India
e-mail: cnsmurthy@gmail.com

P.K. Chande
School of Computer Science & IT DAVV Univercity, Indore, M.P., India
e-mail: pkchandein@yahoo.co.in

2 Approaches Used in Parallel & Multi-core architectures

2.1 Challenges of Multi-core architecture [1,2]

Software may run slower In a single core environment resources are not shared between concurrent software. In multi-core, resources such as memory buses and L2 cache are shared so there is the potential for software on one core to interfere with software running on another core. This is because software on one core could be using a resource that software on the other core requires. In a single core environment this wouldn't cause any delay but in a multi-core environment it can cause not only software to run slower but can also make it less deterministic.

Software may run less efficiently Multi-core CPU's are designed to optimise average-case execution time often at the expense of worst case execution time. This means that the delta between worst case and average execution time becomes much greater, leading software designers to inflate the budgets required to meet worst case timings. This can lead to software actually running

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_77,

less efficiently on a multi-core architecture than a single core one.

(iii) Two cores communicates through the MPI

2.2 Multi-core vs parallel processors

Multi-core systems are a subset of parallel systems. Different systems will have different memory architectures, each with their own set of challenges. How does one system deal with cache coherence? NUMA involved.

UMA- in UMA all processing nodes takes same memory access time, for all the portion of memory.

Drawback:

(I) Problem in cache coherence

(ii) Scalability problem.

NUMA- Here Processor/Memory nodes are on network. Processing node will take long time to access some regions of memory. Memory access time is not same for all portion of the memory.

Advantage: It is Scalable.

Difference between NUMA and Distributed memory architecture[3]. No processor can have mapping to memory connected to other processor in case of Distributed memory architecture. **Multi-core vs Multiprocessor**

- Multiprocessor is any computer with several processors, - Multi-core processor is a special kind of Multiprocessor.
- In Multi-core processor all processors are on the same chip.

Multi-core processors are MIMD Different cores execute different threads(Multiple instructions) operating different part of memory(Multiple data).

Multi-core is a shared memory multiprocessor. All cores share the same memory. Applications with thread level parallelism can be benefited from the mulyicore architecture.

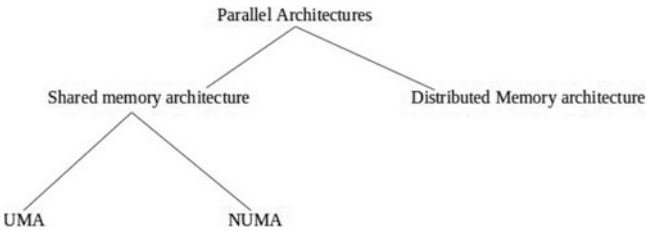


Fig. 1. Parallel architecture types

2.3 Classification of parallel architectures

Why parallel processing?

(I) Sequential architectures reaching physical limitations (speed of light), we can not increase the speed of sequential processors now. Heat problem, cooling requirements.

(ii) Computational requirements are ever increasing, weather-forecasting.

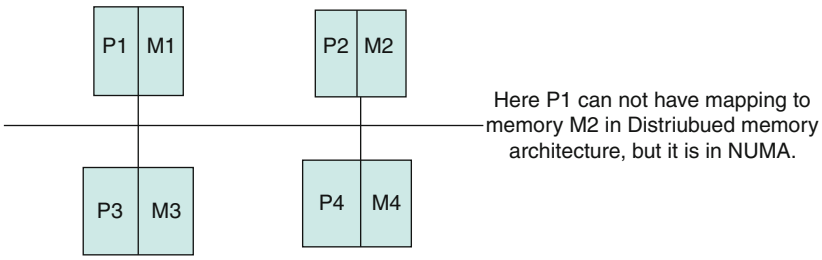


Fig. 2. NUMA vs Distributed memory model

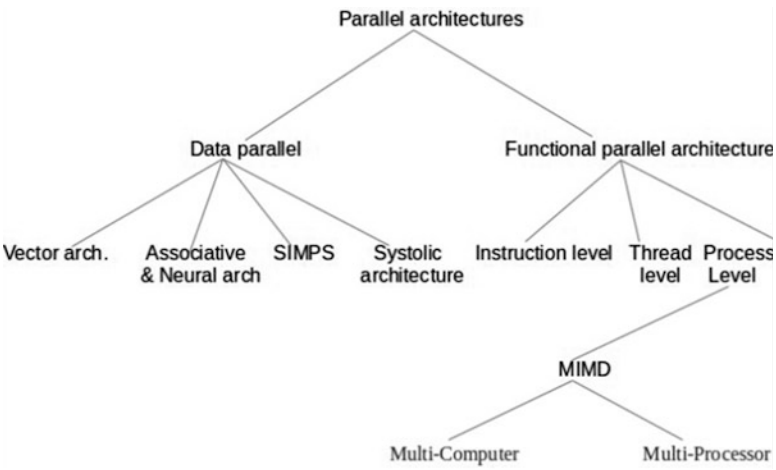


Fig. 3 Detailed classification of parallel architectures

(iii) Hardware improvements like pipe-lining, super-scalar etc. are non scalable & required sophisticated compiler technology.

Heterogeneous vs Homogeneous Multi-core architecture?

Heterogeneous Multi-core architecture: Multi-core architecture where every core is just an image of the other is called homogeneous multi-core architecture.

Heterogeneous Multi-core architecture is a set of cores which may differ in area, performance, power dissipated etc.

Hyper-threading Vs Dual core:

Hyper-threading- here Single physical core but it uses it efficiently. Multi-core capability provide two set of executing resources or two or more cores.

Applications must have to change to take the advantage of the multi-core architecture. Mother-board cost increases as the number of cores increases.

Multi-core processors plug directly into a single processor socket, but the operating system perceives each of its execution cores as discrete logical processor.

Multi-core chips do more work per clock cycle, and can be designed to operate at lower frequencies.

3 Multi-core Basics

it isn't specific to any one multi-core design, but rather is a basic overview of multi-core architecture. Although manufacturer designs differ from one another, multi-core architectures need to adhere to certain aspects.

Closest to the processor is Level 1 (L1) cache; this is very fast memory used to store data frequently used by the processor. Level 2 (L2) cache is just off-chip, slower than L1 cache, but still much faster than main memory; L2 cache is larger than L1 cache and used for the same purpose. Main memory is very large and slower than cache and is used, for example, to store a file currently being edited in Microsoft Word. Most systems have between 1GB to 4GB of main memory compared to approximately 32KB of L1 and 2MB of L2 cache. Finally, when data isnt located in cache or main memory the system must retrieve it from the hard disk, which takes exponentially more time than reading from the mem-ory system.

3.1 Evolution of Multi-core Architectures

Prior to 2003, the traditional methods for boosting processor performance were to increase the clock frequency, add high-speed, on chip cache, and optimize instructions [2]. These traditional methods worked for many years until physical issues limited the processor clock frequency to around 4 GHz in 2003. Although physical issues limited the processor clock frequency, the transistor still continues to shrink. Multi-core architectures capitalise on smaller process sizes and increased transistor counts to provide multiple execution units on a single chip. The following four multi-core processor architectures illustrate the variety and complexity of multi-core architectures.

3.2 Cell Broadband Engine Architecture (CBEA)

The CBEA processor application environment ranges from gaming consoles (e.g., Play Station 3) to high performance supercomputing (e.g., the Roadrunner super computer at Los Alamos National Laboratory). The CBEA provides the Power Processing Element (PPE) as a single, general purpose Simultaneous Multi-threading (SMT) processor core, with two logical cores. The CBEA also provides Synergistic Processor Element (SPE) cores (typically eight) which are designed for computationally intensive tasks. These specialized cores implement a different instruction set from the general purpose SMT core. Each specialized core has its own local memory store.

Each CBEA component is connected via four one-way buses configured in a ring. Two of these rings run clockwise while the other two rings run counter-clockwise. The CBEA uses Direct Memory Access (DMA) to move data between the

The CBEA provides a hardware security architecture where one or more of the specialized cores can be put into a secure processing mode. When in this secure processing mode, the hardware isolates the core so no other core, operating system, or hypervisor can interrogate the internal state of the isolated core.

3.3 TILE64 multi-core

The TILE64 multi-core processor contains 64 independent, general purpose cores. Each core has its own private Level 1 (L1) and Level 2 (L2) cache as well as a distributed Level 3 (L3) virtual cache. The cores are connected using non-blocking switches in an intelligent mesh (iMesh). The iMesh provide extremely low-latency and high-bandwidth communication between the cores, memory, and other I/O. Each core is capable of running an independent operating system, or can be grouped to run a Symmetric Multiprocessing (SMP) operating system.

Fig. 4 CBEA Processor

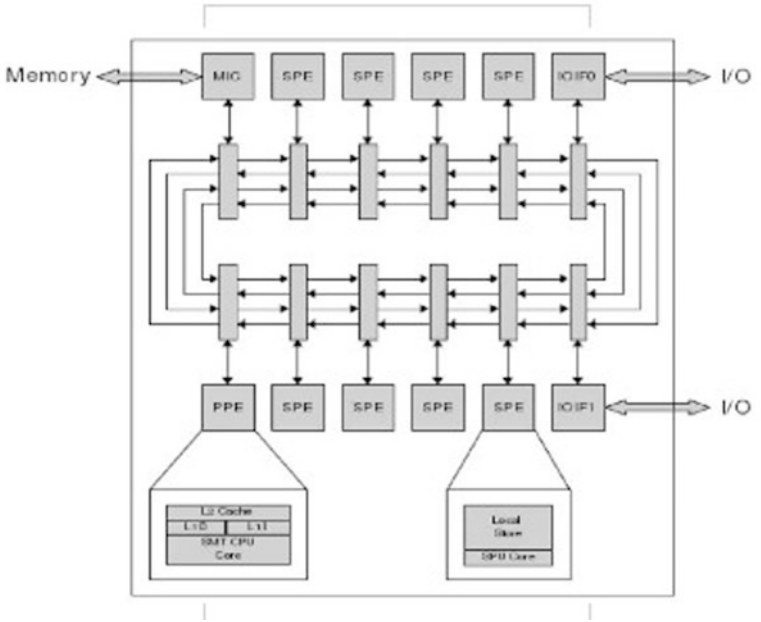
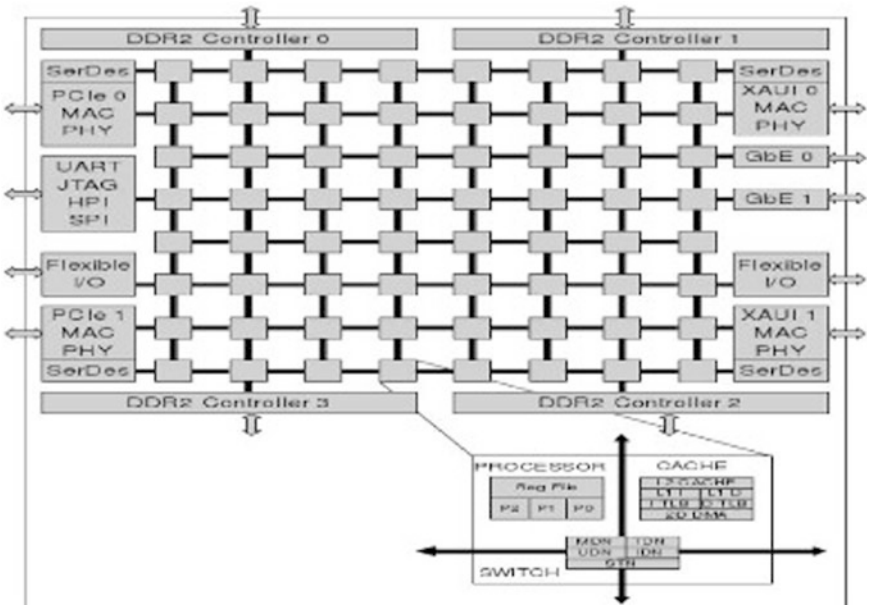


Fig. 5 TILE64 Processor



3.4 Intel Core i7

The Intel Core i7 processor is targeted towards general purpose computing. Multi-core performance improvements are achieved by: (1) replicating the core execution unit, (2) adding additional processor instructions, and (3) improving the communication between the cores. Two major changes for the Core i7 architecture include the large shared L3 cache and the Quick-path interconnect bus which is used for inter-process communication.

3.5 Freescale P4080

The Freescale Semiconductor QorIQ P4080 communications processor provides eight Power Architecture cores, each with integrated L1 and L2 caches. The chip is intended for embedded systems and includes a variety of memory and I/O controllers. This processor supports a multi-megabyte shared L3 cache. Hardware provides acceleration for encryption, regular expression pattern matching, and Ethernet Packet Management. On-chip components are

Fig. 6 intel Corei7 Processor

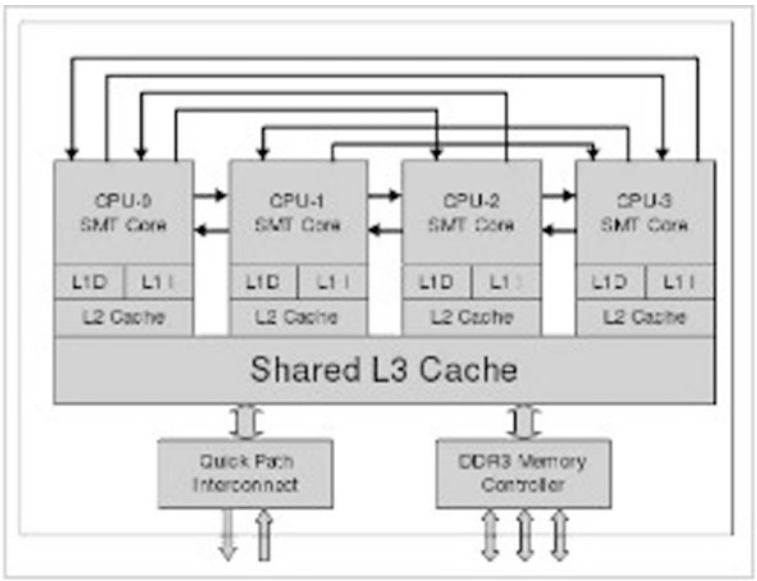
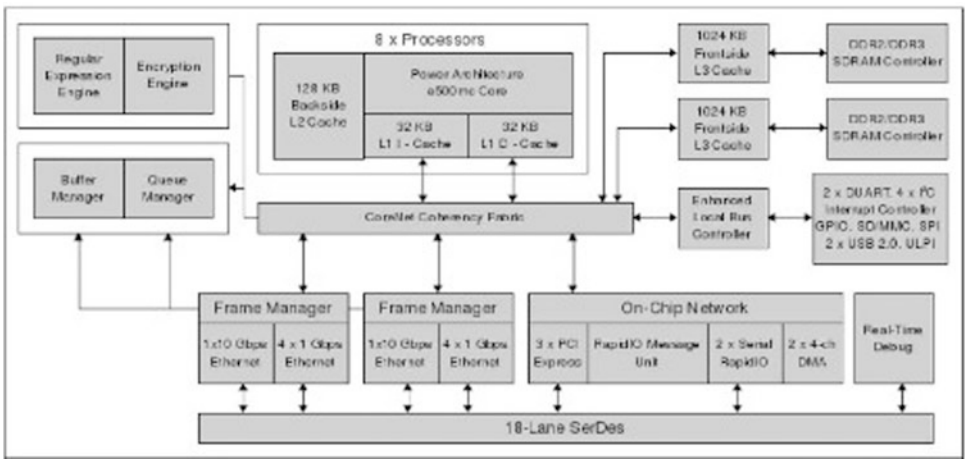


Fig. 7 P4080 Processor



connected by the CoreNet coherency fabric which manages full cache coherency between the caches and point-to-point, concurrent connectivity between the hardware components. Figure 4 shows a block diagram of the P4080.

3.6 Access pattern DMA in Multi-core architecture[5]

DMA is used for intra-chip data transfer in multi-core processors, especially in multiprocessor system-on-chips, where its processing element is equipped with a local memory (often called scratchpad memory) and DMA is used for transferring data between the local memory and the main memory.

DMA Modes of operation-

Burst mode- block of data is transferred at a time from main memory to peripherals

Cycle stealing mode- one word is transferred by the DMA. So in one cycle DMA uses main memory with the help of the system bus and in next time CPU.

Transparent Mode- In this mode DMA uses the system bus when CPU is doing other operation. But it require extra circuitry to detect when CPU is not using the system bus.

DMA can lead to cache coherency problems

DMA on the ISA bus has been stuck at the same performance level for over 10 years. For old 10 MB XT hard disks, DMA was a top performer. For a modern 8 GB hard disk, transferring multiple megabytes per second, DMA is insufficient.

DMAs are used most commonly today by floppy disk drives, tape drives and sound cards.

As an example of DMA engine incorporated in a general-purpose CPU, newer Intel Xeon chipsets include a DMA engine technology called I/O Acceleration Technology (I/OAT), meant to improve network performance on high-throughput

network interfaces, in particular gigabit Ethernet and faster. However, various benchmarks with this approach by Intel's Linux kernel developer Andrew Grover indicate no more than 10% improvement in CPU utilization with receiving workloads, and no improvement when transmitting data.

As an example usage of DMA in a multiprocessor-system-on-chip, IBM/Sony/Toshiba's Cell processor incorporates a DMA engine for each of its 9 processing elements including one Power processor element (PPE) and eight synergistic processor elements (SPEs). Since the SPE's load/store instructions can read/write only its own local memory, an SPE entirely depends on DMAs to transfer data to and from the main memory and local memories of other SPEs. Thus the DMA acts as a primary means of data transfer among cores inside this CPU (in contrast to cache-coherent CMP architectures such as Intel's cancelled general-purpose GPU, Larrabee).

DMA in Cell is fully cache coherent (note however local stores of SPEs operated upon by DMA do not act as globally coherent cache in the standard sense). In both read ("get") and write ("put"), a DMA command can transfer either a single block area of size up to 16KB, or a list of 2 to 2048 such blocks. The DMA command is issued by specifying a pair of a local address and a remote address: for example when a SPE program issues a put DMA command, it specifies an address of its own local memory as the source and a virtual memory address (pointing to either the main memory or the local memory of another SPE) as the target, together with a block size. According to a recent experiment, an effective peak performance of DMA in Cell (3 GHz, under uniform traffic).

reaches 200GB per second.

3.7 Use of different level of caches[6]

The L1 cache is accessed on every instruction cycle as part of the instruction pipeline and is broken into separate instruction and data caches. Thus, the design of the L1 cache should be to maximise the hit rate (the probability of the desired instruction address or data address being in the cache) while keeping the cache latency as low as possible. Intel uses an L1 cache with a latency of 3 cycles.

The L2 cache is shared between one or more L1 caches and is often much, much larger. Whereas the L1 cache is designed to maximise the hit rate, the L2 cache is designed to minimise the miss penalty (the delay incurred when an L1 miss happens).

For chips that have L3 caches, the purpose is specific to the design of the chip. For Intel, L3 caches first made their appearance in 4 way multi-processor systems (Pentium 4 Xeon MP processors) in 2002. L3 caches in this sense greatly reduced delays in multi-threaded environments and took a load off the FSB. At the time, L3 caches were still dedicated to each single core processor until Intel Dual-Core Xeon processors became available in 2006. In 2009, L3 caches became a mainstay of the Nehalem microprocessors on desktop and multi-socket server system

3.8 Performance parameters of Multi-core Architecture [7,8]

Pipe-lining

Multiple pipeline can be added to fetch and issue more than one instruction in parallel, creating supersaclar processing element to increase performance.

Drawback: increasing issue width require extra logic, hazard detection.

(ii)Performance can also be increased by increasing the number of pipeline stages

Drawback: Huge penalty in case of branching instructions

In-order execution-

inorder elements have small die area, low power.

out of order architecture it attempts to dynamically find and schedule multiple instructions "out -of- order to keep the pipeline full. Dynamic scheduling requires very complex circuitry and huge power requirement.

VLIW architecture

VLIW is not useful if compilar will not detect the parallelism.

Memory system

with multi-cores the caches are just one part of the memory system, the other components include the consistency model, cache coherence support and intra-chip interconnect.

Strong consistency model:

strict ordering constraints how in memory read/write operations will be performed.

Strong consistency model is slower and makes memory model complex.

cache configuration

large cache must be used where data is frequently used.

Intra-chip interconnects

there are many styles of interconnects bus, ring, crossbar, each have their advantage and disadvantage. Example bus is easy to design but quickly becomes bandwidth and latency limits.

3.9 Performance issues on multicore Processors

With multi-cores instructions are allowed to run parallel on individual cores simulteniously. It incraeses the amount of parallelism.Total number of cores which can be present into

the single chip are restricted today i.e 8,16 cores.We need to analyze the reason, challenges on it, if we increase the cores then what will be the effect & how performance can be increased.

Some performance factors are which affects the performance

(i) Thermal constraints, overheating effects
(ii) Interconnection network
(iii) Memory issues i.e. Memory latency
(iv) Bandwidt
(v) parallelism

4 Parallel Processors and parallel languages

Types and levels of parallelism:

Available Parallism- parallelism available in program or in the problem solution.

Utilized Parallelism- Parallelism utilized during execution of the program.

Available parallelism has two types -

f**unctional parallelism**- parallelism found in the logic of the problem.

Data parallelism- available in the data structure.

Available parallelism can be utilized by the architecture (instruction level parallel architectures) Compilars- Parallel optimizing compilars. Operating system- Multi-tasking

User level parallelism can be exploited with the help of multi-programming/time sharing

4.1 Parallel programming languages and models for multicore processors [9]

Parallel programming is the spliting of a single task into a number of sub task that can be computed independently. Parallel programming is most often used for tasks that can be easily broken down into independent tasks, such as purely mathematical problems, eg. Factorisation. Problems such as these are known to be embarasingly parallel.

Embarasingly parallel means those problems where finding the parallelism is very easy, they are by nature parallel.

4.2 Parallel and distributed programming models

MapReduce

MapReduce is programming model for large-scale computations.

MapReduce is implemented as runtime library.

MapReduce provides autometic parallelization, load balancing, locality optimization, handeling of machine failure.

MapReduce is a programming model for processing large data sets with a parallel distributed algoritham on a cluster.

"**Map**" **Step:** The master node takes the input, devides it into smaller-problems & then distributed it into worker nodes. A worker node may again do this and forms the tree structure. worker node process the smaller problem and passes the answer to the master node.

"**Reduce**" **Step:** The master node then collects the answers to all the sub-problems and combines them into some way to form the output.

Programs written in this functional style are autometically parallelized and executed on a large cluster of machines. The responsibility of partioning the input data, scheduling the program execution across a set of machines, managing the required inter-machine communication is taken by run-time system.

This helps to those programmers who has not much understanding of parallel and distributed system, can do the programming without any difficulty.

Dryd

- Similar goals as MapReduce
- focus on throughput not latency
- Automatic management of scheduling, distribution, fault tollerance.
- Computation expressed as graphs
- Vertices are computations, edges are communication channel.

Pig latin

Highlevel procedural abstraction of MapReduce contains SQL like premtives.

Green-Marl

- High-level grapgh analysis language/compilar
- Built-in grapgh functions

4.3 A- brief history of languages

When Vector machines were king:

- Parallel languages were (loop annotations) IVDEP
- Performance was fragile
- There was good user support.

When SIMD Machines were king

- Data parallel languages popular and successfull
- languages was CMF, LISP*, C*
- Quit powerful: can handel irregular data(sparse matrix)

When shared memory multiprocessor(SMP) were king

Shared memory model eg. OpenMP
PosixThreads are popular.
When cluster took over
– Message passing(MPI) become popular.

Older languages
– UPC, CAF and Titanium.
– All three uses an SPMD execution model
– advantages: portable, simple performance (some time better than MPI)
3 Newer HPC languages:
– X10, Fortress, and chapel
– All these use a dynamic parallelism model with data parallel constructs.

4.4 Parallel languages: Past, present and future

Data parallelism
Data-parallel languages have been around for a long time, but interest in them has surged recently because of the availability of massively parallel hardware. These languages are attractive because parallelism is not expressed as a set of processes whose interactions are managed by the user, but rather as parallel operations on aggregate data structures.

Flat data-parallel languages
The aggregate data structures of a flat data-parallel language cannot contain members that are themselves aggregates.

I. An examination and comparison of the features found in a number of data-parallel languages [SB91].
II. C*. C extended with domains, poly and mono variable classes, and reduction operations [
III. Fortran 90. Triplet notation for array sections. Operations and intrinsic functions on array sections
IV. HPF. Similar to Fortran 90, but includes data layout specifications to help the compiler generate efficient code
V. Implementation of C* for MIMD machines
VI. APL: the oldest and most influential of data-parallel languages
VII. UC: Another C based data-parallel language
VIII. Apply. Regular communication patterns, mappings for various architectures. Designed for image processing.

Message passing
These languages represent processes that communicate through messages.

A. CSP. Seminal work on formulating parallelism as a network of sequential processes exchanging messages
B. Occam. Real language based on Hoare's CSP
C. Ada tasking and rendezvous facilities
D. A comprehensive survey of message passing languages, both parallel and distributed. This article has an exhaustive list of languages for distributed systems and further references

Functional languages
Functional languages are referentially transparent and have no state. This feature has been exploited in various ways to gain parallelism.

1. Multi-Lisp. Lisp with futures. A *future* is a promise to provide the value of a computation if needed; pointers to the future may be freely manipulated. Run-time system decides when futures should create new processes. An introduction to Multi-lisp [Hal85] and developments in MultiLisp, particularly the difficulties of implementing both futures and explicit continuations

Qlisp. Provides *qlet* and *qlambda* constructs for spawning parallel processes. User

4.5 Parallel algorithams for Multiprocessor Vs Parallel algorithams for Multicore architecture Parallel algorithams [10]

Algoritham which can be executed by more than one processors at a time.algoritham which can not be executed parallel are inherently sequential eg. Newtons method.

Diffrence Between Concurrency and Parallelism?

Concurrency: is a propert of a program(at design level) where to or more taskcan be in progress simulteniously.

Parallelism: is a run time property where two or more tasks are being executed simulteniously.

There is a diffrence between being in progress and being executed since first one is not necessary involve being in execution.

Let
C = Concurrency
P = Parallelism
P is proper subset of C
it means parallelism required concurrency but concurrency does not require parallelism.

Parallel Random access model (PRAM) is used in Parallel algorithams.

PRAM Model
– A number of processors all can access
– a large share memory
– all processors are synchronized.
– all processors running the same program
– each processor have id, pid

Approaches to design parallel algorithams:

- Modify an existing sequential algoritham. Exploit those part of the algoritham that are naturally parallelizable.
- Design a completly new parallel algoritham, that may have no natural sequential algoritham.

where

Synchronized and asynchronous parallel algorithms for multiprocessors[11]

A synchronized parallel algorithm is a parallel algorithm consisting of processes with the following property:

There exists a process such that some stage of the process is not activated until another process has finished a certain portion of its program.

The needed timing can be achieved by using various synchronization primitives.

For example, suppose that we want to compute (A×B) + (C×D×E) by two processes.

We may construct a parallel algorithm by creating process P1 consisting of only one stage, X «- AxB, and process P2 consist- ing of two stages, Y <- C×D×E and S consisting of X+Y. Clearly, the activation of the second stage of process P^ is subject to the condition that process P1 is complete.

Asynchronous Parallel Algorithms

An asynchronous parallel algorithm is a parallel algorithm with the following properties:

(i) There is a set of global variables accessible to all processes.
(ii) When a stage of a process is complete, the process first reads some global variables.

Then based on the values of the variables together with the results just obtained from the last stage, the process modifies some global variables, and then activates the next stage or terminates itself. In many cases, to ensure logic correctness, the operations on global variables are programmed as critical sections. Thus in an asynchronous parallel algorithm, the communications between processes are achieved through the global variables, or shared data. There is no explicit dependency between processes, as found in synchronized parallel algorithms. The main characteristic of an asynchronous parallel algorithm is that its processes never wait for inputs at any time but continue or terminate according to whatever information is currently contained in the global variables. It is called an "asynchronous" parallel algorithm because synchronizations are not needed for ensuring that specific inputs are available for processes at various times

5 Conclusion

It has been observed that parallel architecture techniques and approches, are useful for the growth of the multi-core architecture. There is a need of powerfull programming languages, that should be developed and used by the programmers, so that parallel programs can be developed for the multi-core architecture. In future new techniques and approaches shoould be developed in the area of operating system, scheduling algorithms, memory management for the multi-core architectures. New approches should be developed that are not yet developed in parallel architectures. Development of techniques must be independent with other parallel architectures, e.g. Multiprocessors. We should not always dependent on the approches develpoed for the parallel architectures.

Acknowledgements The authors thanks to management of Chameli Devi Group of Institutions, Indore, M.P. India for providing excellent research enviourment in the college. A Special thanks to the Chairman **Shri Vinod Kumar Agrawal ji**, CDGI Indore, India for motivating to the faculties of the institutioon on research activities.

References

1. Multi-core architectures. Jernej Barbic. 15-213, Spring 2007. May 3, 2007.
2. Ryan Bradetich, et al., “Evaluating Multicore Architectures for Application in High Assurance Systems”, Center for Secure and Dependable Systems, University of Idaho.
3. Blaise Barney, “Introduction to Parallel Computing”, Livermore National Laboratory, Lawrence
4. Nakul Manchanda and Karan Anand, “Non-Uniform Memory Access (NUMA)”, New York University.
5. Selma Saïdi, “Optimizing DMA Data Transfers for Embedded Multi-Cores”, in PhD disseratation, univercity of Grenovale, october 2012.
6. Bryan Schauer, “Multicore Processors – A Necessity”, ProQuest Guide, September 2008.
7. Norm Matloff, “Programming on Parallel Machines”, University of California, Davis
8. Douglas Comer, “Computer Organization”, Computer Science Department Purdue University.
9. Jeffrey Dean and Sanjay Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters”, Google, Inc.
10. Richard M. Karp and Vijaya Ramchandran, “A survey of parallel algorithms for shared-memory machines”, univercity of california at Berkeley.
11. H. T. Kung, “Synchronized and asynchronous parallel algorithms for multiprocessors”, Computer Science Department, Carnegie Mellon University, 1976.

Aligning systems engineering and project management standards to improve the management of processes

Rui XUE, Claude BARON, Philippe ESTEBAN, and Abd-El-Kader SAHRAOUI

1 Introduction

Systems engineering and project management, whatever the terminology used [1], are both critical to achieve the success of projects. Indeed, with the growing scale of projects, the roles of the project manager and systems engineers are more critical than ever. However, for many years, a cultural barrier has been growing between practitioners of systems engineering and of program management. While project management has overall project accountability and systems engineering has accountability for the technical and systems elements of the project, some systems engineers and project managers have developed the mindset that their work activities are separate from each other rather than part of an organic whole. Consequently, work often costs more, takes longer, and provides a suboptimal solution for the customer or end user.

How to bridge the gap between project management and systems engineering recently became of increasing importance in industry. The INCOSE (International Council on Systems Engineering) and the PMI (Project Management Institute) have recognized the importance of integrating systems engineering with project management; together they took out a survey to analyze this question. They identified four methods that can improve the integration of both domains. Exploring the first one, the goal of this paper is to compare the five important systems engineering standards (ANSI/EIA 632 [2], ISO/IEC 15288 [3], IEEE 1220 [4], INCOSE HANDBOOK [5] and SEBoK [6]) and the two significant project management standards (PMBoK [7] and ISO 21500 [8]) in order to evaluate the coherency of standards with regard to the processes they describe and that are involved through the whole project in order to facilitate the management of the projects and improve their chances of success. Our final goal is to identify among them a pair of 'most compatible' standards from the point of view of the management of system engineering processes and to align them.

In section 2, we explain how important integrating systems engineering with project management is and give the background and the research status over this question. Section 3 introduces and compares five systems engineering and two project management standards. Based on the comparison of standards, section 4 gives the conclusion on comparing the most important standards of both domains.

2 Current Situation

Schlager [9] was the first to promote systems engineering in 1950s as a systematic approach for engineering complex industrial systems. On the other hand, it was also in the 1950s that organizations started to systematically apply project management tools and techniques to complex engineering projects [10]. However, aligning systems engineering with project management has only been paid attention in the beginning of 21th Century. The point is that the two disciplines can be disjoint, partially intersecting, or one can be seen as a subset of the other. Sharon put forward that systems engineering involves product domain and project management involves project domain [1]. Howard Eisner in [11] not only defines and describes the essentials of project and systems engineering management but shows the critical relationship and interconnection between project

R. XUE (✉) • C. BARON
CNRS, LAAS, 7 av. du col. Roche, F-31400 Toulouse, France

Univ. de Toulouse, INSA, LAAS, F-31400 Toulouse, France
e-mail: rxue@laas.fr; claude.baron@laas.fr

P. ESTEBAN
CNRS, LAAS, 7 av. du col. Roche, F-31400 Toulouse, France

Univ de Toulouse, UPS, LAAS, F-31400 Toulouse, France
e-mail: philippe.esteban@laas.fr

A.-K. SAHRAOUI
CNRS, LAAS, 7 av. du col. Roche, F-31400 Toulouse, France

Univ de Toulouse, UTM, LAAS, F-31100 Toulouse, France
e-mail: sahraoui@laas.fr

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_78,

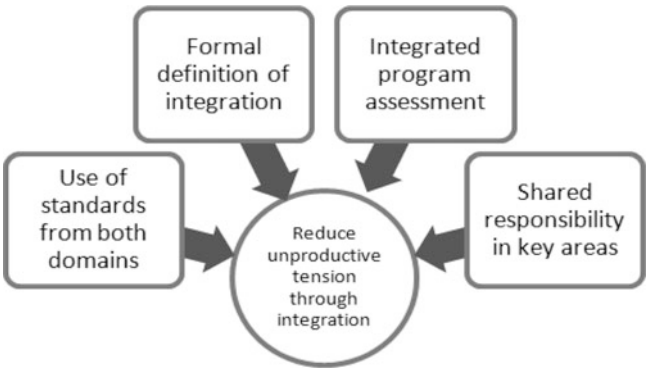


Fig. 1. The result of the survey [12]

management and systems engineering, to enable both engineers and project managers to understand their roles and to collaborate.

One way to address such challenge could be using the systems of systems paradigm [13]. At the normative level, in order to help organizations overcome the resultant inefficiencies of non-collaborative practices, the Project Management Institute (PMI) and the International Council on Systems Engineering (INCOSE) both recognized the need to improve integration of practices between their professional communities [14]. In October 2012, they conducted a survey to better understand how PM and SE are lead and could be integrated within the organizations to improve the probability of projects success [15]. As shown in Fig 1, they proposed four methods to reduce the unproductive tension between the systems engineers and project managers. In 2012 too, the INCOSE Chesapeake chapter formed a SE/PM Working Group in order to enhance program success through the improved integration of shared practices [16].

The goal of this paper is to align SE with PM exploring the first option of Fig. 1: using standards from both domains and try to align them.

3 Analysis and Comparison of Standards

Since 1969, many system SE and PM standards have been elaborated in different application fields, such as military, aeronautics, automatic and management. The ANSI/EIA-632, ISO/IEC-15288 and IEEE-1220, INCOSE HANDBOOK, SEBoK, PMBoK and ISO 21500 play the most important roles. Section 3.1 gives an overview of SE standards, analyzes and compares them considering their ability to manage SE; Section 3.2 introduces and compares PM standards.

3.1 Comparison of the Systems Engineering Standards

Many SE standards have been drawn up in different fields of application, such as military, aeronautics, automatic and management [17–19] since 1969. The first standard widely used is ANSI/EIA 632; after this standard, many organizations developed their own standard. For example, the ISO/IEC 15288 was published by the International Organization for Standardization (ISO), the IEEE [19]. In [19] we introduce the five main SE standards. As some overlap exists between SE and PM, some processes relative to project management domain (Project Management Processes, PMP) can be identified in SE standards. Thus this paper highlights these PMP in each standard.

ANSI/EIA 632. This standard provides a systematic approach to engineering and reengineering a system. It defines 13 processes, 34 requirements (in fact like the subprocess) at total. One of the most useful features of this standard is the close connection between the processes; they are coordinated throughout the project. It defines the processes at an intermediate level and includes the whole systems life cycle, but focuses on the conception, development, production, utilization and support, just referring a little to the retirement. It also offers some PMP, such as the planning process, assessment process and controlling process.

As shown in the Fig 2, the ANSI/EIA 632 organizes its processes into 5 groups. The Technical Management process group is related to the project management aspect and includes the PMP. Although it defines some processes about the project management, but these processes just involve a little to PM, so they are not enough for the project managers to manage the project. This standard although provides the processes for engineering systems, but it does not define the tool or method for how to implement the processes through the whole project.

IEEE 1220. This standard defines the requirements for an enterprise's total technical effort related to development of products and processes that will provide life cycle support for products. It focuses on the technical processes and provides an approach fit for products development in a system context. Compared with the ANSI/EIA 632 and ISO/IEC 15288, it intends to define the SE processes at a high level of detail. It also includes the whole systems life cycle, but focuses on the conception, development and production. It just offers one process, the Quality Management requirement that belongs to PMP. As shown in the Fig. 4, this standard is not organized into process groups. The structure of this standard is not similar to the others. It defines 6 stages for systems engineering, but after we

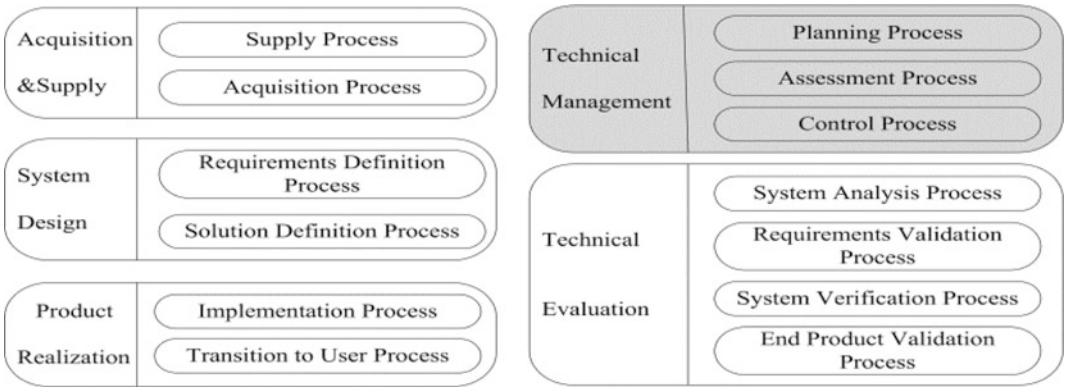


Fig. 2. Hierarchical organization of the ANSI/EIA 632 standard

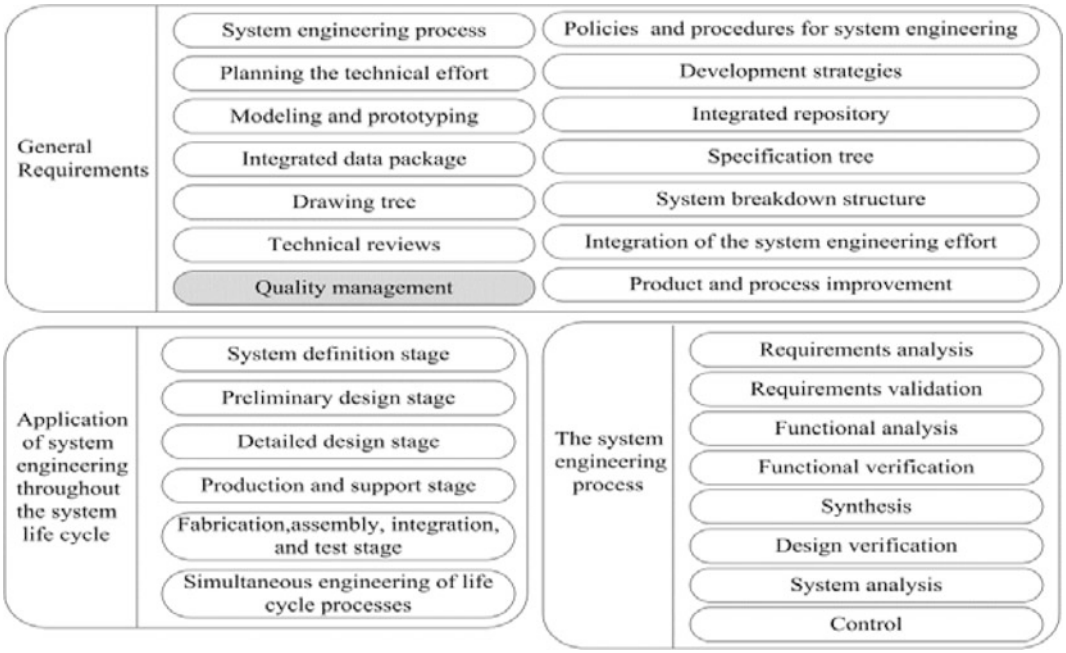


Fig. 3. Hierarchical organization of the IEEE 1220 standard

analyze the definition of those stages, in fact, it defines the same systems life cycle as the ISO/IEC 15288. Its most useful feature is that it can be used in complement to ISO/IEC 15288. This standard just defines one process related to project management processes, the quality management process, because this standard focuses on the development stage of the systems life cycle.

ISO/IEC 15288. This standard provides a comprehensive set of life cycle processes. It defines 26 processes at the process level and includes the whole system life cycle, from conception to the retirement of the systems and the processes cover all the stages. Like the ANSI/EIA 632, it also offers some PMP, such as the configuration management, information management and controlling process. As shown in Fig. 3, the ISO/IEC 15288 defines 4 process groups: agreement processes, enterprise processes, project processes and technical processes. The Enterprise Processes and Project Processes groups offer a partial overlap with project management processes. This standard defines more PMP than ANSI/EIA 632. But it has the same problem that is it does not provide nor specify systems engineering methods or procedures to address detailed process requirements for the application of this standard.

INCOSE SE HANDBOOK. This standard aligns with the ISO/IEC 15288 standard; it defines the same processes as the ISO/IEC 15288, so its organizational structure is the same as the ISO/IEC 15288. It also considers both the technical and management processes. It has been the most detailed standard until SEBoK was published in 2013. The most interesting feature of this standard is that it not only defines the

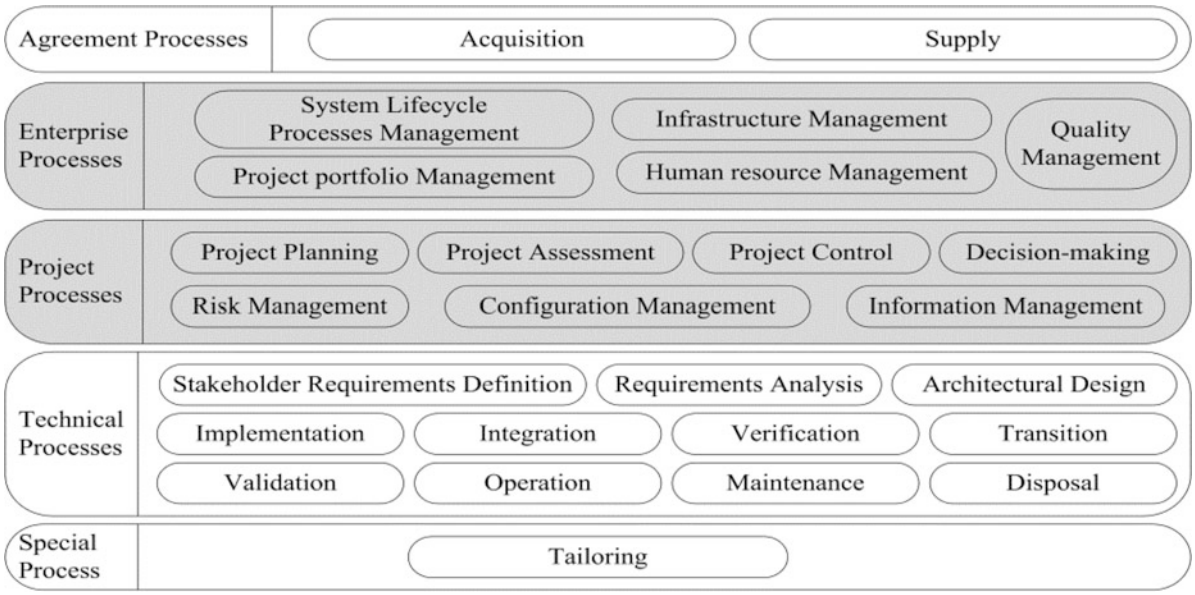


Fig. 4. Hierarchical organization of the ISO/IEC 15288 standard

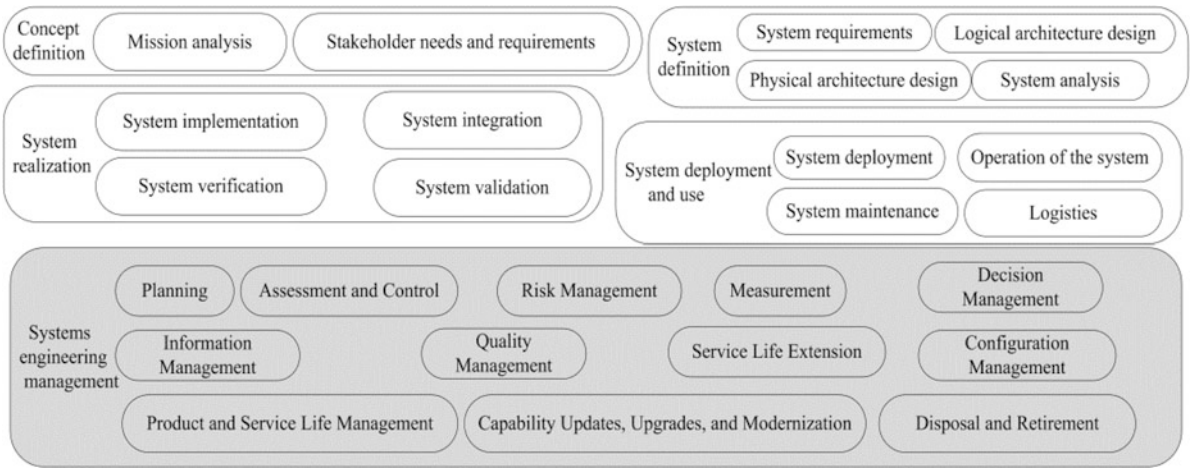


Fig. 5. Hierarchy of the SEBoK standard

systems engineering processes, say "what" to do, but also provides some methods and tools about "how" to do during the whole project.

SEBoK. The SEBoK (Guide to the Systems Engineering Body of Knowledge) refers to all the other systems engineering standards, such as the ANSI/EIA-632, ISO/IEC-15288 and IEEE-1220, but it refers more to the ISO/IEC-15288. This standard is the most detailed standard so far.

As the Fig. 5 shows, there are five groups of processes in this standard, the fifth being Systems Engineering Management. So there are some identical processes, such as the systems implementation process, the system integration process and the system validation process. In fact, all the technical processes of the ISO/IEC-15288 are included in the SEBoK. The SEBoK just changes some process names, divides and integrates some processes. It uses all the processes from the enterprise processes and project processes of the ISO/IEC-15288, for the aspect of technical processes. It breaks the technical processes of ISO/IEC-15288 and assembles them into 14 systems engineering processes.

This standard points out the two domains of systems engineering: the systems engineering domain (as technical processes that we introduced before) and the systems engineering management domain (in this paper we call it project management processes) clearly. It describes the processes for both domains at the same level of details. It recognizes the importance of the implementation of management

Table 1. Profiles of the five SE standards

	ANSI/EIA-632	IEEE-1220	ISO/IEC- 15288	INCOSE HANDBOOK	SEBoK
Content of standard	13 processes 34 requirements	8 processes	25 processes	25 processes	26 processes
Focus of systems life cycle	Conception and development	all the systems life cycle	all the systems life cycle	all the systems life cycle	all the systems life cycle
Number of pages	110	70	70	400	850
Level of details	◆◆◇◇◇	◆◆◇◇◇	◆◆◇◇◇	◆◆◆◆◇	◆◆◆◆◆
Context of application	Program and project environment	Program and project environment	Enterprise environment	Enterprise environment	External environment
Year of publication	1998	2005	2008	2010	2013
Reversion frequency	◆◆◇◇◇	◆◆◇◇◇	◆◆◆◆◆	◆◆◆◇◇	◆◇◇◇◇
Number of SEMP	3	1	12	12	12
SEMP's proportion	3/13	1/14	12/25	12/25	12/26

processes during the systems engineering, so it increases the proportion of the management processes. It is the first SE standard that puts forward the importance of systems engineering management explicitly.

Conclusion on the Systems Engineering Standards. After having analyzed of each SE standard, we compare them according to 9 criteria (in particular considering the System Engineering Management Processes (SEMP)) and defined for each a profile.

In today's environment, there is an ever-increasing need to develop and produce systems that are robust, reliable, high quality, supportable, cost-effective, and responsive to the needs of the customer or user. With the systems increasing complexity, Systems Engineering standards become more and more detailed. The newest standard SEBoK has more than 800 pages and includes the most a lot of disciplines about systems engineering. The more recent SE standards are mainly based on the ISO/IEC 15288. Moreover, as the industrial organizations have paid more and more attention on the project management processes, we can notice that almost half of the processes in the ISO/IEC 15288 are related to project management.

3.2 Comparison of the Project Management Standards

The PM standards have been developed later than the SE standards but more and more attention has been paid lately on project management. Two important PM standards have been drawn up. The first international PM standard is the PMBoK (A Guide to the project management body of knowledge); it was firstly published in 1996, it has been updated five times, and the last version is the fifth version that was published in 2013. The second famous standard is the ISO 21500 that was published in 2012 for the first time by the ISO.

Table 2. Structure of the PMBoK

Five process groups	Ten knowledge areas
• Initiating, • Planning, • Executing, • Monitoring and controlling, • Closing	• Project integration management, • Project scope management, • Project time management, • Project cost management, • Project quality management, • Project human resource management, • Project communications management, • Project risk management, • Project procurement management, • Project stakeholder management.

PMBoK. Published by the Project Management Institute, the purpose of this standard is to provide the knowledge, processes, skills, tools and techniques that have a significant impact on the project success [7]. It defines 5 process groups and 10 knowledge areas. It not only defines the processes for PM, but also provides the tools and methods for how to implement the processes. The processes cover the whole systems life cycle, from conception to the retirement of the systems. This standard is the most famous project management guide for the project managers.

ISO 21500. The ISO 21500 refers to many standards, such as the PMBoK and the IPMA (International Project Management Association).This standard defines five process groups and ten subjects (the same as the knowledge area in PMBoK). There is less difference between the two PM standards, just the level of detail; the PMBOK is more detailed than ISO 21500.

According the table 2 and table 3, we can find that there are very little difference between the two standards, just some names of the process groups and the knowledge area. In the ISO 21500, it uses the implementing instead of executing in PMBoK, the controlling instead of the monitoring and controlling and the project resource management instead of the project human resource management.

Table 3. Structure of the ISO 21500

Five process groups	Ten subjects
• Initiating, • Planning, • Implementing • Controlling, Closing	• Project integration management, • Project scope management, • Project time management, • Project cost management, • Project quality management, • Project resource management, • Project communications management, • Project risk management, • Project procurement management, • Project stakeholder management.

Table 4. Conclusion on the two project management standards

	PMBoK 5th	ISO 21500
Content of standard	5 process groups, 10 knowledge areas	5 process groups, 10 subjects
Number of pages	616	47
Level of details	◆◆◆◆◇	◆◇◇◇◇
Year of publication	2013	2012
Revision frequency	◆◆◆◆◆	◇◇◇◇◇

Conclusion on the Project Management Standards. Table 4 compares all the SE PM standards according to 5 criteria. After comparing the two PM standards, we get the conclusion that the contents of the standards are almost the same. But the ISO 21500 is only 47 pages and is limited to the introduction of the process, their inputs and their outputs. The ISO 21500 processes are more likely to be oriented towards a cascade approach of the scope definition rather than an interactive approach. So the ISO 21500 is maybe less attractive for agile organization. The first edition of PMBoK was published very early and it has been updated for four times, so it is more famous and popular than ISO 21500.

3.3 Synthesis

All SE standards describe good systems engineering practices. They may suggest a life cycle to provide a context for their recommendation. In short, the EIA-632 standard is more suitable for engineering enterprise-based systems; it focuses more on the technical management, validation and verification aspects. The IEC-15288:2008 standard is more suitable for engineering complex systems, especially projects that cover an entire system life cycle; there are some other SE standards that are based on it. The IEEE-1220 standard is more suitable for smaller systems and focuses on the development stage rather than the system life cycle or the technical management aspects. The INCOSE handbook has the same processes as ISO 15288, but it provides some tools and methods for systems engineering. The SEBoK is the most detailed SE standards, but it also refers more to ISO 15288. So for the systems engineering, the ISO15288 is the most suitable SE standard for meeting our requirement.

After comparing the two PM standards, we get to the conclusion that the content of the standards are almost the same. But the ISO 21500 is only 47 pages and is limited to the introduction of the process, their inputs and their outputs. The ISO 21500 processes are more likely to be oriented towards a cascade approach of the scope definition rather than an interactive approach. So the ISO 21500 is maybe less attractive than PMBoK for agile organization. Moreover, we can notice that the first edition of PMBoK was published very early and that since it has been updated for four times; it is more famous and more popular than the ISO 21500.

4 Conclusion

To develop the system quickly and efficiently, it is necessary to correctly implement SE processes and PM processes. One first goal of this paper was to provide enterprises with objective elements to choose two reference standards for their projects from the SE standard and the PM standard. This goal is reached after our analysis of standards from both domains. However, the originality of the paper relies in that it presented an analysis of five SE standards and the two PM standards, considering not only the SE processes, but also the PM ones, in order to be able to align practices and make them more coherent. This paper thus compared the standards from both domains in order to evaluate the compatibility of standards with regard to how they manage Systems Engineering processes. Considering the analysis, the choice of the most compatible ones to guide Systems Engineering on one side and Project Management on the other seems to be the ISO /IEC 15288 and the PMBoK. Choosing these standards should thus improve the probability of success of projects. However, aligning standards remains one solution to coordinate practices and processes; according to [12], other means must be associated to this alignment of processes to improve the success of projects, such as sharing responsibilities and decisions for example.

References

1. Sharon, A., de Weck, O. L., Dori, D.: Project management vs. systems engineering management: A practitioners' view on integrating the project and product domains. Syst. Eng. 14, 427–440(2011)
2. ANSI/EIA: Standard for Processes for Engineering a System. ANSI/EIA (1998)
3. ISO/IEC/IEEE Systems and Software Eng. - System Life Cycle Processes. IEEE (2008)
4. IEEE Standard for Application and Management of the Systems Eng. Process. IEEE 1999)
5. INCOSE Systems engineering handbook V3.2: A "How to" Guide for all engineers. INCOSE (2010)

6. Pyster, A., Olwell, D. and el. :Guide to the Systems Engineering Body of Knowledge (SE-BoK). BKCASE (2013)
7. PMI: A Guide to the Project Management Body of Knowledge: PMBOK V5. PMI (2013)
8. ISO 21500 Guidance on Project Management A Pocket Guide. Van Haren Pub (2013)
9. Schlager, K. J.: Systenas Engineering-Key to Modern Development. IRE TRANSACTIONS ON ENGINEERING MANAGEMENT (1956)
10. Kwak, Y.H., Carayannis, E.G.: A brief History of Project Management-The story of managing projects. Greenwood Publishing Group (2005)
11. Eisner, H.: Essentials of project and systems eng. management. John Wiley & Sons (2008)
12. Conforto, E., Rossi, M., Rebentisch, E., Oehmen, J., Pacenza, M.: Survey Report: Im-proving Integration of Program Management and Systems Engineering. MIT Consor-tium for Engineering Program Excellence (2013)
13. Sahraoui, A.E.K., Buede, D.M., Sage, A.P.: systems engineering research. Journal systems science and systems engineering, Springer, Vol17, No3, pp 319-333 (2008)
14. PMI and Incose Align To Help Organizations Improve Program Success | Project Man-agement Institute, http://www.pmi.org/About-Us/Press-Releases/PMI-and-Incose-Align-To-Help-Organizations-Improve-Program-Success.aspx
15. Oehmen, J., Oppenheim, B. W., Secor, D., Norman, E., Rebentisch, Sopko, E., J. A., Steu-ber, M., Dove, R., Moghaddam, K., McNeal, S.: The Guide to Lean Enablers for Man-aging Engineering Programs. Joint MIT-PMI-INCOSE Community of Practice on Lean in Program Management (2012).
16. Systems Engineering – Project Management (SE – PM) Working Group — INCOSE Chesapeake Chapter Chesapeake Chapter of INCOSE.
17. An Overview of the Systems Engineering Knowledge Domain - Recherche Google
18. Sahraoui, A.: Processes for Engineering a System an Introduction to Processes, Methods and Tools. In Information and Communication Technologies, pp. 2760–2765. (2006)
19. Xue, R., Baron, C., Esteban, P.: Managing systems engineering processes: a multi-standard approach. In 8th annual IEEE international systems conference, in press.

Effect of the groove dimensions and orientation on the static and dynamic performance of non recessed hybrid journal bearing

Vijay Kumar Dwivedi, Satish chand, and K. N. Pandey

1 Introduction

Non recessed hybrid journal bearing with grooves are often used when applied load and shaft speed are high. Lubricant is usually supplied in the journal bearing through a hole or groove in the low pressure region of the lubricant film. Three types of grooves namely; axial, circumferential and spiral feeding grooves are commonly used for lubricant feeding in the bearing. The lubricant is often supplied at a prescribed pressure through one or two axial groove. Twin groove journal bearings are widely used, especially when the journal is expected to assume both direction of rotation.

Morton et al. [1] presented the influence of grooves in bearing on the stability and response of rotating systems. They concluded that grooved bearing modify the journal locus by increasing the attitude angle and it also change the cavitation boundary at low eccentricity ratios i.e. high speeds. Basri and Neal [2] were the first to include a separate measurement of the flow rate at each groove of twin groove journal bearing. Ma and Taylor [3, 4] presented experimental results for twin bearing but omitted the main lubricant properties. They have used unusually large groves bearing. Keogh et al. [4] performed CFD based work on twin groove journal bearing an included flow and energy calculations within the groove region. Chun and Ha [5] analyzed the effect of mixing at the grooves for the turbulent flow occurring in high speed journal bearings and found that groove mixing should be always included in the analysis. Costa et al. [6] found that single groove bearing configuration perform better with a groove located along the load line. They also commented that twin groove journal bearing are widely used especially when the shaft is expected to assume both direction of rotation. Jeddi et al. [7] used the FEM to study the effect of feeding pressure and the runner velocity on the thermo hydrodynamic behavior of the lubricant groove of the journal bearing. They found these two factors to change significantly the flow and thermal pattern within the groove region. Brito et al. [8] presented experimental comparison of the performance of a journal bearing with a single and twin axial groove bearing. They were found that under heavy loaded operation the twin groove configuration might actually deteriorate the bearing performance when compared with the single groove arrangement due to uneven lubricants feed through each groove.

In this paper supply condition such as the groove geometry, location of groove and dimensions of groove for twin groove journal bearing have been presented.

2 NOMENCLATURE

ø = attitude angle
ε = (e/c) eccentricity ratio
h = minimum film thickness
ζ = angular location of groove
w = groove width
μ = viscosity
e = eccentricity
c = radial clearance
a = groove length
L = bearing length
D = diameter of bearing
p = pressure

3 Theoretical Model/ Analysis

This section describes the theoretical model developed for the analysis of the performance of two groove journal bearing. Figure 1 shows the basic bearing geometry.

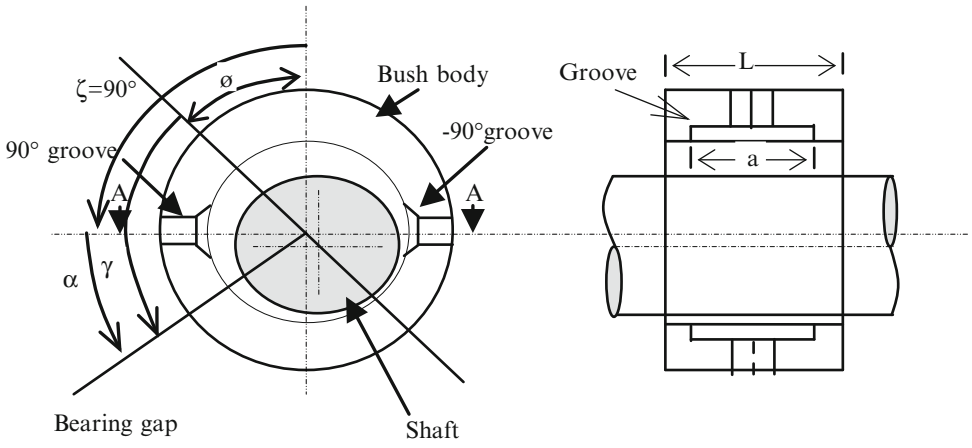
V.K. Dwivedi (✉)
GLA University, Mathura, Mathura, U.P., India

S. chand
Vishveshwarya Group of Institutions, G.B. Nagar, U.P., India

K.N. Pandey
MNNIT, Allahabad, Allahabad, U.P., India

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_79,

Fig. 1 Bearing geometry



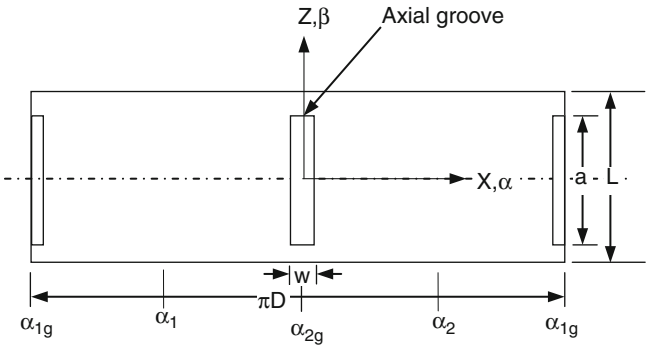


Fig. 2 Fluid domain (unwrapped bearing geometry) including coordinate system

The operational performance of a journal bearing can be divided into static and dynamic aspects. Static performance includes a bearing's load capacity (fluid film reaction) and dynamic performance includes stiffness and damping coefficients of the fluid film. A bearing load carrying capacity is often measured using eccentricity ratio. Figure 2 and Fig 3 show the unwrapped bearing geometry in two dimensions and three dimensions, respectively, considering fluid film thickness very small. $\alpha_1$and α_2 in these figure show the extent of positive fluid film pressure. α_{1g} and α_{2g} in represent the location of centre line of the grooves.

Flow field equations: The non-dimensional Reynolds equation which governs the flow of lubricating oil in the clearance space of a hydrodynamic journal bearing using linearized turbulence theory of Constatantinescu (1967) is given by Eq. (3.5)

$$\frac{\partial}{\partial\alpha}\left[\frac{\overline{h}^3}{\overline{\mu}\frac{\partial\overline{p}}{\partial\alpha}}\right]+\frac{\partial}{\partial\beta}\left[\frac{\overline{h}^3}{\overline{\mu}\frac{\partial\overline{p}}{\partial\beta}}\right]=\frac{1}{2}\overline{\Omega}\frac{\partial\overline{h}}{\partial\alpha}+\frac{\partial\overline{h}}{\partial\overline{t}} \tag{1}$$

The non dimensional fluid-film thickness ($\overline{h}$) for parallel axes case is given by

$$\overline{h}=1.0-\overline{X_j}cos\alpha-\overline{Z_j}sin\alpha \tag{2}$$

Boundary condition: The boundary conditions pertinent to the problem are given as

$$(i)\frac{\partial\overline{p}}{\partial\beta}=0at\beta=0 \qquad (ii)\overline{p}=0at\beta=\pm\frac{L}{D}=\pm\lambda$$

$$(iii\overline{p}=0at\left(\left(-\frac{a}{D}\langle\beta\langle\frac{a}{D}\right)and\left(\left(\alpha_{1g}-\frac{w}{D}\right)\langle\alpha\langle\left(\alpha_{1g}+\frac{w}{D}\right)\right)\right)$$

$$(iv)\overline{p}=0at\left(\left(-\frac{a}{D}\langle\beta\langle\frac{a}{D}\right)and\left(\left(\alpha_{2g}-\frac{w}{D}\right)\langle\alpha\langle\left(\alpha_{2g}+\frac{w}{D}\right)\right)\right) \tag{3}$$

Result and discussion The analysis and solution algorithm were used to compute the pressure profile, fluid film reaction or load capacity, fluid film stiffness and damping coefficients. These studies are conducted by taking bearing aspect ratio (L/D) 0.5 and 0.25. Assuming bearing and journal axes parallel and ratio of nominal clearance to the journal radius 0.001 (C/R = 0.001).

Validation of results To establish the validity of the analysis, solution algorithms and the computer program, the dimensionless load capacity for different groove locations obtained from the present short bearing approximation were compared with results available in literature [10]. In Fig 4, the comparison of dimensionless load capacity with groove location is shown for laminar flow condition of fluid film at L/D ratio of 0.50, C/R of 0.004 and eccentricity of 0.7 between present short bearing analysis and results available in literature [10]. Results compare well with some minor inevitable variation due to different solution methodology. After validation of solution procedure, now results of the study of different groove dimensions, location and superlaminar flow conditions are presented in subsequent paragraph.

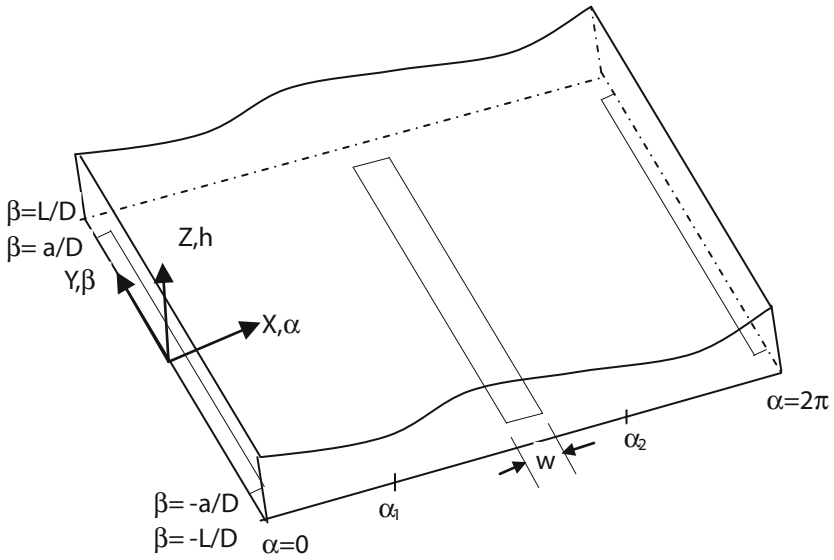


Fig. 3 Development of fluid film between bearing and journal surface (3-D) view

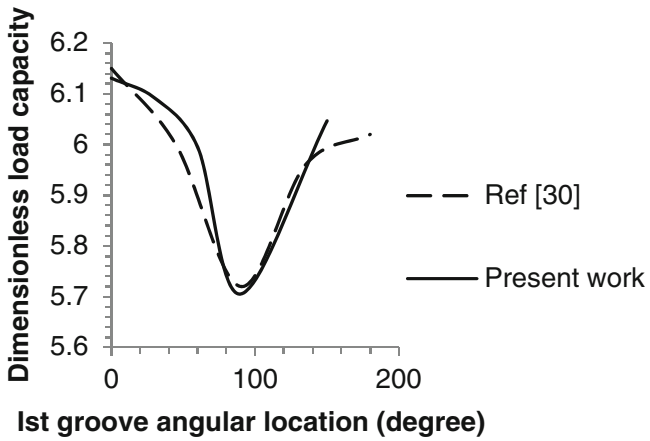


Fig. 4 Eccentricity vs Sommerfeld Number for L/D = 0.50 for variation of Reynolds number

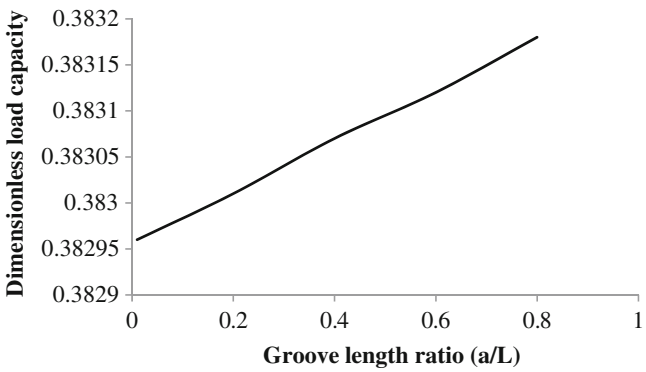


Fig. 5 Influence of groove length ratio on load carrying capacity

Influence of the groove length ratio (a/L):

In most practical cases the ideal axial groove length would coincide with bearing length, (a/L=1). This would, in principle, provide the most uniform delivery of lubricant to the whole axial extension of the bearing. However, this is not always possible, or at least not practical to do due to design constraints. Larger grooves normally require bigger pumping systems as they tend to increase oil flow rate. Larger grooves also tend to reduce the structural stiffness of the system, which might be an important parameter in many applications. Groove length ratio affects the dynamic characteristics of the bearing.

Influence of the groove length ratio (a/L) on dimensionless load capacity The variation of dimensionless load capacity with groove length ratio is shown in Fig 4.11 for a fixed value of w/D = 0.2. The dimensionless load capacity increases with increase in groove length ratio.

Influence of the groove length ratio (a/L) on stiffness coefficient The variation fluid film stiffness coefficients $\overline{K}_{XX}, \overline{K}_{XZ}, \overline{K}_{ZX}, \overline{K}_{ZZ}$ with groove length ratio are shown in Figs 6 - 9. It is observed from Fig 6 and Fig 7 that $\overline{K}_{XX}$ and $\overline{K}_{XZ}$ decreases linearly with increase in groove length ratio. The magnitude of cross-coupled damping coefficient $\overline{K}_{ZX}$is negative over the entire range of groove length ratio.

It is observed from Fig 8 that $\overline{K}_{ZX}$ first decreases with increase in groove length ratio then stabilized at a certain value. The lowest value of $\overline{K}_{ZX}$ is equal to – 4.13. Direct stiffness in Z direction, $\overline{K}_{ZZ}$ shows trend similar to the $\overline{K}_{ZX}$ i. e. first it decreases before stabilizing to a fixed value. The minimum value of $\overline{K}_{ZZ}$ is equal to 1.8.

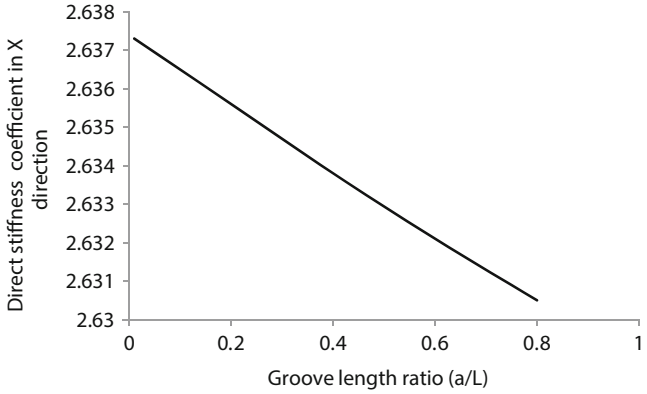


Fig. 6 Influence of groove length ratio on direct stiffness coefficient

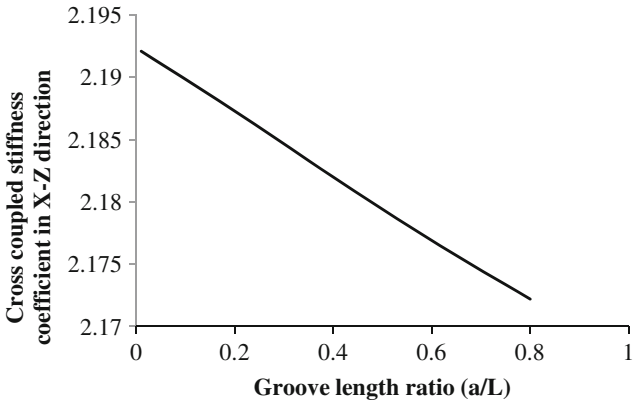


Fig. 7 Influence of groove length ratio on cross coupled stiffness coefficient

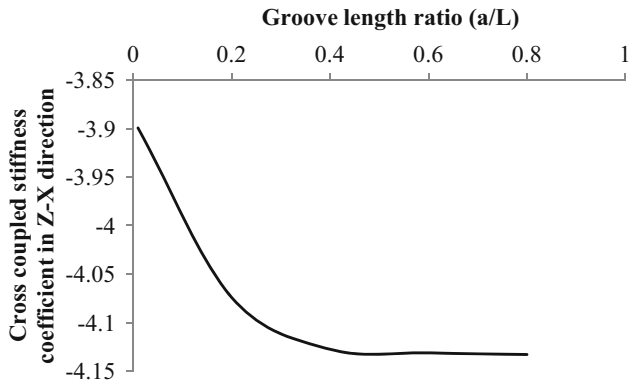


Fig. 8 Influence of groove length ratio on cross coupled stiffness coefficient

Influence of the groove width ratio (w/D) The influence of the circumferential extension of groove, also called the groove width (w) upon bearing performance is very significant. The effect of w/D ratio on pressure profile, stiffness coefficient and damping coefficient has been discussed in subsequent paragraphs.

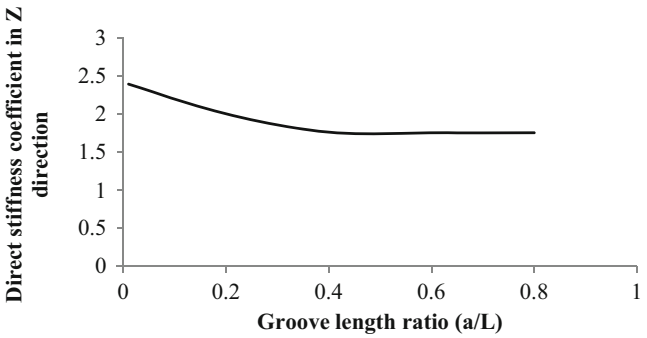


Fig. 9 Influence of groove length ratio on direct stiffness coefficient

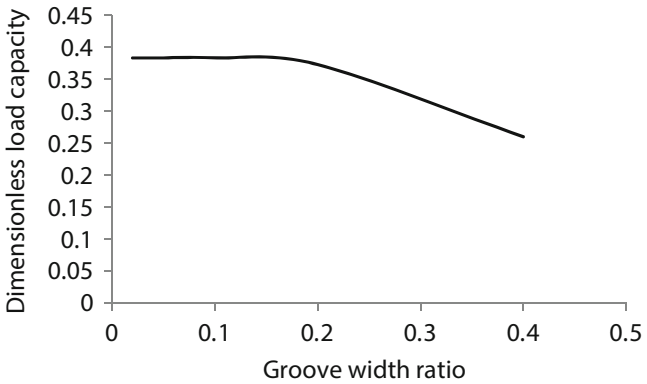


Fig. 10 Influence of groove width ratio on load carrying capacity

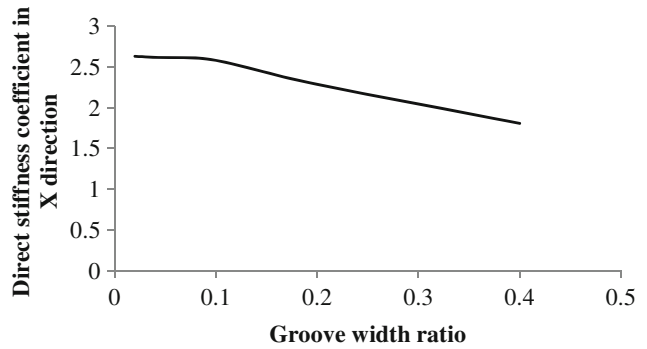


Fig. 11 Influence of groove width ratio on direct stiffness coefficient

Influence of the groove width ratio (w/D) on dimensionless load capacity Variation of dimensionless load capacity of the bearing with groove width ratio is given in Fig 10. It is observed that the dimensionless load capacity is initially constant up to groove width ratio w/D = 0.15 then decreases with increase in groove width ratio.

Influence of the groove width ratio (w/D) on stiffness coefficient The fluid film stiffness coefficients $\overline{K}_{XX}, \overline{K}_{XZ}$, $\overline{K}_{ZX}, \overline{K}_{ZZ}$ are shown in Figs 11 to 14 with variation of groove

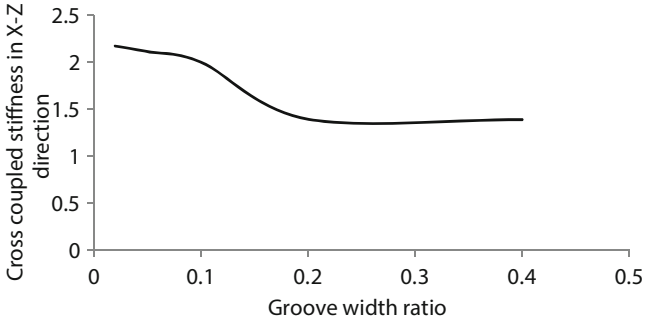


Fig. 12 Influence of groove width ratio on cross coupled stiffness coefficient

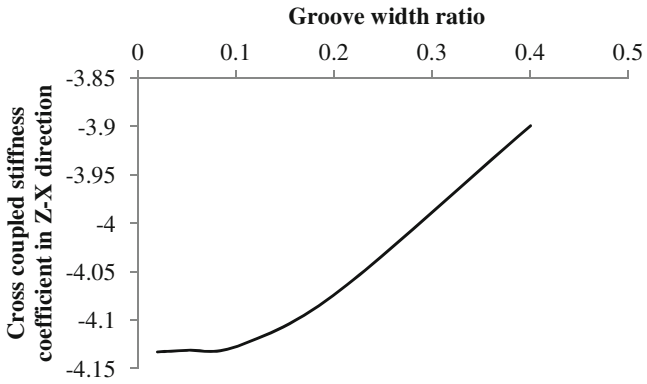


Fig. 13 Influence of groove width ratio on cross coupled stiffness coefficient

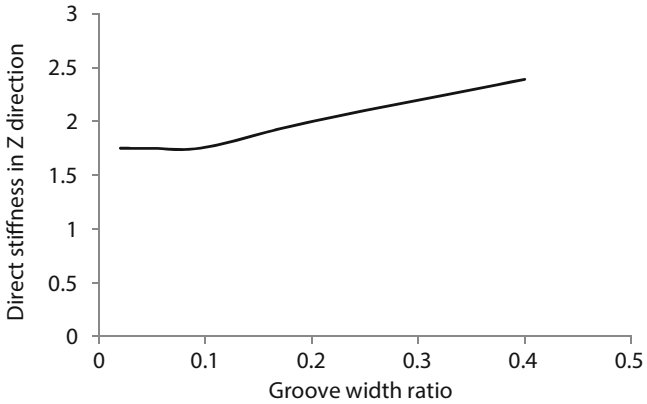


Fig. 14 Influence of groove width ratio on direct stiffness coefficient

width ratio. It is observed from Fig 11 that direct stiffness coefficient $\overline{K}_{XX}$ is initially constant up to groove width ratio w/D = 0.10 then decreases with increase in groove width. Figure 12 presents that cross coupled stiffness coefficient $\overline{K}_{ZX}$ decreases with increase in groove width ratio before settling to a fixed value. The minimum value of $\overline{K}_{XZ}$ is equal to 1.43. The magnitude of the cross coupled stiffness coefficient $\overline{K}_{ZX}$ is negative over the entire range of groove

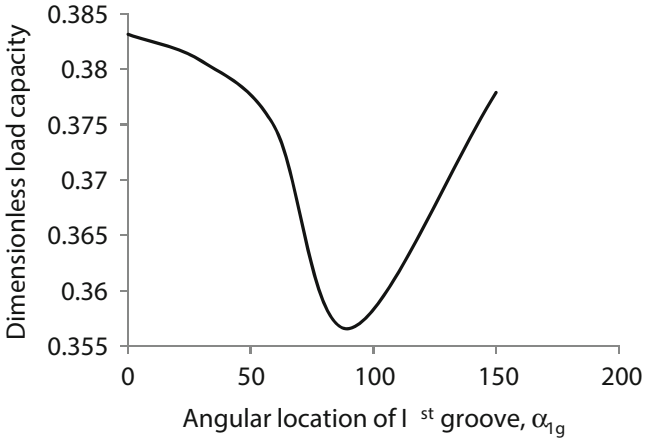


Fig. 15 Influence of angular location of groove on Fluid film reaction

width ratio. It is observed from Fig 13 that $\overline{K}_{ZX}$ increases with increase in groove width ratio. Figure 14 shows that direct stiffness coefficient in Z direction $\overline{K}_{ZZ}$ is initially constant up to groove width ratio equal to 0.10 then increases with increase in groove width ratio.

Influence of the location of the grooves with respect to the load line The angle between the axis of the grooves and the load line, designated as the groove angle (ζ) is measured in the direction of rotation of shaft. This angle is normally 90° in twin groove journal bearings. The highest values are obtained with the most extreme values of ζ (30° and 150º). The angle between the grooves of twin groove bearing is 180^0.

Influence of location of the grooves on dimensionless load capacity Groove location is an important parameter considered by designers because it affects static as well as dynamic characteristics of a journal bearing. The variation of dimensionless load capacity of the bearing with angular location of Ist groove, α_{1g} is presented in Fig 15. The dimensionless load capacity initially decreases with increase in angles of groove and then it starts increasing from $\alpha_{1g}= 90^0$. The fluid film reaction is almost same for $\alpha_{1g}= 30^0$ and 150^0. The minimum value of dimensionless load capacity is 0.3567 at $\alpha_{1g}= 90^0$.

Influence of location of the grooves on stiffness coefficient The variation of fluid film stiffness coefficients $\overline{K}_{XX}, \overline{K}_{XZ}, \overline{K}_{ZX}, \overline{K}_{ZZ}$ with location of Ist groove are shown in Figs 16-19. The location of IInd groove is 180^0 ahead to Ist groove. It is observed from Fig 16 that the variation of direct stiffness coefficient $\overline{K}_{XX}$ increases with increase in location of first groove upto α_{1g} equal to 90^0. An optimum value of Ist groove location $\alpha_{1g}= 90^0$ is observed at which stiffness coefficient is maximum, i.e. $\overline{K}_{XX} = 2.83$.

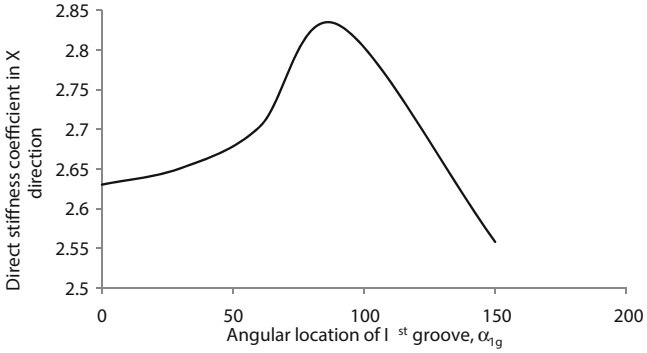


Fig. 16 Influence of angular location of groove on direct stiffness coefficient

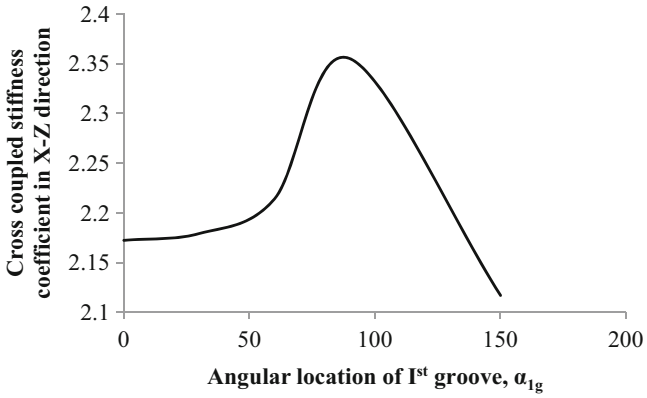


Fig. 17 Influence of angular location of groove on cross coupled stiffness coefficient

Fig 17 shows the cross coupled stiffness coefficient $\overline{K}_{XZ}$. The cross coupled stiffness coefficient $\overline{K}_{XZ}$ shows the similar behavior as observed in the case of direct stiffness coefficient $\overline{K}_{XX}$ with Sommerfeld number. The optimum value of cross coupled stiffness $\overline{K}_{XZ}$ is 2.353 at Ist groove location 90^0. The magnitude of cross-coupled stiffness coefficient $\overline{K}_{ZX}$ is negative for the entire range of groove location, Fig 18. The optimum value of $\overline{K}_{ZX}$ is equal to 4.2931 for α_{1g} equal to 90^0. Fig 19 shows the variation of direct stiffness coefficient in Z direction, $\overline{K}_{ZZ}$ against angular position of Ist groove of twin axial groove bearing. It is observed that $\overline{K}_{ZZ}$ decreases up to $\alpha_{1g} = 90^0$ then starts $\overline{K}_{ZZ}$ increasing up to $\alpha_{1g} = 150^0$. The lower optimum value of $\overline{K}_{ZZ}$ is equal to 1.7107 for $\alpha_{1g} = 90^0$.

4 Conclusion

A parametric study of lubricant supply conditions on the performance of non recessed hybrid journal bearings has been carried out using the analytical model. On the basis of proposed analytical model following conclusions were drawn:

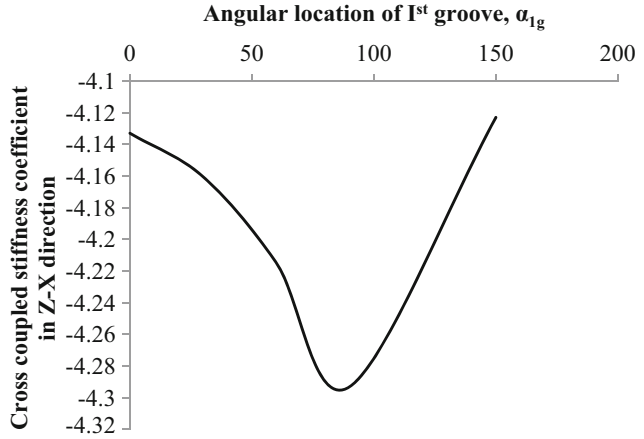


Fig. 18 Influence of angular location of groove on cross coupled stiffness coefficient

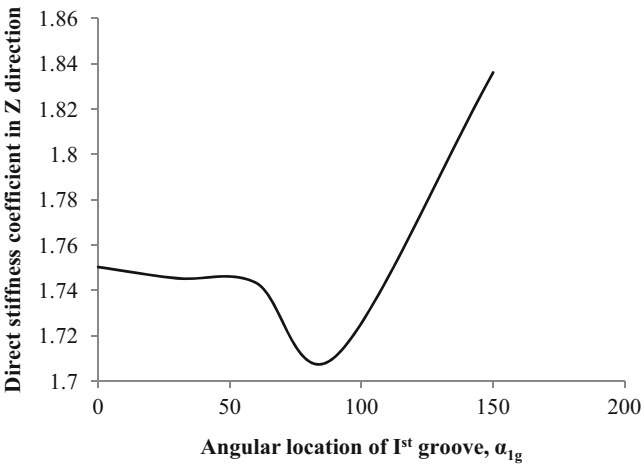


Fig. 19 Influence of angular location of groove on direct stiffness coefficient

i. The use of smaller length grooves (lower *a*/*L*) yielded the decrease in load carrying capacity. Whereas the use of smaller groove width increases the dimensionless load carrying capacity.
ii. Non-dimensional load capacity increases with increase in groove length ratio (a/L) whereas it decreases with increase in groove width ratio (w/D). The optimum dimension of groove for better static and dynamic characteristic of bearing were obtained as a/L = 0.8, w/D = 0.02 and Ist groove angle $\alpha_{1g} = 0^0$ (90^0 from load line).
iii. The location of the grooves with respect to the load line is found to affect strongly most performance parameters due to the strong interference of the grooves in the hydrodynamic pressure field. The optimum location of the groove axis was obtained to lie between 60° and 90° ($\alpha_{1g} = 0^0$) to the load line.

It may be concluded with confidence that the lubricant supply conditions play an important role in the performance of hybrid journal bearings and therefore should not be neglected in bearing analysis. Their optimization can result in substantial energy savings and reduced environmental impact. Truly, in some cases the supply conditions might even dictate the occurrence or the avoidance of bearing seizure.

References

1. Morton, P.G., Johnson, J.H., Walton, M.H., 1987, "The influence of grooves in bearing on the stability and response of rotating systems", Tribol. Series, 11, pp. 347-354.
2. Basri, H. and Neal P.B., 1990, "Oil flow in axial groove journal bearings". Proc. of the Seminar on Developments in Plain Bearings for the '90s, IMechE Seminar, Tribology Group, May 1990, pp. 11–17.
3. Ma, M.T., Taylor, C.M., 1992, "A theoretical and experimental study of thermal effects in a plain circular steadily loaded journal bearing", IMechE Seminar: Plain Bearings- Energy Efficiency and Design, London: Mech. Engg. Pub. Pvt. Ltd.
4. Ma, M.T., Taylor, C.M., 1996, "An experimental investigation of thermal effects in circular and elliptical plain journal bearings, Tribol. Int., 29(1), pp.19-26.
5. Chun, S.M. and Ha, D.H., 2001, "Study on mixing flow effects in a high-speed journal bearing", Tribol. Int., 34, pp.397-405.
6. Costa, L., Mirinda, A. S., Fillon M., Claro, J.S.P., 2003, "An analysis of the influence of oil supply condition on the thermo-hydrodynamic performance of a single groove journal bearing", IMechE, PartJ: J. Engg. Tribol., 217, pp.133-144.
7. Jeddi, L, El Khlifi, M. and Bonneau D, 2005, "Thermo-hydrodynamic analysis for a hydrodynamic journal bearing groove", IMechE, PartJ: J. Engg. Tribol., 219, pp. 263-274.
8. Brito, F.P. Miranda, A.S., Claro, J.C.P., Fillon M., 2012 "Experimental comparison of the performance of a journal bearing with a single and a twin axial groove configuration", Tribol. Int., 54, pp. 1-8.
9. Constantinescu V. N., 1967 "The pressure equation for turbulent lubrication", Proc. of IMechE, at London, 182(3A) pp.383-400.
10. Gethin, D.T. and Deihi El, M.K.I., 1987, "Effect of loading direction on the performance of a twin axial groove cylindrical bore bearing", Tribol.Int., 20 (4), pp.179-185.

Understanding Asynchronous Distributed Collaboration in an Enterprise Systems Engineering Context

Gary L. Klein, Jill L. Drury, and Sherri L. Condon

1 Introduction

Systems engineering (SE) efforts have been growing increasingly complex (Johnson and Yonas, 2006; Stevens, 2011). While systems engineers are sometimes asked to develop a single, well-defined system with a few designated user groups, often SE efforts are complex in multiple dimensions. New SE efforts may, in fact, build systems-of-systems (SoS; a class of mega-systems), within which some of the constituent systems are not under the control of the engineers, and may include extending and evolving an existing, heterogeneous, mission-critical system while not disrupting current operations. Many times, these SoS have multiple, diverse sets of intended end users and large numbers of interconnections within and beyond the SoS boundaries. Further, the intended functionality of the SoS may affect how the end users will work and may change their organization's structure, policy, or processes.

There has been a renewed emphasis on incorporating user participation into all phases of the SE process to help understand users' needs, to make engineering judgments and tradeoffs in users' best interests, and to increase users' sense of ownership. This emphasis can be seen in a description of Enterprise Systems Engineering (ESE) that implies close coupling with users: "In performing enterprise systems engineering, we engineer the enterprise and we engineer the systems that enable the enterprise. In particular, we help customers shape their enterprises, aligning technology to support goals. We support their business planning, policy-making, and investment strategies" (MITRE, 2012, p. 34).

The significance of end users and other stakeholders in the success of ESE is further evident in the *Guide to the Systems Engineering Body of Knowledge (SEBoK)*, where it is explained that the "primary purpose of an enterprise is to create value for society, other stakeholders, and for the organizations that participate in that enterprise" and "[t]he ability of an organization to create value is critically dependent on the people it employs, on what they know, how they work together, and how well they are organized and motivated to contribute to the organization's purpose" (Pyster et al., 2012). These claims motivate not only collaboration among end users and other stakeholders, but also understanding the characteristics of the enterprise's organizations.

Paradoxically, the larger the enterprise, the greater the need for user participation; but also the greater the difficulty to achieve this participation via face-to-face activities (Drury and Cuomo, 1997). Users are more likely to be dispersed geographically and across time zones. There may be only a few members in some of the user groups, and they may perform critical roles that impede travel to the ESE team. Even when the ESE team members are willing to travel to the users' locations, they may be working in facilities that are difficult for those outside the organization to access (for example, in nuclear submarines or in air defense command and control centers).

Therefore, addressing this paradox requires an alternative to face-to-face collaboration with users by employing technologies that enable collaboration across distance and time: distributed asynchronous collaboration. Especially with the advent of new social media platforms that support group work in professional settings (e.g., Holtzblatt et al., 2013; Muller et al., 2012), there are more mechanisms than ever for working with users remotely and asynchronously via software that enables threaded discussions, blogging, microblogging, sharing files, and/or collaborative authoring via wikis.

Indeed, we assert that because such technological support for collaboration is so foundational to enabling ESE teams to work with end users, the quality of such tools constitutes an "ESE Success Variable" (ESV). Furthermore, the quality of these tools will also impact other ESVs, such as the appropriate application of organizational change management

G.L. Klein (✉) • J.L. Drury • S.L. Condon
The MITRE Corporation, McLean, VA and Bedford, MA, USA
e-mail: gklein-at-mitre.org; jldrury-at-mitre.org; scondon-at-mitre.org

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_80,

techniques to "engineer the enterprise" (MITRE, 2012) because all such ESE techniques that involve users will require remote asynchronous collaboration.

This paper presents for the first time our model of ESVs and includes an assessment of the impact of collaboration support on the models' components. We created this model to ground our research in understanding the characteristics of an environment that leads to successful ESE. We are continuing our investigation by interviewing ESE teams to determine the factors that contribute to their success or failure. We are using the model components to generate questions for our interview protocol, and our method for doing so is explained below.

2 ESE Success Variables (ESVs)

We identified four major categories of ESVs for ESE systems acquisitions that fall into the complex portion of Snowden's Cyneform framework (Snowden, 2000) and that include direct user interaction with the finished systems (versus embedded computing efforts). As described above, *collaboration support* during the ESE process is foundational. Next, there are many possible techniques and approaches that can be used in executing ESE tasks, so the variability in *ESE technique application* affects the success of the ESE effort. The *characteristics of the system* that is being acquired also affect success. For example, a complex system requiring many interfaces with other systems is inherently more difficult to engineer than a less complex one with fewer interfaces. Finally, success is affected by the degree to which *organizational characteristics* need to be engineered to match the demands of the new system.

2.1 Collaboration Support

Proponents of collaborative engineering (e.g., Inoue et al., 2013; Lu et al., 2007) promote the need to work more effectively with all stakeholders across cultural, disciplinary, geographic, and temporal boundaries. Much of the motivation for collaborative engineering emerged from the concurrent engineering approach, which emphasizes simultaneous work on different phases of the product lifecycle and requires collaboration among diverse development teams that represent those different stakeholders (Koufteros et al, 2001; Umemoto et al., 2004). Under a remote collaborative environment, this perspective further emphasizes the need for high quality collaboration support and consequently some way for evaluating the quality of that support.

One approach would be to apply Herbert Clark's psycholinguistics work that determined successful communications tend to exhibit eight different behaviors (Clark, 1996). These were codified in the Collaboration Evaluation Framework (Klein, Adelman and Kott, 2008) as:

- Connection – locating with whom to collaborate and how to contact them
- Transmission – sending a message
- Notification – alerting the intended party of an incoming transmission
- Identification – designating the sender, receiver, and subject of a transmission
- Confirmation – notifying the sender of a transmission that it has been received
- Synchronization – orchestrating actions to facilitate joint action
- Common Ground Preservation – establishing and maintaining a shared context and shared meanings in transmissions

An example of support for *connection* is a directory of collaborators and their roles. The ability to send email or text messages supports *transmission*. *Notification* could consist of a text message saying that a new shared file had been posted. Email messages normally contain *identification* information regarding the sender, receiver, and subject; and an email receipt provides *confirmation*. A group calendar could aid in *synchronization*. *Common ground preservation* is challenging because it pertains to people's abilities to truly understand what each other means. Video can possibly assist because it provides richer awareness cues derived from facial expressions and postures compared to voice or text/email.

2.2 ESE Technique Application

Part of our ultimate research goal is to explore the assertion that four supporting ESE techniques are critical for engineering complex systems, as follows.

Cognitive Systems Engineering (CSE) is "an approach to the design of technology, training, and processes intended to manage cognitive complexity in sociotechnical systems. In this context, 'cognitive complexity' refers to activities such as identifying, judging, attending, perceiving, remembering, reasoning, deciding, problem solving, and planning" (Militello et al., 2010, p. 263). CSE is needed for designing user interaction that takes cognition into account.

To effectively achieve CSE, ESE team members need to execute activities such as cognitive task analyses to understand in detail how users currently think about and perform their tasks. These analyses require some combination of observing users in their normal workplace, interviewing, and analyzing job aids such as specialized systems' users' manuals. Observation at a distance is difficult to achieve but

could possibly be approximated via video and/or screen sharing; while interviews can occur via phone or VTC. Note that both of these activities require synchronous collaboration at a distance, and there are few options for asynchronous CSE activities.

The primary emphasis of *Participatory Design (PD)* consists of direct end user involvement during the design portion of the ESE effort. "Participatory design (PD) is a set of theories, practices, and studies related to end-users as full participants in activities leading to software and hardware computer products and computer-based activities" (Muller and Druin, 2012, p. 1125). PD can help to ensure that users' needs are taken into account during all performance trade-off and design decisions. There are a few PD activities that can occur asynchronously (e.g., retrospective video analysis as described in Muller, 1991), but most are designed for synchronous use.

One definition for *Organizational Change Management (OCM)* is: "activities involved in (1) defining and instilling new values, attitudes, norms, and behaviors within an organization that support new ways of doing work and overcome resistance to change; (2) building consensus among customers and stakeholders on specific changes designed to better meet their needs; and (3) planning, testing, and implementing all aspects of the transition from one organizational structure or business process to another." (Government Accountability Office, undated). OCM is needed to ensure that technology adoption occurs smoothly, and that the enterprise's members are aided in adapting their business processes, roles, responsibilities, and/or policies to complement the new system.

OCM preparations can take advantage of CSE and PD activities, which should yield a rich array of insights regarding how users currently work and the ways in which their work processes may need to change with the new system. Additional collaboration support such as wikis for joint asynchronous editing may be needed to create products such as policy statements or training manuals.

Systems-of-Systems (SoS) engineering "deals with planning, analyzing, organizing, and integrating the capabilities of a mix of existing and new systems into an SoS capability greater than the sum of the capabilities of the constituent parts" (DoD, 2004). SoS techniques are needed to help ensure that a new system will function effectively as a part of a larger set of interoperating systems.

Asynchronous collaboration is needed during SoS tasks to extract information about the systems and organizations that users interact with on a regular basis. Many of the collaboration support mechanisms would likely be appropriate for these tasks, such as joint authoring via wiki pages.

2.3 System Characteristics

Characteristics of the system that is being acquired constitute a significant set of factors that impact both the need for collaboration with end users and, we believe, the success of the ESE effort. There are many ways to characterize systems, but the degree of complexity of the system is frequently cited as a major dynamic in models and descriptions of engineered systems (Calantone and Di Benedetto, 2000; Childs and McLeod, 2013; Valle and Vázquez-Bustelo, 2009). Honour (2006) has argued that "the essence of systems engineering is that it is the engineering of complexity." Snowden's Cynefin framework distinguishes four types of contexts ranging from *simple*, in which best practices are applied to clear cause and effect relations, to *chaotic*, in which "[t]he relationships between cause and effect are impossible to determine because they shift constantly and no manageable patterns exist" (Snowden and Boone, 2007, p. 3). In *complicated* contexts, cause and effect relations are clear, but may require considerable analysis and expertise to understand, and there are multiple good solutions rather than one best one. Complicated systems such as Microsoft Word combine many different functions or subsystems, but their interactions are predicable. For complicated systems, collaboration with end users can contribute important knowledge and expertise to the process of constructing solutions, and users' preferences and constraints can guide selections among multiple options.

Complex contexts, where "much of contemporary business has shifted," are characterized by unpredictability and flux, and right answers are replaced by "instructive patterns" (Snowden and Boone, 2007, p. 5). In complex systems such as Adobe PhotoShop, the many different functions can be combined in multiple ways to produce an infinite range of results. Honour (2006) views systems engineering practices as methods of managing the interacting parts, reflexive feedback mechanisms, emergent properties, and adaptability of complex systems. There is a sense in which ESE is always complex: when users and their organizations adapt to the introduction of new systems into their practices, they are likely to develop new expectations for the systems, new ways of using the systems or integrating them into their workflows, and new ideas for the next updates. Because ESE includes these kinds of interactions and feedback mechanisms, solutions can be moving targets, and problems tend to be managed rather than solved. As unpredictability increases and systems grow complex or veer toward chaotic, input from users and their organizations becomes essential to manage emergent issues and outcomes.

Other properties of systems also constitute significant factors for both engineering success and the need for

collaboration with users. The size of the system, even if it does not introduce complexity, can present challenges for control and integration. Large systems that perform many functions are likely to impact more users with varying goals and more users from different parts of the organization with different needs and perspectives. When developing systems that incorporate new technologies, engineers cannot rely on established practices or experience and there are no precedents from which to anticipate impacts on users or their organizations. Systems with very high performance demands for qualities such as throughput or security complicate the engineering process and may have unforeseen consequences for users. All of these factors motivate close collaboration with users and their organizations. Moreover, Honour (2004) observes that engineers often need to make tradeoffs among these factors, and collaboration with users can generate valuable input for decision-making as well as help users understand why solutions were selected.

2.4 Organizational Characteristics

Many characteristics of the end users' organization(s) may affect the success of the ESE efforts, but we present here one class of characteristic as an illustration: Thompson's three types of interdependence (1967).

Under *pooled* interdependence, each person or unit provides a discrete contribution to the whole so that the result is a collated (or pooled) set of information and knowledge. An example is an air defense system, in which the database of automatically tracked aircraft is refined incrementally by individual operators' manual entries as they learn about the aircraft via external sources and/or apply their a priori knowledge. In this example, the air defense system database itself becomes an asynchronous collaboration tool for end users because any operator can access and filter it.

Under *sequential* interdependence, the product of one person or unit depends upon the output of another. An example can be seen in an en route air traffic control (ATC) system, where an aircraft transits through one controller's region before it can be handed over to the controller of another region. The controllers collaborate via the ATC database, display, and voice intercom to make the necessary transitions.

Under *reciprocal* interdependence, people or units pose critical contingencies for each other that have to be resolved before taking action. For example, operations and logistics often have a reciprocal interdependence. Whether an operation can occur depends on the availability of specific resources, and the availability of those resources depends on previous and planned operations. Personnel engaging in reciprocal interdependence often require real-time (synchronous) collaboration support to negotiate tightly coupled tasks.

ESE teams should determine the type(s) of interdependence that the new system needs to facilitate. The task of determining interdependence can occur using a range of asynchronous collaboration technologies employed among ESE team members and end users, particularly during the requirements elicitation and design phases.

3 A Model of ESE Success

The four ESVs discussed above form the primary components of our model of a successful ESE ecosystem, as seen in Figure 1. This model also takes into account the interactions among these ESVs, shown as junctions in the figure.

Organizational factors—such as whether tasks are performed in a pooled, sequential, or reciprocal manner—can affect the outcomes of CSE, PD, OCM, or SoS efforts. For example, consider the reciprocal interaction between sourcing of materials and production operations: the setup of the assembly line is dependent on the types of materials being supplied, while the types of materials to be acquired reflect the products and form of assembly. This requires negotiation between acquisition and production departments and constant information sharing, which means that systems, organizational structures, and procedures must support an intensive two-way relationship. Clearly, successful enterprise systems engineering under reciprocal interdependence requires application of all four ESE techniques to address support this interdependence at all levels because changes in acquisition will affect production and vice versa.

4 The Model as Interview Protocol

We created questions that addressed the four major components of the ESE model (CSE, PD, OCM, and SoS; shown on the left-hand side of the model illustration in Figure 1). For each component, we began with a question designed to elicit whether that component was relevant to the respondent's SE effort. If the response was positive, we asked what techniques (if any) were used to address that component, inquiring about how the techniques were used to design both the technology and the processes for how that technology would be used. Further, we asked for the respondent's assessment of the success of using those techniques. Finally, after going through the four components, we asked for an assessment of the success of the effort as a whole. For example, the questions regarding CSE were the following (notes in square brackets are directed towards the reader and not the interviewee):

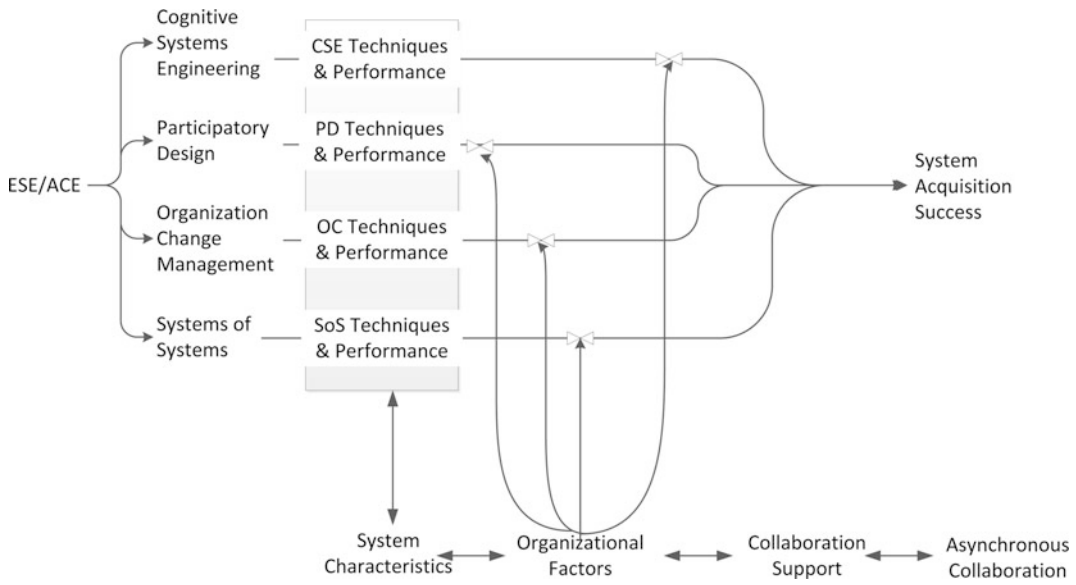


Figure 1. A Model of a Successful ESE Ecosystem

- Does your project involve designing user interaction that takes into account how the user thinks about their task? [If "yes," ask further:]
  - What techniques did you use to accomplish this design?
  - How successful was the use of those techniques?
- Does your project involve redesigning how the users' tasks fit into their overall workflow?
  - What techniques did you use to accomplish this redesign?
  - How successful was the use of those techniques?

Additional questions focused on the system characteristics to determine its level of complexity. Examples are:

- Did the system involve coordinating many different functions or subsystems?
- How large was the system in the following ways: how many functions? How many types of output? How much data did it process?

Other questions probed organizational characteristics such as:

- Did the system include technologies, groups of people or individuals who provide discrete contributions to the work process or products? For example, individual maids cleaning their floor of a hotel result in the whole hotel being clean. [This is probing for examples of pooled interdependence.]
- In the system being acquired, was the product of one person or group of people dependent on the output of another? For example, requests for information would fall into this category. [Addresses sequential interdependence.]
- Was the system designed such that there would be situations in which the product of two or more people or groups of people would have to be negotiated among them? For example, operations and logistics people must coordinate closely because the type of operations that can be executed depends on the current resources available. [Addresses reciprocal interdependence.]

The interview protocol contained questions about the project's support for collaboration behaviors, such as:

- During your work, did you have information or support regarding whom you should be communicating with? [*connection* support]
- Did technology support you in establishing and maintaining a common understanding with your collaborators? [*common ground preservation* support]

We piloted the interview protocol and examined the results. In particular, our pilot interviewee had used a number of ESE techniques and was able to explain the ways in which each had helped (or not). We assessed the interviewee's use of the techniques in terms of the degree of formalism utilized and the thoroughness or breadth of their application.

We then revised the protocol to make the wording clearer and more concise, delete questions that didn't elicit unique material, and added questions that addressed the interaction among the ESVs (the junctions in Figure 1). An example of a question addressing an interaction is:

- We have talked about ESE techniques, organizational characteristics, support for collaboration, and qualities of the technology.
  - Do you think that any of these factors affected each other?
  - Did that impact the success of the project?

5 Discussion and Future Work

The process of developing the model made it clear that discerning the impact of organizational factors on better versus worse outcomes of ESE efforts is indeed a complex endeavor. This impact is not only mediated by type and quality of ESE techniques employed, but also the type and quality of user participation. That participation is often a function of support for asynchronous collaboration, especially in large enterprises. We intend to incrementally explore the interactions represented in the model. Because ESE efforts will be more successful if their collaborative systems support the eight collaborative behaviors defined above, we plan to conduct controlled experiments in which experienced systems engineers solve ESE problems asynchronously in conjunction with distributed users played by experimenters while varying the support for collaborative behaviors from simple email exchanges to an enhanced multifunction virtual collaboration environment. These functions may be as simple as a directory authored on a wiki page, which could aid in connection when team members don't know each other at project startup. Another example is a utility that will send text message alerts to aid in urgent notification for members who are not connected to their work email account at all times. By varying this support for collaborative behaviors, and varying the types of ESE scenarios to reflect different combinations of the modeled factors we can demonstrate some of the interactions implied by the model. Our ultimate aim is to increase the likelihood of success of ESE efforts.

Acknowledgements We thank M. Francesca, R. Modeen, and G. Rebovich for helpful peer review. This work was supported by MITRE Corporation Systems Engineering Innovation Project 19MSR080-BA. Approved for Public Release; distribution unlimited; case #14-1336.

References

1. Calantone, R. J., & Di Benedetto, C. A. (2000). Performance and time to market: accelerating cycle time with overlapping stages. *Engineering Management, IEEE Transactions on*, *47*(2), 232-244.
2. Childs, S., & McLeod, J. (2013). Tackling the wicked problem of ERM: using the Cynefin framework as a lens. *Records Management Journal*, *23*(3), 191-227.
3. Department of Defense (2004). System of Systems Engineering, in *Defense Acquisition Guidebook*, Washington, DC, October 14, 2004.
4. Drury, J. & Cuomo, D. (1997). Usability issues in complex government systems. In NIST Special Publication 500-237, *Symposium Transcription, Usability Engineering: Industry-Government Collaboration for System Effectiveness and Efficiency*. Gaithersburg, MD: National Institute for Standards and Technology.
5. Government Accountability Office. (undated). BPR Glossary of Terms. [Online]. Avail-able: http://www.gao.gov/special.pubs/bprag/bprgloss.htm
6. Holtzblatt, L., Drury, J. L., Weiss, D., Damianos, L., & Cuomo, D. (2013). Evaluating the uses and benefits of an enterprise social media platform. *J. of Social Media for Organizations*, Vol 1(1). Available: www.mitre.org/jsmo.
7. Honour, E. C. (2004). Understanding the value of systems engineering. In *Proc. of the INCOSE International Symposium*, 1-16.
8. Honour, C.E. (2006). Systems engineering and complexity. In *Proc. of the Conference on Systems Engineering Research*, Los Angeles, CA 2006, Paper 193-4.
9. Inoue, M., Nahm, Y. E., Tanaka, K., & Ishikawa, H. (2013). Collaborative engineering among designers with different preferences: Application of the preference set–based design to the design problem of an automotive front-side frame. *Concurrent Engineering*, *21*(4), 252-267.
10. Johnson, C. & Yonas, G. (2006). The wicked future of systems engineering. [Online] Available: www.incose.org/enchantment/docs/06Docs/06DecWickedSE.ppt
11. Klein, G. L., Adelman, L., & Kott, A. (2008). Enabling collaboration: Realizing the potential of network-enabled command. *Battle of Cognition, A. Kott (Ed.), Westport, CN: Praeger Security International*.
12. Koufteros, X., Vonderembse, M., and Doll, W. (2001). Concurrent Engineering and its consequences. *Journal of Operations Management*, 19, 97-115.
13. Lu, S. C-Y., ElMaraghy, W., Schuh, G., and Wilhelm, R. (2007). A scientific foundation of collaborative engineering. *CIRP Annals-Manufacturing Technology*, 56(2), 605-634.
14. Militello, L. G., Dominguez, C. O., Lintern, G. and Klein, G. A. (2010). The role of cognitive systems engineering in the systems engineering design process, *Systems Engineering*, vol. 13, no. 3.
15. The MITRE Corporation, *Systems Engineering Guide* (2012) Ebook published 5 June 2012 under ISSN 2169-9372. Available: http://www.mitre.org/publications/systems-engineering-guide/systems-engineering-guide
16. Muller, M. (1991). PICTIVE: An exploration in participatory design. In *Proc. of the CHI '91 Conference on Human Factors in Computing Systems*.
17. Muller, M. J. & Druin, A. (2012). Participatory Design: The third space in HCI, in *The Human-Computer Interaction Handbook*, 3rd ed., J. Jacko, Ed. New York, NY: Taylor & Francis.
18. Muller, M., Erlich, K., Matthers, T., Perer, A., Ronen, I., & Guy, I. (2012). Diversity among enterprise online communities: Collaborating, teaming, and innovating through social media. In *Proc. of the CHI Human Factors in Computing Systems Conference*.
19. Pyster, A., Olwell, D. H., Hutchison, N., Enck, S., Anthony Jr, J. F., & Henry, D. (2012). Enterprise Systems Engineering. *Guide to the Systems Engineering Body of Knowledge (SEBoK)*. Available: http://www.sebokwiki.org/wiki/Enterprise_Systems_Engineering
20. Snowden, D. (2000). Cynefin: a sense of time and space, the social ecology of knowledge management, in Despres, C. &

Chauvel, D. (Eds), *Knowledge Horizons: The Present and the Promise of Knowledge Management*, Oxford: Butterworth-Heinemann.
21. Snowden, D. J., and Boone, M. E. (2007). A leader's framework for decision making. *Harvard Business Review*, *85*(11), 68.
22. Stevens, R. (2011). *Engineering Mega-Systems: The Challenge of Systems Engineering in the Information Age*, Boca Raton, FL: Auerbach Publications.
23. Thompson, J. D. (1967). *Organizations in action*. NY, NY: McGraw-Hill.
24. Umemoto, K., Endo, A., & Machado, M. (2004). From sashimi to zen-in: the evolution of concurrent engineering at Fuji Xerox. *Journal of Knowledge Management*, *8*(4), 89-99.
25. Valle, S., & Vázquez-Bustelo, D. (2009). Concurrent engineering performance: Incremental versus radical innovation. *International J. of Production Economics*, 119(1), 136-148.

A Design Model for Rapid Transit Networks Considering Rolling Stock's Reliability and Redistribution of Services During Disruptions

Esteve Codina, Ángel Marín, and Lídia Montero

1 Introduction

Designing a Rapid Transit Network (RTN) or even extending one that is already functioning, is a vital subject due to the fact that they reduce traffic congestion, travel time and pollution. Usually a RTN is in operation with other transportation systems such as private transportation (car) and this makes that the design must take into account this factor. Another factor that needs to be considered is the capability of the newly designed system to keep operating under more or less suitable conditions under a set of predictable disruptions.

In Bruno G. *et al*. [1], a RTN design model is presented where the user cost is minimized and the coverage of the demand by public network is made as large as possible. Marín in [8], studies the inclusion of a limited number of lines. Also, Laporte G. *et al* in [6] build robust networks that provide several routes to passengers, so in case of failure part of the demand can be rerouted. Connections between two-stage stochastic programming network design and recovery-robustness in railway networks planning models have been studied by Cicerone *et al*. in [4] and by Cacchiani *et al*. in [2]. Also, in [3] Cadarso and Marín develop a two-stage stochastic programming model for rapid transit network design in which disruptions probabilities are assumed known a priori, illustrating some of its recoverable robustness properties.

This paper presents a conceptual scheme that permits to incorporate a probability model for the disruptions of a RTN. It is assumed that disruptions arise when transportation units present some failure during operation leaving a link blocked. Other sources of disruption with their associated scenarios could be added, but this is not done for ease of exposition. As a consequence of this, the disruption probabilities will depend on the level of traffic on the network links. The probabilities of failure follow the following hypothesis: a) disruptions are due to a single event and scenarios with several simultaneous disruptions are discarded a priori as they are assumed to have a much lower probability, b) a preselected set of scenarios is considered, c) the number of failures that a train unit may experience along a large number of services distributes accordingly to a geometrical law and the individual probability of failure of a service is constant along the planning horizon and depends only on the train unit characteristics (e.g., quality of material and maintenance). The resulting model has a bilevel structure and it is solved by a specific heuristic method.

2 Structure and elements of the design model

In this RTN design model it is assumed that the location of potential stations is known. There already exists a current mode of transportation (for example, private cars or an alternative public transportation is already operating in the area) competing with the new RTN to be constructed. The aim of the model is to design a network, i.e. to decide at which nodes to locate the stations and how to connect them covering as many trips by the new network as possible.

– A potential network (N, A) is considered from which the optimum rapid transit network is selected. The node set is composed by centroids (N_c) and stations at RTN (N_r), the node set is then $N = N_c \cup N_r$. Links will be denoted either by a single subscript (e.g., a) or by a double subscript (i.e., (i, j)) when considered convenient. Because both riding directions are always considered, the set of potential links is so that $(i,j) \in A \Leftrightarrow (j,i) \in A$. $E(i)$ and (i) are the set outgoing and incoming nodes to node i respectively.

E. Codina (✉) • L. Montero
Dept Statistics and Operations Research, BarcelonaTECH-UPC, Carrer Jordi Girona 1-3, 08034 Barcelona, Spain
e-mail: esteve.codina@upc.edu; lidia.montero@upc.edu

Á. Marín
Dept. of Applied Mathematics and Statistics, UPM, Plaza Cardenal Cisneros, 3., 28040 Madrid, Spain
e-mail: angel.marin@upm.es

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_81,

– Each feasible link (i, j) has a generalized travel cost which may depend on the scenario of disruption. This is further discussed in section 4.
– For simplicity, in this model it will be assumed that the planners have selected a priori a set of candidate $\widehat{L}$ lines, being $\widehat{L}$ a large number. Lines will be considered an ordered chain of n links $\{a_1, a_2, ..., a_n\}$ all of them appearing only once in the sequence. Lines with circulation in both directions will be treated as two separate lines $\{a_1, a_2, ..., a_n\}, \{a_n, a_{n-1}, ..., a_1\}$. To take into account the recovery of the disruptions that may arise in a link $a \in \ell$, for each line $a \in \widehat{\ell}$, an additional set of lines must be considered that will operate only in the scenarios of disruption. These lines will be referred to as the recovery lines, whereas lines in $\widehat{L}$ will be referred to as the primary lines or also as the candidate lines. Thus, if $\ell = \{a_1, a_2, a_3\}$, for a disruption in link a_1, the recovery line must be $\{a_2, a_3\}$. For a disruption in link a_2, the recovery lines that must be considered are $\{a_1\}$ and $\{a_3\}$ and finally, for a disruption in link a_2, the recovery line that must be considered is $\{a_1,a_2\}$. The set of all recovery lines will be denoted by $\widehat{L}'$ and the set of recovery lines for line $\ell \in \widehat{L}$ will be denoted by $\widehat{L}'(\ell)$. The number of recovery lines in $\widehat{L}'(\ell)$ for a line $\ell \in \widehat{L}$ is $\left|\widehat{L}'(\ell)\right| = 2(|\ell| - 1)$, where $|\ell|$ is the number of segments in line ℓ. If $L = \widehat{L} \cup \widehat{L}'$, the total number of lines is then $|L| \leq \left|\widehat{L}\right| + 2\sum_{\ell \in \widehat{L}} (|\ell| - 1)$, from which only ν will be finally included in the solution ($\nu \leq \left|\widehat{L}\right|$). Further definitions are:
– $\widehat{L}(a) \subseteq \widehat{L}$ is the subset of candidate or primary lines containing segment $a \in A$.
– $\widehat{L}'(a)$ is the subset of recovery lines associated to the set of primary lines containing link a.
– The model considers a set of scenarios associated to regular conditions of operation of the transport system (i.e., morning peak period, afternoon, night, holidays,. . .). Each of these scenarios is assumed to extend during a given time period of length H_r (i.e., 3 hours for morning peak periods). This set of scenarios will be denoted by S_0 and for any $r \in S_0$, there will be associated a weight or probability $q_r > 0$ associated to its relevance, so that $\sum_{r \in S_0} q_r = 1$. A typical way of evaluating the weights q_r is accordingly to their associated total demands, i.e.: $q_r = g^r/\widehat{G}$, where $g_r = \sum_{w \in W} g_w^r$ and $\widehat{G} = \sum_{r \in S_0} g_r$. For any scenario $r \in S_0$ a set of possible disruption scenarios $D(r)$ will be considered with probabilities $p_s > 0$, $s \in D(r)$, so that $p_r + \sum_{s \in S(r)} p_s = 1$. These disruption scenarios will be associated each with a breakdown of a service at a link. The set of that links for a regular scenario $r \in S_0$ will be denoted by $\widehat{A}(r)$. For each $a \in \widehat{A}(r)$, $s(a)$ will denote the associated disruption scenario and for each scenario $s \in D(r)$, $r \in S_0$, $a(s)$ will denote the disrupted link. Finally, by S it will be denoted the set of all possible scenarios, i.e., $S = S_0 \cup_{r \in S_0} D(r)$.
– Users may choose between two transportation modes: a private mode (typically car) or the public transportation mode comprising a set of lines, some of them already in operation and some others that will be the outcome of this design model. The model's demand will take into account differences between scenarios $r \in S_0$. The total demand (private + public transport) for scenario $r \in S_0$ is given by the trip matrix $G^r = (g_w^r)$, where g_w^r is the total number of trips from origin $o(W)$ to destination $d(W)$. For a particular scenario $s \in S$, the trip travel time for o-d pair W through the private transportation network is given by the matrix $U_c^s = (u_c^{w,s})$ and the trip travel time for using public transportation is given by the matrix $\Lambda^s = (\lambda^{w,s})$. The model assumes a modal choice for each o-d pair given by a logit model, i.e., the proportion of trips ξ_w^s using the private transportation mode is given by:

$$\xi_w^s = \frac{exp\left(-\beta_c^w - \eta u_c^{w,s}\right)}{exp\left(-\beta_c^w - \eta u_c^{w,s}\right) + exp\left(-\beta_{PT}^w - \eta \lambda^{w,s}\right)} \qquad (1)$$

where β_{PT}^w is proportional to the price of fares for public transport in the planning period, β_c^w is proportional to the parking cost plus the cost of gasoline for the trip from $o(W)$ to $d(W)$ and η is proportional to the user's value of time.
– Let c_x and c_ψ denote the link vector costs and the node vector of location costs respectively.

The design model has two stages or levels: a) in the first "planning" stage, the decision variables x,y are chosen, i.e., the topology of the network is set and b) in a second stage, at a given scenario, the passenger flows make use of the network designed in the first stage, taking into account the scenario characteristics.

2.1 Variables and constraints in the 1st stage

A link-line incidence matrix $(\delta_{a,\ell})$ will be assumed known with elements $\delta_{a,\ell} = 1$ if candidate line ℓ contains link a and 0 otherwise. Let h_ℓ, $\ell \in L$ be a binary variable indicating whether candidate line ℓ is chosen or not. Let also χ_a be a binary variable so that $= 1$ if arc a is located and $= 0$, otherwise. The following constraints force that link a must be built if some line ℓ using it is chosen:

$$M\chi_a \geq \sum_{\ell \in \widehat{L}} \delta_{a,\ell} h_\ell, \quad a \in A \qquad (2)$$

The binary variables x_a^l state whether link a is required because line l is included in the solution or not. These variables are related to the variables h_l through the following constraints:

$$h_\ell \le x_a^\ell, \quad a \in \ell, \quad \ell \in \widehat{L} \tag{3}$$

A limitation on the number of lines can be imposed by $\sum_{l \in L} h_l \le v$. Let now ψ_i be a binary variable so that $= 1$ if station i is located and $= 0$, otherwise. Then, variables X and ψ are linked by:

$$\begin{aligned} &\chi_a \le \psi_i, \quad \forall i \in N, \forall a = (i,j) \in A \\ &\chi_a \le \psi_j, \quad \forall j \in N, \forall a = (i,j) \in A \end{aligned} \tag{4}$$

2.2 *Variables and constraints of the 2nd stage*

- $v_a^{w,s}$, is the passenger flow on link $a \in A$ for origin destination pair w under scenario $s \in S$. By $v^{w,s} = (..., v_a^{w,s}, ...; a \in A)$ it will be denoted an arc flow vector per o-d pair w and scenario $s \in S$.
- $v_a^{w,s}$, is the flow for o-d pair w using private transport in scenario s. By $v_c^{w,s} = (..., v_c^{w,s}, ...; w \in W)$ it will be denoted the flow vector of passengers using private transportation in scenario $s \in S$.

The balance constraints for flows at a given scenario s will be:

$$\sum_{j \in E(i)} v_{ij}^{w,s} - \sum_{k \in I(i)} v_{ki}^{w,s} = \begin{cases} v_{PT}^{w,s} & if\, i = p(w) \\ -v_{PT}^{w,s} & if\, i = q(w), i \in N, w \in W, s \in S \\ 0 & otherwise \end{cases} \tag{5}$$

where $v_{PT}^{w,s} \ge 0$ are the flows using public transport mode at o-d pair $w \in W$, that must verify:

$$v_c^{w,s} + v_{PT}^{w,s} = g_w^s, w \in W, s \in S \tag{6}$$

Also, link flows $v_a^{w,s}$ for scenario s will be subject to the location-allocation constraints, which in fact are equivalent to suppress the links for which the decision variables X_a annul:

$$v_a^{w,s} \le M \chi_a, \quad a \in A, \quad s \in S, \quad w \in W \tag{7}$$

2.3 The conceptual model for modal split

Formulations in this subsection do not include links a for which decision variables $X_a = 0$. Also flow variables are assumed for a generic scenario $s \in S$ and this superscript will be omitted.

Borrowing ideas from combined modal split-assignment models in transportation planning (see, for instance [5]) the following convex problem

$$\begin{aligned} Min_{v, v_c, v_{PT}} \quad & \sum_{w \in W} \left\{ d^T v^w + v_c^w \left(log v_c^w - 1 - \beta_c^w - \eta u_c^w \right) + \right. \\ & \left. + v_{PT}^w \left(log v_{PT}^w - 1 - \beta_{PT}^w \right) \right\} \\ s.t.: \quad & v, v_c, v_{PT} \ge 0 \text{ and verify constraints } (5), (6) \\ & \sum_{w \in W} v_a^w \le m_P \sum_{\ell \in \widehat{L}} z^{\ell, r} x_a^\ell, \; a \in A \end{aligned} \tag{8}$$

provides solutions verifying the modal split accordingly to (1). A linearization of this model is used by López in [7] in order to reformulate it as a mixed linear integer problem. Capacity constraints arise because the capacity of the public transport lines operating on the network links as a function of the number of services z^l of the lines. Variables X are considered implicitly and because of that, the solution set of previous problem (8) will be denoted by $V^{s,*}(z^s, X)$, when specified for a specific scenario $s \in S$.

Next section describes a simplified model which provides the required number of services on the lines to attend the passenger's demand, accordingly to the scenarios that are considered.

3 Service setting for normal operational conditions and recovery of disruptions

A model that states the number of services that must operate on each link is required for both non-disrupted scenarios $r \in S_o$ and scenarios corresponding to a disruption $s \in S \setminus S_o$. Let v^r the passenger vector flow on each of the network links for a non-disrupted scenario $r \in S_o$. Let v^r be the vector of total link flows which can be expressed as $v_a^r = \sum_{w \in W} v_a^{w,r}$ $a \in A$. Let γ^l the individual cost of a service on line $l \in L$ and let C^l be the time required to perform a complete service on l by a transport unit. A total of n_v transport units are assumed to operate on the network. Also, assume that the maximum number of services on link a for scenario $r \in S_o$ is $\hat{Z}_a^r$. Then, the following simple covering model will be used to determine the number of services for each line:

$$\begin{aligned}\Gamma_0^r(v^r,x) = Min_z \quad & \sum_{\ell\in\widehat{L}} \gamma^\ell z^{\ell,r} \\ & \hat{Z}_a^r \geq \sum_{\ell\in\widehat{L}} z^{\ell,r} x_a^\ell \geq \frac{1}{m_P} v_a^r, \;\; a\in A,\; r\in S_0 \\ & \sum_{\ell\in\widehat{L}} C^\ell z^{\ell,r} \leq n_v H_r, \\ & Z \ni z^{\ell,r} \geq 0\end{aligned} \tag{9}$$

The solution set of previous problem (9) will be denoted by $Z_r^*(v^r,x), r\in S_o$. If $Z^{r,*}$ is the vector of the optimal number of services for the lines in the regular scenario r given by the previous problem (9), then the total number of services θ_a^r on link a will be given by:

$$\theta_a^r(v^r) \stackrel{\Delta}{=} \sum_{\ell\in\widehat{L}} z^{\ell,r,*} x_a^\ell, \quad a\in A, \quad r\in S_0 \tag{10}$$

For a disruption on link $a\in\widehat{A}$, those lines $\ell\in\widehat{L}$ containing that link, i.e., $\ell \ni a$, can be put partially in operation as a recovery strategy using their disruption lines $\widehat{L}'(\ell)$ and in this case the model will evaluate a new set of services for all the lines in the system. Then, for a scenario $s\in S\backslash S_0$ corresponding to a disruption in link a, the set of lines that can potentially be operating is $\tilde{L}(a) = \widehat{L}'(a)\cup(\widehat{L}\backslash\widehat{L}(a))$.

The following problem establishes the services, z_s^l, that must be assigned for the lines operating in the scenario $s\in D(r)$ for a disruption of the regular scenario r:

$$\begin{aligned} & r\in S_0, s\in D(r): \\ & \Gamma_s^r(v^s,x) = Min_z \sum_{\ell\in\widehat{L}(a(s))} \gamma^\ell z^{\ell,s} \\ s.t.: \hat{Z}_b^s \geq & \sum_{\ell\in\widehat{L}\backslash\widehat{L}(a(s))} x_b^\ell z^{\ell,s} + \sum_{\ell\in\widehat{L}(a(s))} \sum_{\ell\in\widehat{L}'(\ell)} \hat{x}_b^\ell z^{\ell,s} \geq \frac{1}{m_P} v_b^s, b\in A, \\ & \sum_{\ell\in\widehat{L}(a(s))} C^\ell z^{\ell,s} \leq n_v H_r, \\ & Z \ni z^{\ell,s} \geq 0, \;\; \ell\in\widehat{L}(a(s)) \end{aligned} \tag{11}$$

where m_p is the maximum number of passengers that a unit may allocate. The solution set of previous problem (11) will be denoted by $Z_s^*(v^s,x), s\in D(r)$.

4 A model for probabilities of disruption

The probability p_s of each scenario is considered dependent on the use that is made on the designed network. By means of a failure model it will be possible to find an expression for the probability that a link presents a disruption during the operational horizon of the transit network. It will be assumed that the probability of failure of a service is mainly determined by the type of units operating in the service and the characteristics of the link. Let T be the set of type units operating on the network. Let $\pi_{a,\tau}$ be the joint individual probability that a service carried out by a unit of type $\tau\in T$ presents a disruption on link $a\in A$. By examining annual disruption reports from transit operators, the fraction of disrupted services with a disruption time of 20 minutes or more over the total number of services on a line is between $1.5\cdot 10^{-4}$ to $5.0\cdot 10^{-4}$, i.e. 1 disruption each 2000 or 6600 services. Assume that by analyzing statistically the previous mentioned annual disruption reports the probabilities $\pi_{a,\tau}$ have been determined. Let now $\theta_{a,\tau}$ be the total number of services of type τ on link a during the operational horizon used for our planning model (for instance, peak morning period). Let $T(a)$ be the set of unit types that operate on link $a\in A$. Let also $\tilde{\theta}_{a,\tau}$ be the total number of services with a relevant disruption out of the $\theta_{a,\tau}$ and $\tilde{\theta}_a \sum_{\tau\in T(a)} \tilde{\theta}_{a,\tau}$ the aggregated number of disrupted services on link a. It is assumed that $\tilde{\theta}_{a,\tau}$ follows a binomial distribution with probability $\pi_{a,\tau}$, i.e.: $\tilde{\theta}_{a,\tau} \sim Bino(\theta_{a,\tau}, \pi_{a,\tau})$. Thus, the probability $\widehat{P}_a$ that link $a\in A$ has at least one disrupted service from any unit type $\tau\in T(a)$, as a function of the number of services $\theta_{a,\tau}$ of type τ operating on that link is:

$$\begin{aligned} \widehat{P}_a \stackrel{\Delta}{=} P(\tilde{\theta}_a \geq 1) &= 1 - \prod_{\tau\in T(a)} P(\tilde{\theta}_{a,\tau} = 0) = \\ &= 1 - \prod_{\tau\in T(a)} (1-\pi_{a,\tau})^{\theta_{a,\tau}} \\ &= 1 - exp(-\sum_{\tau\in T(a)} \alpha_{a,\tau}\theta_{a,\tau}) \end{aligned} \tag{12}$$

where $\alpha_{a,\tau} = -log(1-\pi_{a,\tau})$. For small probabilities $\pi_{a,\tau}$, then $\alpha_{a,\tau} \approx \pi_{a,\tau}$. Also, the probability of having no disruption on link a of any of the type units $\tau\in T(a)$ is $\widehat{Q}_a \stackrel{\Delta}{=} 1 - \widehat{P}_a$.

Because the probability of more than one link with disruptions simultaneoulsy is small, the disruptions that will be considered are failures of a single link a within the set of links $\widehat{A}$ considered candidates for a disruption. Let $a(s)$ denote the link associated with scenario $s\in D(r), r\in S_0$. For ease of notation let $\widehat{A}_a = \widehat{A}\backslash\{a\}$. The probability of each scenario s corresponding to a disruption in link $a(s)$ will be evaluated now by a given function $F_r : \Re^{|\widehat{A}|} \to \Re^{|D(r)|+1}$ of the number of services $\theta_a^r, a\in\widehat{A}$ on the links candidates for a disruption. If there is a single type of units operating in the network then, the function $F^r(\cdot)$ for the probabilities p_s^r and p_0^r that will be adopted is:

$r \in S_0$:

$$p_s^r \triangleq F_s^r(\theta) = \frac{\exp(\alpha_{a(s)}\theta^r_{a(s)}) - 1}{1 + \sum_{b \in \widehat{A}} (\exp(\alpha_b \theta_b^r) - 1), s \in D(r)} \tag{13}$$

$$p_0^r \triangleq F_0^r(\theta) = (1 + \sum_{b \in \widehat{A}} (\exp(\alpha_b \theta_b^r) - 1))^{-1} \tag{14}$$

In case that probabilities $\pi_{a,\tau}$ are very small, then probabilities p_r^r of no disruption are much higher than the probabilities p_s^r associated with the disruption on a link.

5 A stochastic 2-stage model and a heuristic solution

Conceptually, the model could be formulated as the following bilevel programming problem:

$$\begin{aligned}
Min_{\substack{y,\chi,h \\ \Psi,v,z}} \quad & c_x^{\mathrm{T}}\chi + c_\psi^{\mathrm{T}}\Psi + \sum_{r \in S_0} \left\{ y_0^r \Gamma_0^r(v^r, x) + \sum_{s \in D(r)} y_s^r \Gamma_s^r(v^s, x) \right\} + \\
& + \vartheta \sum_{r \in S_0} \left[y_0^r \sum_{w \in W} \left\{ d^{r\,T} v^{w,r} + u_c^{w,r} v_c^{w,r} \right\} + \right. \\
& \left. + \sum_{s \in D(r)} y_s^r \sum_{w \in W} \left\{ d^{s\,T} v^{w,s} + u_c^{w,s} v_c^{w,s} \right\} \right] \\
s.t.: \quad & \text{constraints } (2),(3),(4),(7) \\
& y_0^r = q_r F_0^r(\ldots, \theta_a^r(v^r), \ldots; a \in \widehat{A}),\ r \in S_0 \\
& y_s^r = q_r F_s^r(\ldots, \theta_a^r(v^r), \ldots; a \in \widehat{A}),\ s \in D(r) \\
& \text{where } \theta_a^r(v^r) \text{ is defined in } (10) \\
& \text{resulting from lower level problem} (9) \text{ and } (11) \\
& \text{Also, from } (9) \text{ and } (11): \\
& z^r \in Z_r^*(v^r, x),\ v^r \in V^{r,*}(z^r, \chi),\ r \in S_0 \\
& z^s \in Z_r^*(v^s, x),\ v^s \in V^{s,*}(z^s, \chi),\ s \in D(r)
\end{aligned} \tag{15}$$

In order to solve heuristically the previous problem (15) the following mixed linear integer programming problem (16) needs to be considered. In this problem it is assumed that probabilities y_s^r, y_0^r are fixed and also that the total amount of transport trips $v_{PT}^{w,s}$ in public transport are known.

$$\begin{aligned}
Min_{\substack{\chi,h \\ \Psi,v,z}} \quad & c_x^T\chi + c_\psi^T\Psi + \sum_{r \in S_0} \left[y_0^r \sum_{\ell \in \widehat{L}} \gamma^\ell z^{\ell,r} + \right. \\
& \left. + \sum_{s \in D(r)} y_s^r \sum_{\ell \in \widehat{L}(a(s))} \gamma^\ell z^{\ell,s} \right] + \\
& + \vartheta \sum_{r \in S_0} \left[y_0^r \sum_{w \in W} \left\{ d^{r\,T} v^{w,r} \right\} + \right. \\
& \left. + \sum_{s \in D(r)} y_s^r \sum_{w \in W} \left\{ d^{s\,T} v^{w,s} \right\} \right] \\
s.t.: \quad & \text{constraints } (2),(3),(4),(5),(7) \\
& + \text{constraints in problems } (9) \text{ and } (11)
\end{aligned} \tag{16}$$

The previous model (15) will be solved using the following heuristic algorithm:

Algorithm

0. Calculate initial vector of probabilities $y^{(0}$; set $\Lambda^{(-1} = 0$; $k = 0$; take initial $0 < v_{PT}^{w,s,(0} < g_w, w \in W, s \in S$.
1. For the probability vector $y^{(k}$ and the number of trips for public transport $v_{PT}^{w,(k}$ solve problem (16). Let $\chi_a^{(k}, \psi_i^{(k}, x_a^{\ell,k}$ the design decision variables. Also let $\Lambda^{(k} = c_x^T\chi^{(k} + c_\psi^T \Psi^{(k}$ the building costs.
2. If $\left|\Lambda^{(k} - \Lambda^{(k-1}\right| \le \varepsilon\Lambda^{(k-1}$ & $|| y^{(k} - y^{(k-1} ||\le \varepsilon'$ & $| v_{PT}^{w,s,(k} - v_{PT}^{w,s,(k-1} |\le \varepsilon'' v_{PT}^{w,s,(k-1}, \forall_w \in w$ then STOP.
3. With the solutions $v^{w,s,(k}$ and $z^{\ell,(k}, \ell \in L$ of previous problem (16), evaluate the mean travel times for public transport for scenario $s \in S$, as $\lambda_{PT}^{w,s,(k} = \left(d^T v^{w,s,(k)}\right)/v_{PT}^{w,s,(k}$ and use logit formula (1) to evaluate a new modal split: $\widehat{v}_c^{(w,s)}$ and $\widehat{v}_{PT}^{(w,s)} = g_w^s - \widehat{v}_c^{(w,s)}, w \in W, s \in S$.
4. Taking into account the number of services $\theta_a^r(v^r) = \sum_{l \in \widehat{L}} z^{l,(k)}{}_{x_a^l}, a \in A, r \in S_0$, reevaluate the failure probabilities $\widehat{p}_0^{r,(k)} = F_0^r(\theta_a^r(v^{r,(k)})), s \in S$ and compute a probability vector $\widehat{y}^{(k)}$.
5. Perform an MSA step (using, for instance, $\alpha^{(k} = 1/(k+1)$). Then, increase the iteration counter $k = k + 1$.

$$\begin{aligned}
& y^{(k+1} = y^{(k} + \alpha^{(k}(\widehat{y}^{(k} - y^{(k)}) \\
& v_{PT}^{w,s,(k+1} = v_{PT}^{w,s,(k} + \alpha^{(k}(\widehat{v}_{PT}^{w,s,(k} - v_{PT}^{w,s,(k}),\ w \in W,\ s \in S
\end{aligned} \tag{17}$$

6 Computational tests

The computational proofs have been carried out on the network reported in [3] and in [8] with 9 nodes, 15 edges and 72 origin-destination pairs. The network parameters (construction costs for nodes and links, i.e. c_i and c_a, respectively), the o-d demand matrix and the o-d costs for the alternative mode of transportation, (u_c^w), have also been taken from that references. In all computational tests a maximum of $|L| = 5$ lines has been allowed in the solution from a pool of 46 candidate lines for operation in a single undisrupted scenario $S_0 = \{0\}$ for which there are 15 disruption scenarios (one for each link in the network; $|D(0)| = 15$) The recovery of the affected lines is carried out using 230 recovery lines. Table 1 shows in column *#it* the number of iterations necessary to converge and column $|pr|$ displays the error $y^k - y^{(k-1)}$ in the last iteration. By means of the tests it is possible to analyze the influence of the service probability failure π in the reliability of the designed RTN. For higher values of π the algorithm oscillates, converging very slowly. The more reliable the system is the smaller the total costs, being these represented in the *objfun* column. Also, the probability p_0 of no disruption increases as π is smaller and the attractiveness of the public transportation system increases as the system becomes more reliable. This is illustrated in columns *PTt* and *Ct*, (total time in public and private transport, respectively) showing that the total time spent by all public transport users increases whereas the expected time spent in the competing mode decreases.

If the probability of failure is high (i.e., $\pi = 0.01$), then, the scenario with no disruptions has smaller probability than other scenarios corresponding to disruptions. If the probability π is below a given threshold, the no disruption scenario becomes the most likely situation. In our test example this seems to happen for $\pi \approx 5 \cdot 10^{-4}$. The tests also show that in this case, the topology of the designed network does not change, i.e. it is *as if* the failure scenarios would not need to be taken into account in the design of the transportation system. This is achieved when $\pi = 5 \cdot 10^{-6}$, where disruption scenarios have almost no relevance in the model.

7 Conclusions

A two-stage stochastic model has been developed for the design of rapid transit systems taking into account the rate of failures of the transportation units. Also taken into account in the design is the number of the services during a disruption, assuming that the affected lines can operate at both sides of the link out of service. The probabilities assigned to the disruption scenarios are consistent with a probability distribution model that arises as a consequence of failures in the transportation unit services. By means of the tests it is possible to analyze the influence of the service probability failure π in the reliability of the designed RTN and determine its admissible levels for which the disruptions are at an acceptable level. A heuristic solution method is examined for small to medium networks demonstrating the computational viability of the approach.

Acknowledgments This research was supported by project grants TRA2011-27791-C03-01/02 by the Spanish Ministerio de Economía y Competitividad.

References

1. Bruno G., Gendreau M., and Laporte G. (2002). A heuristic for the location of a rapid transit line. Computers and Operations Research, 29, 1-12.
2. Cacchiani V., Caprara A., Galli, L., Kroon, L., Maróti, G., and Toth, P. (2011). Railway Rolling Stock Planning: Robustness against large disruptions. Transportation Scie., 1-16.
3. Cadarso L. and Marín, A. (2012). Recoverable Robustness in Rapid Transit Network Design. 15th EWGT, Paris, 10-13 September 2012.
4. Cicerone, S., D'Angelo, G., Di Stefano, G., Frigioni, D., Navarra, A., Schachtebeck, M., Schöbel, A. (2009). Recoverable robustness in shunting and timetabling, R. Ahuja, R. Möhring, C. Zaroliagis, eds. *Robust and On-Line Large Scale Optimization* Lectures Notes in Computer Science., Vol 5868. Springer-Verlag, Berlin, 28-60.

Table 1 Outputs for different values of π

π	$\hat{objfun}$	Λ	PTt	Ct	p_0	$\|pr\|$	$\#it.$
0.01	355.26	77.5	974.97	600.12	$2.0957e-01$	$1.4311e-02$	(*)21
0.005	349.97	62.7	101.61	698.08	$3.5839e-01$	$1.4104e-02$	(*)21
0.0005	344.85	77.5	123.77	221.74	$8.7240e-01$	0.0	6
0.00005	344.48	77.5	126.30	170.11	$9.8569e-01$	$4.3368e-18$	6
0.000005	344.21	77.5	126.59	164.26	$9.9855e-01$	$1.7889e-18$	6

(*) Maximum number of iterations

5. Evans SP (1975) Derivation and analysis of some models for combining trip distribution and assignment. Transportation research 10 pp 37-57 1975
6. Laporte G., Marín Á., Mesa J.A. and Ortega F.A. (2011). Designing robust rapid transit networks with alternative routes. Journal of Advanced Transportation, 45, 5-65.
7. López-Ramos F. (2014) Conjoint design of railway lines and frequency setting under semi-congested scenarios. PhD Thesis. Departament d'Estadística i Investigació Operativa. Universitat Politècnica de Catalunya.
8. Marín Á. (2007). An extension to rapid transit network design problem. TOP, 15, 231–241.

Management System Architecture for 3D Audio Evaluation Database

Jaemin Hwang, Jeonghyuk Kim, and Sanggil Kang

1 Introduction

Recently, development of sound equipment industry and stereophonic sound system fused to a 3D multimedia has attracted an attention. The 3D audio core algorithms such as the sound source localization [1], artificial reverberation [2], source separation [3], cross-talk cancellation [4], spatially mapped GCC function [5], and sweet spot generation modeling [6] play a core role on developing stereophonic sound system. For developing the core algorithms, we need three audio sources such as raw music instrument sound [7], HRTF (Head related transfer function) [8] for determining the position of sound, RIR (Room impulse response) [9] for creating an artificial reverberation.

We describe as one audio database (DB), A binaurl room impulse response database [9] has provided audio source for evaluation. In this study, they provide the information that magnitued squared cohrence (MSC), measruement system, maximum length sequence, and measering room for analysis. In addition, they presented the evaluation method through the Speech to reverberation modulation energy ratio (SRMR)[9] and the Segment signal-to reverberation ratio (SegRR) [9]. An evaluation system is required for evaluating the core algorithms developed by researchers or scholars in order to investiage whether they are well devloped or not.

Conventional evaluation systems have [10] been carried out by comparing the 3D audio sources with filtered sources by applying Source-to-distortion-ratio (SDR), Source to-interference-ratio (SIR), and Source to-artifact -ratio (SAR). For more accurate analysis, performance mearsurement in BASS(Blind audio source separation)[11] is presented Time-varying gains allowed distortion(TI Gain), Time-invariant filters allowed distortion (TI Filt), and Time-varying filters allowed distortion(TVFilt). they provide Matlab tool box and API to users in a manual manner. Usually, the metadata of 3D audio sources and filtered sources about system, environment, and measuring position, is stored in XML format.

In these studies, the metadata stores environmental information not evaluation results. users have to learn usage of evaluation algorithm and set up additional environment. In addition, it is ineffient to use and search audio sources because audio soureces are not indexed in general. In order to solve these problems, we propose an automatic evaluation system for evaluating performance of 3D audio core algorithms. To do that, we design the architecture of 3D audio core algorithm evaluation DB enabling to automatically evaluate core algorithms using database management system (DBMS).

The remainder of this paper is organized as follows. Section 2 explains the operation of management system architecture for 3D audio core algorithm evaluation. And, we describe our proposed XML metadata scheme in Section 3. In Section 4, we explain the additional description of XML scheme. In Section 5, we show implement of our prototype system. In Section 6, we conclude our work.

2 Structure of Our Evaluation DB

Fig. 1 is the structure of our performance evaluation DB for 3D realistic stereo sound systems consisting of three parts such as UI manager, Performance evaluation DB, and Database management system. We design the DB to enable to rate the performance of the algorithms developed for 3D realistic stereo sound systems. UI manager receives audio data such as raw data and sends it to Data control interface in Performance evaluation DB.

J. Hwang (✉) • J. Kim • S. Kang
Department of Computer and Information Engineering, Inha University, 253 Younhyun-dong, Nam-gu, Incheon, South Korea
e-mail: nulpis1@gmail.com; sspwiz@inha.edu; sgkang@inha.ac.kr

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_82,

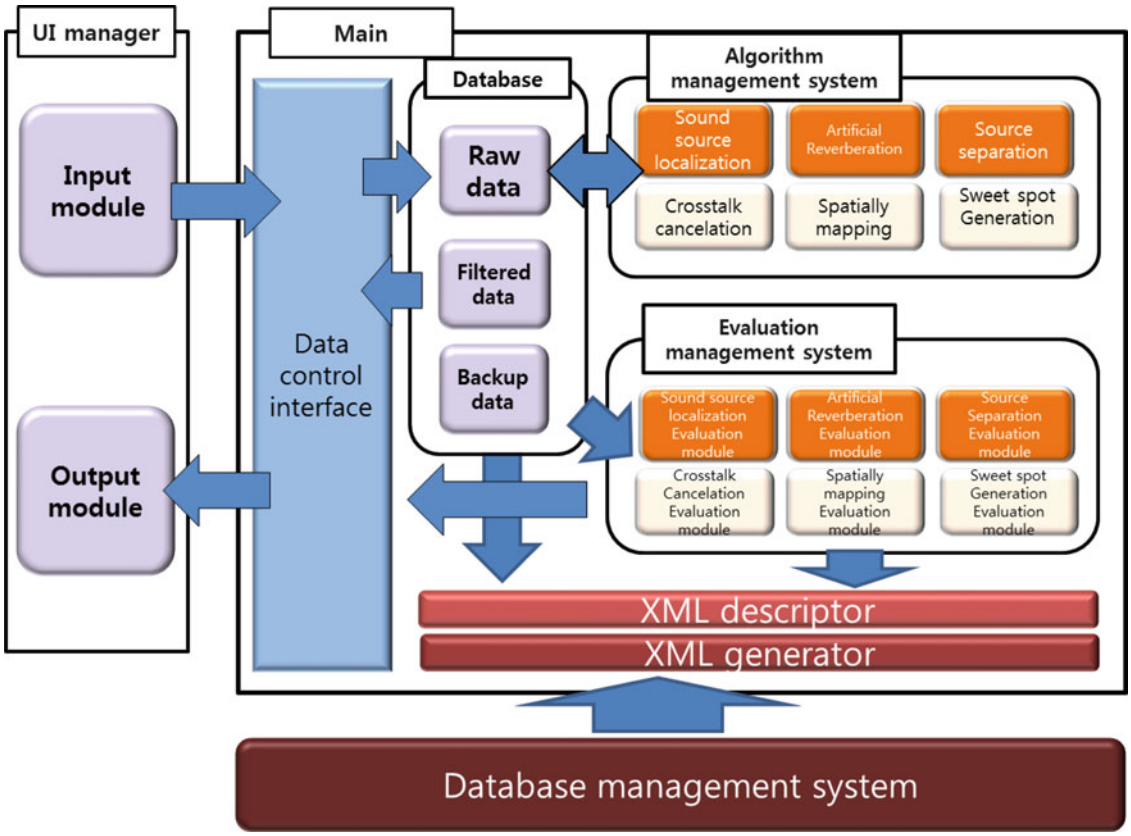


Fig. 1. Structure of our evaluation DB

An automatic 3D audio performance evaluation database system consists of three parts such as UI manager, Main, Database management system. Database module stores the audio data received through Data control interface and then it sends data to the 3D audio core algorithms system. Then, the result, i.e. filtered data, obtained from each algorithm is stored into Database module in a metadata format [12]. The filtered data is sent to the evaluation module and is stored to enable to evaluate the performance of various 3D audio algorithms. The evaluation results are coded in metadata format using XML (Extensible Markup Language) descriptor and recorded in a file format by XML generator. Not only the evaluation results but also all audio data stored in DB are coded in XML. The stored metadata is managed by Database Management System (DBMS). There are two reasons why we use XML: one is that XML is Object Management Group (OMG) standard for exchanging metadata information via Extensible Markup Language (www.omg.org). The other is XML has been widely used for storing audio metadata. So when we are storing data by XML format, it can be used widely in several studies. The evaluation results are sent to Data control interface and then presented to user through the output module in UI manger.

3 XML Scheme of Audio Source

3.1 Main Structure of XML Scheme

Fig. 2 is an example of XML scheme in which objects are described in a hierarchical structure by XML descriptor. In the example, audio data can have three objects such as acoustic, environment, and evaluation. Also, Acoustic object consists of three sub-objects such as general, listener, and source. The solid boxes mean essential object, while the dotted ones are optional objects. The acoustic and environment object defines acoustic information of the audio data,

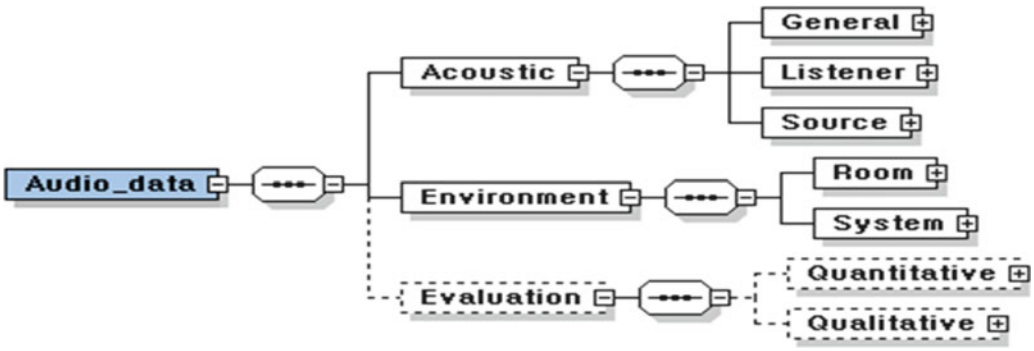


Fig. 2. Main structure of XML scheme

environment of measured audio data, respectively. Also, the evaluation object defines information related to the evaluation. Fig. 3 is an example code of audio data described with XML scheme explained as above.

The ID of each has a unique name. If the element is optional, it was shown by a dotted line. Non evaluated audio source is not essential to describe evaluation object. The list of object that can be placed in Acoustic and Environment is described in the section 4.

3.2 Evaluation

Fig. 4 is the evaluation object parameters which tells about evaluation information. The evaluation consists of two types such as quantitative evaluation and qualitative evaluation. The quantitative evaluation includes the source to distortion ratio (SDR), the source to interference ratio (SIR), and the source to artifact ratio (SAR). In the case further description is required, it is possible to record the time-invariant gains allowed distortions (TI Gain), the time-invariant filters allowed distortions (TI Filt), and the time-varying filters allowed distortions (TV Filt) [11]. In the qualitative evaluation parameter, evaluators are placed separately in layman and expert. The expert records the name of evaluator from Evaluator1 to Evaluator10 and the layman has the name of evaluators from Evaluator11 to Evaluator20 in order to avoid the duplicate names. The expert object consists of three parts such as Sound_diffusion, Sound_depth, and Sound_stage and they are optional. The layman object is similar to expert object, too.

4 Audio Source XML Schema Diagram

Fig. 5 is the general object parameters for audio source in XML schema. The general object describes the general information of the audio source. The sub-object, Title records the name of the project that has acquired the audio source, such as "Information Technology Research Center." The sub-object Date records the date when the audio source is obtained. The sub-object Type records the type of audio source such as raw audio source, RIR, or HRTF. The parameter Note describes additional information about the project.

Fig. 6 is the listener object parameters for listeners' information such as ID, position of listeners, and description about listeners. ID is listener's unique subject name. The sub-object Position is coordinates of the listener's location which is recorded in three-dimensional by dividing into X, Y, and Z. For HRTF, it records the position of the listener HRTF sound. However, it records the location of the acquired audio source for other audio sources. For Description object, we can describe listeners' characteristics such as hair style or singularities of the listener.

Fig. 7 is the source object parameters about acoustical information from the evaluation DB. The object ID records the name of an audio source. The object URL records the location of the audio source in WWW. The object Position records the position of audio source. The object Directivity records a omnidirectional pattern on the basis of the audio source. Its position should be described in the O-format impulse response [13]. The object Instrument records the type of instruments used to measure audio source.

Fig. 8 is the room object parameters which describes reverberation in a particular environment. The sub-object ID records the name of the room or its identifier. The Perceptual_parameter records reverberation, presence, and envelopment of the room. The sub-object Reverb describes the frequency-dependent reverberation time. The sub-object Perceptual and Reverb_time are optional, except ID object.

Fig. 9 is the system object parameters which is the name of the equipment that was used to record the audio source.

```
- <AudioData>
  - <Acoustic>
    - <General>
        <Title>Summer2014session</Title>
        <Date>20140304</Date>
        <Type>HRTF</Type>
        <Notes>No photo</Notes>
      </General>
    - <Listner>
        <ID>Subject02</ID>
        <Position />
        <Description>short hair</Description>
      </Listner>
    - <Source>
        <ID>ITRC02</ID>
        <URL>Http://inha.ac.kr/IMLdatabase/soure3.cda</URL>
      </Source>
    </Acoustic>
  - <Enviroment>
    - <Room>
        <ID>Anechoic room</ID>
        <PerceptualParameters />
        <ReverbTime />
      </Room>
    - <System>
      - <Software>
          <Name>Max/MSP</Name>
          <Version>1.0</Version>
        </Software>
      - <Microphone>
          <Type>Sennheiser</Type>
        </Microphone>
        <Amp>Yamaha</Amp>
      - <Speaker>
          <Speaker_type>Tannoy System</Speaker_type>
        </Speaker>
      </System>
    </Enviroment>
  - <Evaluation>
    - <Quantitative_Evaluation>
      - <Source1>
          <SDR>2.31</SDR>
          <SIR>23.48</SIR>
          <SAR>5.64</SAR>
        </Source1>
      </Quantitative_Evaluation>
    - <Qualitative_Evaluation>
      - <Expert>
        - <Evaluator_1>
            <Sound_diffusion>3.9</Sound_diffusion>
            <Sound_depth>4.05</Sound_depth>
            <Sound_stage>4.1</Sound_stage>
          </Evaluator_1>
          <Evaluator_2 />
        </Expert>
        <Layman />
      </Qualitative_Evaluation>
    </Evaluation>
</AudioData>
```

Fig. 3 Example of XML scheme

It includes the type of software, microphone, amp, and speaker. The sub-object Distance records the distance of audio source. The sub-objects Measuring_type and Signal_type record the type of measurement and signal such as blocked ear and sine sweep, respectively.

5 Implementation

Fig. 10 shows the prototype of our evaluation system implemented for the source separation. The specifications of the prototype are as follows: CPU 3.4GHz i7, 8 GB RAM, C#, C, and Oracle Database. As seen in the figure, our prototype consists of four parts such Metadata displayer, Source selection, Quantitative evaluation, and Qualitative evaluation.

Metadata displayer placed on the bottom part in the evaluation system shows metadata information about audio source from XML file. For example, IRC02 means the subject's name, the summer session is a title of session and so on. An audio source stored in the DB can be selected in Metadata displayer as seen in the figure. We also provide the selection function of raw source, truth value of the source, and output value of the source are selected in Source selection which is placed on the left part in the evaluation system. Here, the truth value of the source is true output of source separation measured in university's anechoic room. The output value of the source is the output value obtained from the simulated source separation algorithm. By comparing the values of the output and the truth, we can evaluate the performance of the simulated source separation in a quantitative estimation, as seen in the right part. SDR, SAR, and SIR of the truth value are -2.76, -8.02, and 8.71 respectively. And, SDR, SAR, and SIR of the output value are 2.31, 5.64, and 11.48 respectively. As a reference, we provide a qualitative evaluation for sound diffusion, sound depth, and sound stage of truth value which are -2, 7, and 9 respectively.

6 Conclusion

In this paper we designed the management system architecture for evaluating performance of 3D audio core algorithms using a DBMS which enables to automatically evaluate core algorithms. Also, we developed new metadata scheme for evaluation system, which contributes to 3D audio core algorithm evaluation in the sense of enhance convenience, by integrating the metadata format and applying automatic system. Because of this, we expect that evaluation DB will contribute to the research fields about 3D audio core algorithms by providing our evaluation system to the corresponding researchers.

Acknowledgment "This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2013-H0301-13-4005) supervised by the NIPA(National IT Industry Promotion Agency)"

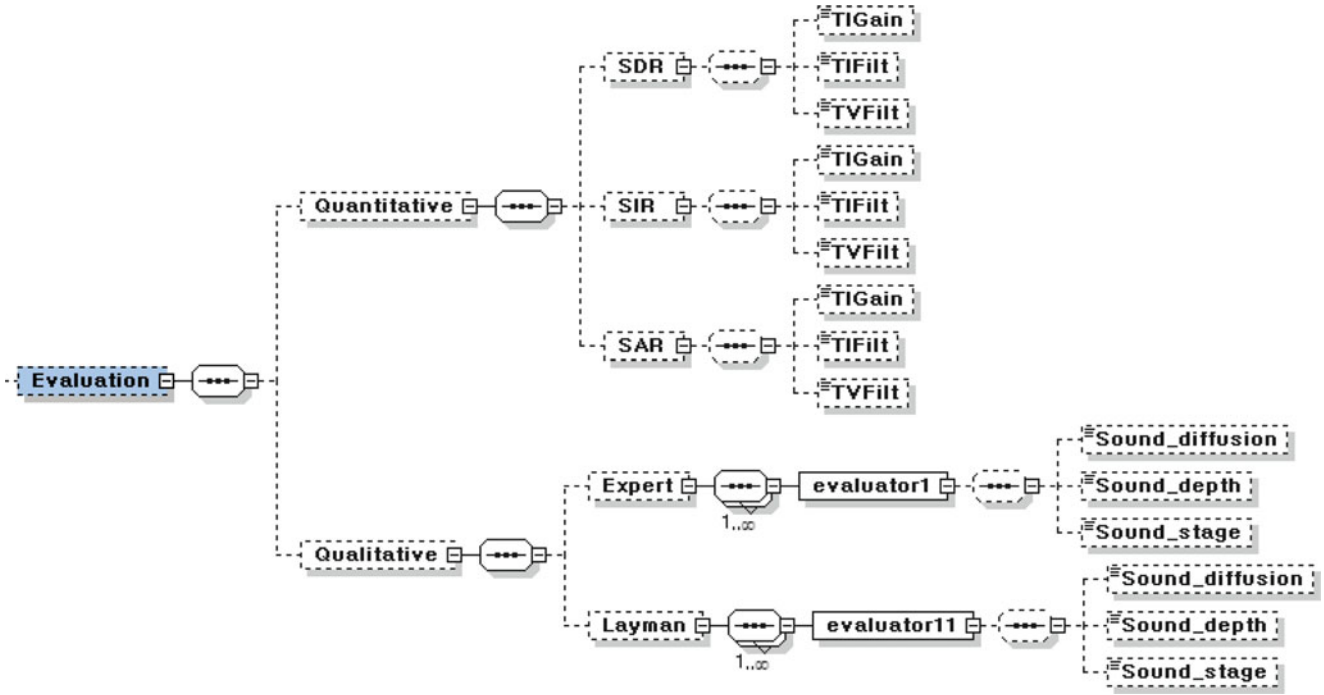


Fig. 4 Evaluation object parameters

Fig. 5 General object parameters

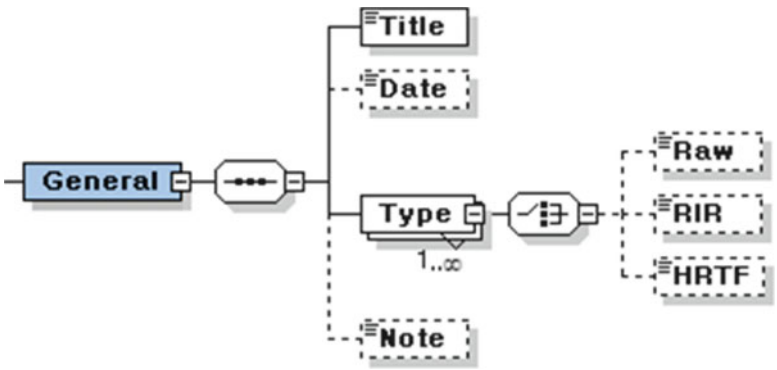
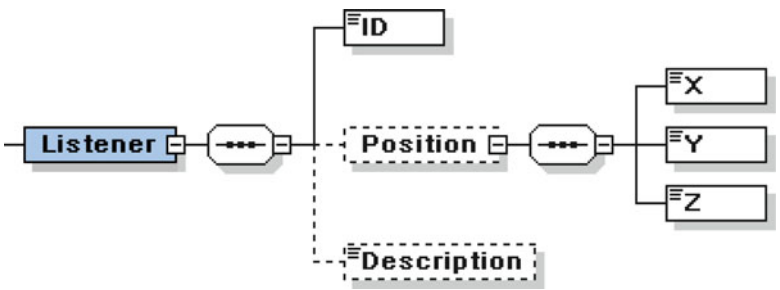


Fig. 6 Listener object parameters



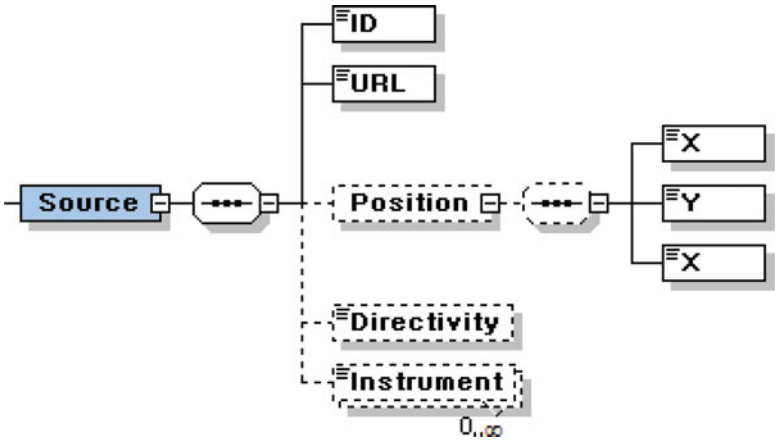


Fig. 7 Source object parameters

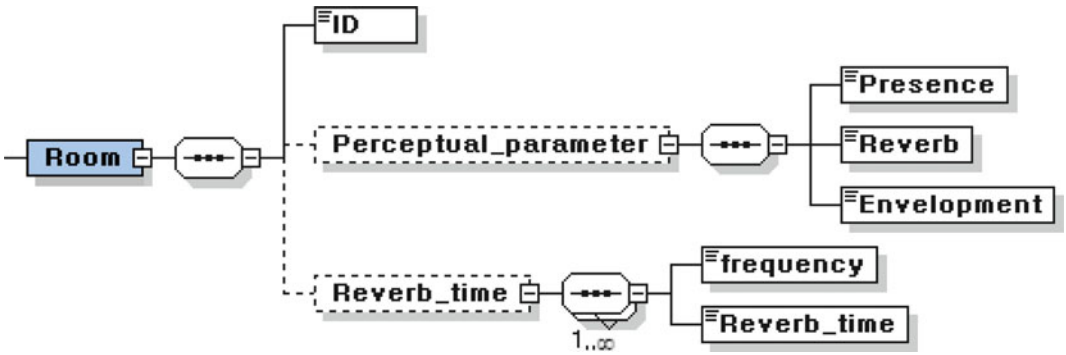


Fig. 8 Room object parameters

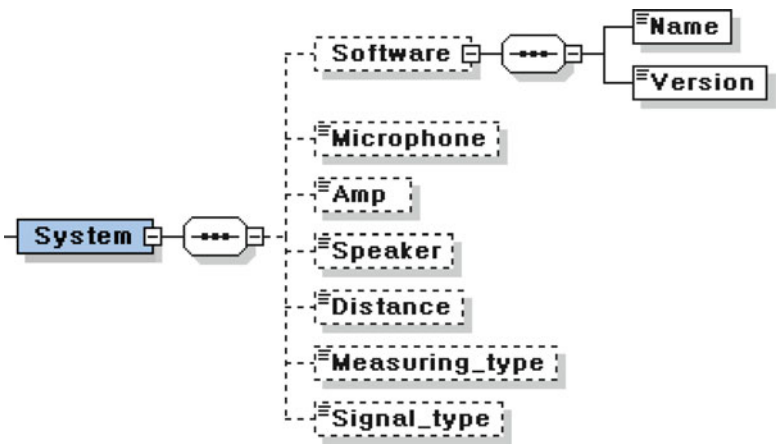


Fig. 9 System object parameters

Fig. 10 Prototype of evaluation system

References

[1] Abhijit Jukjarni, H, Steven Colburn, "Role of spectral detail in sound-source localization", Hearing Research Center and Department of Biomedical Engneering, Boston University, Vol 396, pp. 747-749(1998)

[2] Jean-Marc Jot, "An Analysis/sybthesis approach to real-time artificial reverberation", Studer Digitec S.A, France Telecom Paris, pp. 221-224(1992)

[3] Adel Belouchrni,Karim Abed-Meraim,Jean_francois Cardoso, Eric Moulines, "A Blind Source Separation Technique Using Second-Order Statistics", IEEE, Vol45, pp. 434-443(1997)

[4] Siijeong Lee, Gabken Choi, SoonHyob Kim, "A method of the cross-talk cancellation for an sound reproduction of 5.1 channel speaker system", The Institute of Electronics and Information Engineers, Vol42, pp. 159-166(2005)

[5] Byoungho Kwon, Youngjin Park, Youn-sik Park, "Mutiple sound sources localization using the spatially mapped GCC function", ICROS-SICE International Joint Conference,pp. 1773-1776, (2009)

[6] John Rons, Philip Nelson, Boaz Rafaely, and Takashi Takeuchi, "Sweet spot size of virtual acoustic imaging systems at asymmetric listener locaations", Institue of Sound and Vibrations Resrearch,pp. 1992-2002,(2002)

[7] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, Ryuichi, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database", National Institute of Advanced Industrial Science and Technology (2003)

[8] V.R. Algazi, R.O Duda, D.M Thompson, C. Avendano "THE CIPIC HRTF DATABASE", Creative Advanced Technology Center (2001)

[9] Marco Jeub, Magnus Schafer, Peter Vary, "A binaural room impulse response database for the evaluation of derveration algorithms", Instituete of Communication Systems and Data Processing, IEEE, (2009)

[10] Remi Gribonval, Laurent Benaroya, Emmaual Vincent, Cedric Fevotte, "Proposal for performance measurement in source separation", Symp on Independent Componet Anal and Blind Signal Separation, '',pp. 763-768(2003)

[11] Emmanuel Vincent, Remi Gribonval, Cedric Fevotte, "Performance Measurement in Blind Audio Source Separation", IEEE Transaction on audio and language processing, Vol14, pp. 1462-1469,(2006)

[12] Guillaume Potard, Ian Burnett, "An XML-base 3D audio scene metadata scheme", University of Wollongong, (2004)

[13] D.G. Malham, "3-D sound spatialization using Ambisonics techniques", Computer Music J., vol. 10, no.4, pp. 58-70, Winter 1995.

A Generic Metamodel for Context-Aware Applications

Imen Jaouadi, Raoudha Ben Djemaa, and Hanene Ben Abdallah

1 Introduction

Nowadays new needs in information systems have appeared. The user of application wishes to have information whenever and wherever he is located. This has prompted developers to integrate mobile devices into their applications, creating new information systems called pervasive or ubiquitous. In such systems, an adaptation of the application to a set of parameters (type of terminal, connection status, etc.) must be provided to ensure a comfortable use. All these parameters are a particular context element. In different contexts, users access the same data and the same services but can receive different responses. These systems, said context-sensitive, are able to, on the one hand, provide the personalized and relevant information and, on the other hand, self adapt to the variation of the conditions' execution descended from ubiquitous computing.

The design and development of context aware applications is particularly complex. Context acquisition is not an easy process. Context information which can be acquired from heterogeneous and distributed sources (sensors, files, applications) may be dynamic and may require an additional interpretation in order to be meaningful for an application. So, to facilitate the development of such applications, it is necessary to provide a metamodel of a generic context that is dynamic, manageable by different applications.

The rest of the paper will be structured as follow: the Section 2 will present context metamodels proposed in literature. In Section 3, we discuss theses metamodels. Then, we propose our context metamodel in Section 4. Section 5 illustrates its usage through case study. Finally, we conclude and outline our future works.

I. Jaouadi (✉) • H.B. Abdallah
MIRACL, FSEG, Sfax, Tunisia
e-mail: Jaouadi.Imen@fsegs.rnu.tn; Hanene.BenAbdallah@fsegs.rnu.tn

R.B. Djemaa
MIRACL, ISIMS, Sfax, Tunisia
e-mail: Raoudha.Bendjemaa@isimsf.rnu.tn

2 State of Art

A context metamodel is defined as the semantics of the key concepts that can be used to define the context. This section includes an overview of the main metamodels of contexts as proposed by researchers in the literature. A challenge, in the context modeling, is to identify what are the concepts that must be considered and how they are connected. We examine these metamodels based on the following characteristics: <entity/property/association>. We use "-" to denote a criteria that was not supported by a metamodel.

Henricksen et al. in [6] propose a graphical modeling notation called Context Modeling Language (CML), conceived as an extension to the Object Role modeling. In their work, the formality of models is considered, diverse context sources are addressed, validity and quality of context entities are provided and dependencies on context types are regarded (Table 1). The approach is slightly hindered by the absence of a context modeling editor. Moreover, CML isn't widely used by developers.

Farias et al. propose in [4] a context metamodel formally described using the specification Meta Object Facility (MOF) [8] to allow a precise syntax and an abstract representation common to all the models that are developed. The authors provide a metamodel independent of the application domain and mainly based on the work of Henricksen et al. [6] (Table 2). However, it does not provide a height level abstraction for context elements to express conceptual characteristics. For instance, it does not provide concepts such as task, focus, quality, etc.

The UML based Context Modeling Profile (CMP) as defined by Simons et al. [9] allows modeling the context for mobile distributed systems. UML stereotypes have been defined for modeling the context domain and Object Constraint Language (OCL) constraints are enforced to ensure

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_83,

Table 1. Approach proposed by Henricksen [6]

Approach	Entity type	Property of entity	Association type	Property of association
Henricksen	Generic	Generic	• Static • Dynamic (derived, profiled, Sensed)	• Multiplicity (simple, collection, alternative) • Temporal Constraint • Dependency constraint • Quality

Table 2. Approach proposed by Farias [4]

Approach	Entity type	Property of entity	Association type	Property of association
Farias	Generic	history of values	• Static • Dynamic (profiled, sensed, derived)	• Multiplicity (simple, collection, alternative) • Temporal Constraint (relative interval, fixed interval) • Dependency constraint

Table 3. Approach proposed by SIMONS [9]

Approach	Entity type	Property of entity	Association type	Property of association
Simons	Generic		• sensed, • Derive • User provided	• Access association: (owner, restricted, group, all) • Validity : (permanent, infrequent, frequent, volatile) • Derivation rules

the validity of the models. For a further definition of the types of information, the authors has added modeling access rights of the context elements and the degree of validity of context information (Table 3). However, Simons et al. are only interested for the context of the user. In addition, he proposed a profile that is not generic and that responds to the needs of systems meeting case study. The approach benefits from the use of UML, since the CMP can be used in various UML tools. Despite that fact, these tools do not provide a standard way to model stereotypes and enforce constraints.

In [1], the authors propose an approach MDD (Model Driven Development) to model context-aware applications independent of the platform (Table 4). However, it is focused on the contextual elements that allow the collection of context information and the identification of the context states that are relevant to the adaptation of a given application. Thus, it is not based on another work. It proposes a new model and not an extension of existing models. For example, it did not provide concepts such as temporal aspects of context [4], [6], and access rights [9].

Table 4. Approach proposed by Ayed [1]

Approach	Entity type	Property of entity	Association type	Property of association
Ayed	Generic	• collection process • **q**uality • **c**ontext **s**tate <context type, operator, value >	• collection • alternative	-

Vieira et al. [10] present a domain-independent context metamodel, which guides a context modeling in different applications. This metamodel has added new concepts such as Focus and the level of relevance that are very important to build more systems that are adaptive, usable and friendly (Table 5). However, this model is not including concepts already proposed in the literature such as access rights [9], quality of context information [1], [6]. In addition, this metamodel does not conform to MOF.

In [7], the authors have proposed a metamodel of context which is based mainly on the definition of Dey [3]. A further definition of information types, the quality, the validity and the right access of information are provided in this metamodel (Table 6). However, the proposed model is generic but is not completed. It has not represented equally important concepts to build more adaptive systems as the focus and the types of association which are defined by the work discussed above.

In 2009, Hachani et al. [5] proposed a context metamodel which respond only to the needs related to the language and interaction device. Thus, in this metamodel, the authors define only the context information types and model the context properties with a generic manner (Table 7).

3 Synthesis

Focusing on stat of art (Section 2), we can conclude that context metamodel must consist of one or more entities that represent context elements that are considered relevant to the interaction between the user and the application. In the literature, all authors have presented in their models the concept of entity (Entity), but they used different terminology as *ContextItem* in [9]. Hachani et al. [5] have used the concept *ElementContext* to describe the entity. Each entity must be described by properties (*Property*). Most metamodels proposed in the literature have considered this concept in their metamodels, but some of them have presented a generic property [5, 6] and others have considered entities and properties in their context metamodel. The entities can be atomic or composite. For example, the position can be characterized by three properties width, height and length. Therefore, several authors [5], [7], [10] have distinguished between two types of entities in their metamodels: atomic or composite. Then, the context is

Table 5. Approach proposed by Vieira [10]

Approach	Entity type	Property of entity	Association type	Property of association
Vieira	• Who (identity) • What (activity) • Where (location) • When (time) Why (motivation)	Focus <agent, task>	Generic	• Acquisition type: (sensed, profiled, user defined, derived, queried) • Validity (permanent, infrequent, frequent, volatile) • Transformation process • Relevance Constraint < conditions, actions>

Table 6. Approach proposed by Morfeo project [7]

Approach	Entity type	Property of entity	Association type	Property of association
Morfeo project	• Atom • composite	• source • quality • Validity • measurement unit • timestamp • Dependency constraints	-	-

Table 7. Approach proposed by Hachani [5]

Approach	Entity type	Property of entity	Association type	Property of association
Hachani	• device/ user/ enviroment • Atomic/ composite	Generic	-	-

dependent of the application. Indeed, information can be considered as part of the context in one area and not in another. In literature, several metamodels are restricted to narrow classes of context. In particular, there are authors who represent only the sensed context information and its derivation [1]. Thus, there are authors who proposed metamodels that respond only to specific needs. In [9], Simons et al. have proposed a metamodel that only responds to the meeting system requirements. Hachani et al. [5] have proposed a metamodel that meets only to language and interaction device. Therefore, a generic context metamodel is necessary to captures various types of context information and to be used by different applications in different fields. For this reason, we suggest to construct a generic context metamodel formed by a generic entities characterized by generic properties.

In addition, a context metamodel can be used on different platforms using different technologies. For this, we must have a context metamodel compliant to MOF (Meta Object Facility) to ensure the coherence between the different representations of context used by applications. According to the study that we did, several authors have proposed context metamodels not conform to MOF [1], [6], [10].

Moreover, each model must be formed by associations that connect entities together. Since, a generic context metamodel must be able to express the concept of associations. However, in the literature there are authors who didn't represent any association in their metamodels [5]. The context information can be characterized as static or dynamic. The static context information describes aspects of a pervasive system that are invariant; for example, date of birth. Dynamic information describes the information that changes over time. Therefore, the association links between entities can be of two types either static or dynamic. Most studies have distinguished between these two types of associations in their metamodels [4], [6]. Others have represented the dynamic associations describing only the validity period of association [7], [9]. Several distinguished the subtypes of dynamic information. In [4], [6], [10], the authors have distinguished three dynamic associations' classes: sensed, profiled and derived that represent the source of context information. Other authors have described a generic information source [7]. In fact, a global context metamodel must represent all these types (profiled, user defined, sensed and derived) as subclasses of the dynamic association representing the source of context information.

In addition, each association must be described by a cardinality that represents the occurrences number of the entity participation. Based on the study of metamodels proposed in the literature (Section 2), we find that the authors [1], [4], [6] agreed on three types of the association-end: simple, collection and alternative. We suggest to representing these three types of multiplicity in our context metamodel to describe the multiplicity of the association.

We also conclude that all proposed approaches which represent derived information of context also describe either constraints or dependencies rules on derived associations [7], [10], [19], [21]. The derived association is dependent on one or more associations which describe how the information is obtained from one or more other pieces of information. For example, the activity of a person can be inferred from the location of a person and the stories of his previous activities in the past.

In pervasive systems, a context can be dynamically changed if other information change over time i.e a change in an association may cause a change in other associations. However, the dependence of associations can exist

independently of derived associations. This concept is important in context-aware systems to provide updated and relevant information. But, it is proposed in the literature only by Henricksen et al. in [6]. For this reason, we need to model in our model the dependence on the dynamic associations.

Then, derived information can be inferred from the stories of other context elements. For example, the activity of a person can be inferred from the location of a person and the stories of his previous activities in the past. Therefore, a context metamodel should be able to represent the stories of context element. In the literature, only Farias et al. [4] proposed a solution in their metamodel to represent this concept. In our metamodel, we will represent the stories of properties values. We will also specifying for each value a time at which the context information is acquired (Timestamp).

From the study of the proposed context metamodel, we also find that most authors have presented the temporal aspect of context. Some are presented by defining temporal constraints of associations [4], [6]. These constraints indicate a valid interval of time for the use of relevant contextual information. Other authors indicating the validity period of contextual information value [7], [9], [10]. They distinguished four types of validity period of information: permanent, frequent, infrequent and volatile. Only in [4], the authors have distinguished two types of temporal constraints in their metamodel: Temporal relative Interval and Temporal Fixed Interval. However, the authors in [4] have associated the temporal constraints at the ends of associations (AssociationEnd), whereas the temporal aspect represents the entire dynamic association and not just the cardinality. Indeed, in our metamodel we choose temporal constraints to represent the temporal aspect of contextual information. Each dynamic association may have one or several constraints. Thus, we also represent the two types of temporal constraints that are added by Farias et al. (FixedInterval and relativeInterval).

Then, it is essential that applications have a context-sensitive means by which to judge the reliability of information. For this reason, we need to incorporate certain measures of sensed information quality in our context metamodel. Parameters types are dependent on the nature of the association. For example, the quality of the user's location information can be characterized by its accuracy as measured by the standard error of the location system. In our metamodel, sensed association may be annotated by one or more quality parameters. The quality of contextual information is added by some authors in the literature [1], [6], [7].

The context-user can contain personal information; hence privacy issues have to be regarded. So, a context metamodel should allow modeling the access rights to contextual elements. This concept is added only by Simons et al. [9] where he presented four access types to contextual information: owner, restricted, and all groups. These types must be represented in our metamodel.

Modeling the focus of context is very important in the context-sensitive applications. It determines which primitive contexts are being considered when the current context dynamically occurs. However, this concept is indicated only in [10]. In our metamodel we will represent the context composed of a set of focus. Each Focus is composed of a set of rules.

4 Our context metamodel

Based on the requirements mentioned in Section 3, we propose a metamodel compliant to MOF shown in Fig. 1.

A context- aware application is formed by one or more focus *(Focus*). It describe by an agent (*AgentName*) who interact with application and the task to execute (*TaskDescription*). A Focus is defined by a set of rules *(Rule)*. A Rule consists of a set of conditions *(Condition)* and actions *(Action)*. In addition, a context- aware application composed of several context primitives that represent super class of Contextual Entity *(ContextualEntity),* contextual property *(ContextualProperty)* and contextual association *(ContextualAssociation)*. ContextualEntity corresponds to physical or conceptual objects from which contextual information is captured such as: person, device and places. A contextualEntity is characterized by at least one ContextualProperty. Each property is described by a dataType. The reflexive association (*isComposedBy*) represents an attribute that can be atomic or composed of other properties. The class (*HistoryProperty*) represents the historical values of property that can assume during its lifetime. The attribute (*Timestamp*) indicates the time during which the value has been stored. The ContextualAssociation is used to define the relationships between entites and properties and the relationships among entities themselves. Each association is described by a (*name*) and type of access *(access)* that can be:

Owner, Restricted, Group or All. *Owner* is used to model the elements of the private context such as the information of the credit card. The associations applied with a restricted access type indicate a user-dependent access. *Group* attribute denotes the access to members of a group. *All* attribute indicates an unrestricted access. The reflexive relationship (*dependsOn*) indicates that an association may be dependent on one or more associations. The class (*AssociationEnd)* represents the cardinality of information. It has three types. An association is *simple* if each entity does not participate more than once as an owner of the association, such as the name of a person. The *collection* association can represents an entity that may be simultaneously associated with multiple attribute values and/or other entities; for example one person can work with several others. The alternative

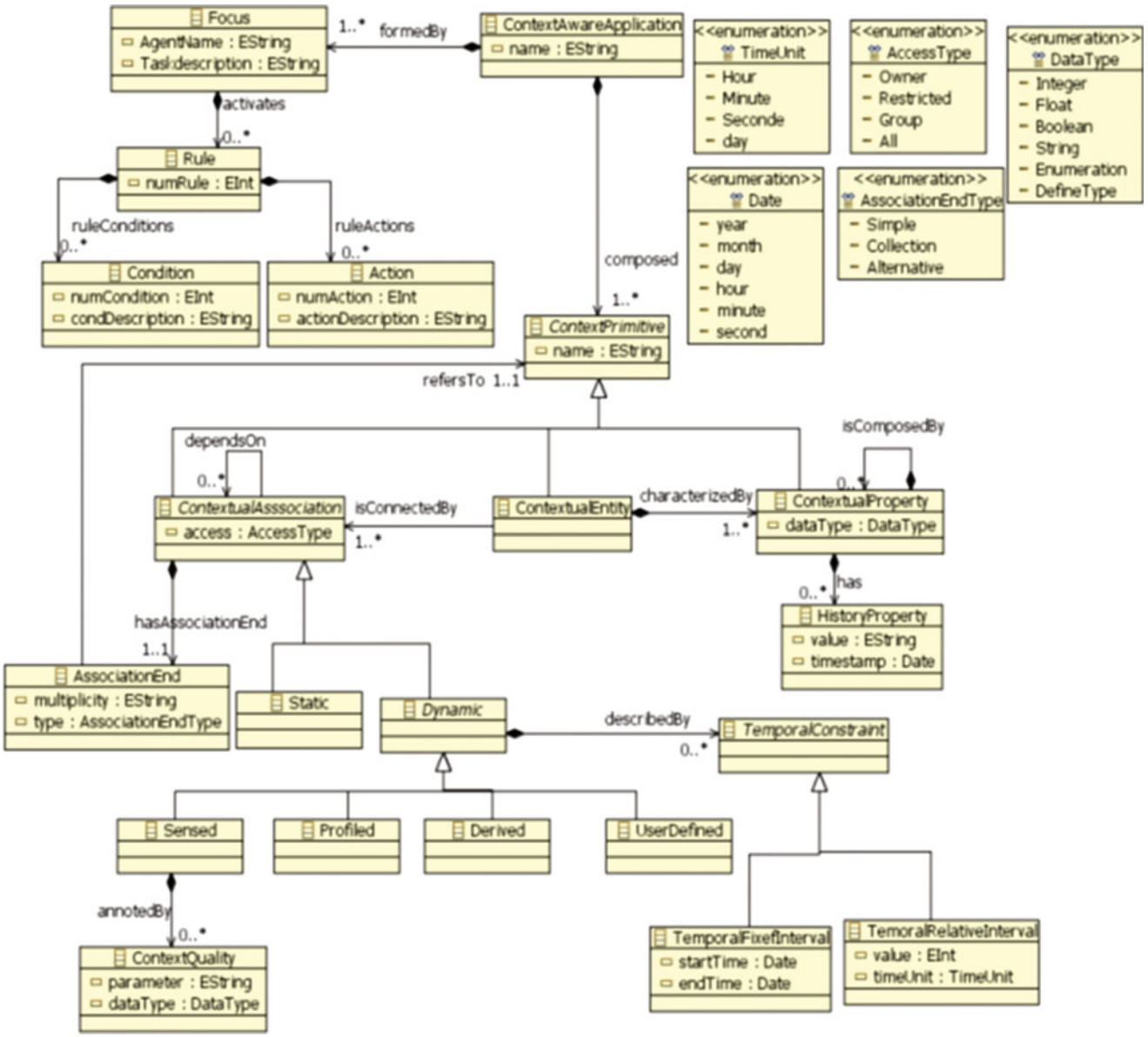


Fig. 1. Our context metamodel designed by eclipse EMF editor

association which indicates a collection of mutually exclusive values for example; a channel requires only one device associated therewith. Each (*associationEnd)* refers to only one *ContextPrimitive*. The associations are classified into two groups: *static* and *Dynamic*. *Static* associations are relationships that remain fixed throughout the life of the entity that has such as date of birth. *Dynamic* are all associations that are not static. They are classified into four types: *a sensed association represents* information for entities obtained through sensors. *Derived* association depends upon one or more associations. *Profiled association* represents information provided by application users by means configurable parameters. *UserDefined* association is directly informed by the agent through a dialog interface. A *Sensed* association can be described by *contextQuality* parameters. *TempralConstraint* is an abstract metaclass can be used to define temporal constraints in an association. The metaclass *RelativeInterval* is used to define a time interval based on current time. The metaClass *FixedInterval* is used to explicit define the valid interval.

5 Case study

The application of our context metamodel is illustrated using the healthcare epilepsy system has been mentioned in [2]. “Mr. Janssen is an epileptic patient and despite his medications, he still suffers from seizures. Because of his medical condition, Mr. Janssen is unemployed, homebound,

Table 8. Examples rules

Rule number	Conditions	Actions
Rule1	If value Activity = driving	send SMS "please stop the car as soon as possible you may have an epileptic seizure
Rule 2	if heart Rate > threshold	lunched alarm
Rule 3	If (caregiver statut = "on call"or caregiver statut = "emergency only")	situation of caregiver is avable
Rule 4	If distance (location caregiver coordinates, location patient coordinates) <100	situation caregiver = within range

and his situation requires constant vigilance to make sure healthcare professionals are alerted of a severe seizure.

Recently, Mr. Janssen has been provided with a tele monitoring context-aware application capable of monitoring epileptic patients and providing medical assistance moments before and during an epileptic seizure. Measuring heart rate variability and physical activity, this application predicts seizures and contacts nearby relatives or healthcare professionals automatically. In addition, Mr. Janssen can be informed moments in advance about the seizure, being able to stop ongoing activities, such as driving a car or holding a knife. The aim is to provide Mr. Janssen with both higher levels of safety and independence allowing him to function more freely in society despite his disorder".

The first main activity is to identify the focus of the application. Therefore, we must analyze the agents that interact with system and the tasks that agents could perform. We identify the roles which a person (Epileptic Patient or Caregiver) can play in this scenario. An Epileptic Patient represents the person who suffers from an epilepsy medical condition and need to notify upcoming seizure. The Caregiver represents the persons who have volunteered to assist epileptic patients having an epileptic seizure. The below table (Table 8) represent examples rules that can be enabled to Focus (Epileptic Patient receives notifications upcoming seizure).

Analyzing the application scenario, we identify entities, properties and association types necessary to model the healthcare application. Five contextual entities are identified. The epileptic patient (*EpilepticPatient*), and the caregiver (*CaregiverPatient*) agents that interact with system. Geographical location (*Geolocation*) described by the latitude, longitude, and the altitude of the person's current location. The device (*Device*) carried by the patient. The detection of epileptic seizure generates seizure alarm (*EpilepticAlarm*). These devices collect patient's biosignals (*BioSignalPatient*) in order to predict an epileptic seizure. Caregivers can set their status to (i) onCall, which specifies they are currently available to receive requests for helping patients, (ii) notOnCall, which specifies they are not available for receiving requests for help, (iii) busy, which specifies they are currently receiving requests, but are busy at the moment; (iv) emergencyOnly, which specifies they are currently available for receiving requests only on emergency situations. An epileptic patient may be also doing a potentially hazardous activity, which is captured by a Boolean attribute of the contextual property (*Activity*). The hazardous activity value is derived from the history activities recorded in historic property (*HistoryActivities*). Fig. 2 shows a context model of the healthcare epilepsy system designed by the eclipse EMF editor which ensures that our proposed model is compliant with our metamodel proposed in Fig. 1

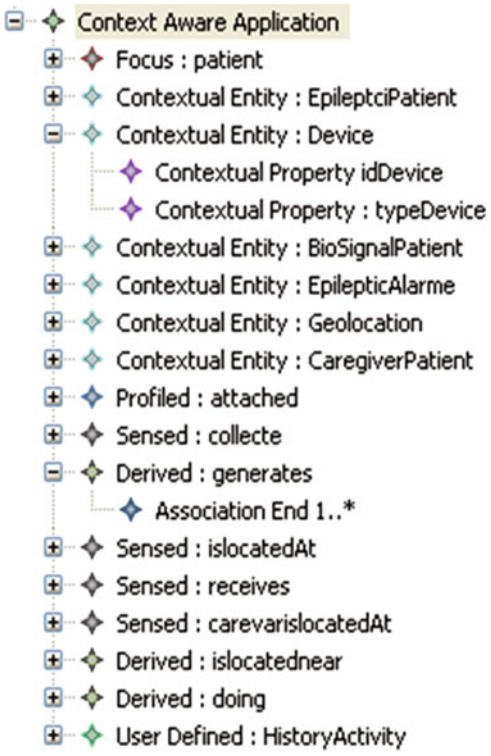


Fig. 2. Context model of the healthcare epilepsy system designed by the eclipse editor EMF

In order that our context model will be clearer to the reader, we decided to present it as a class diagram. Fig. 3 shows a context model designed by the PowerAMC.

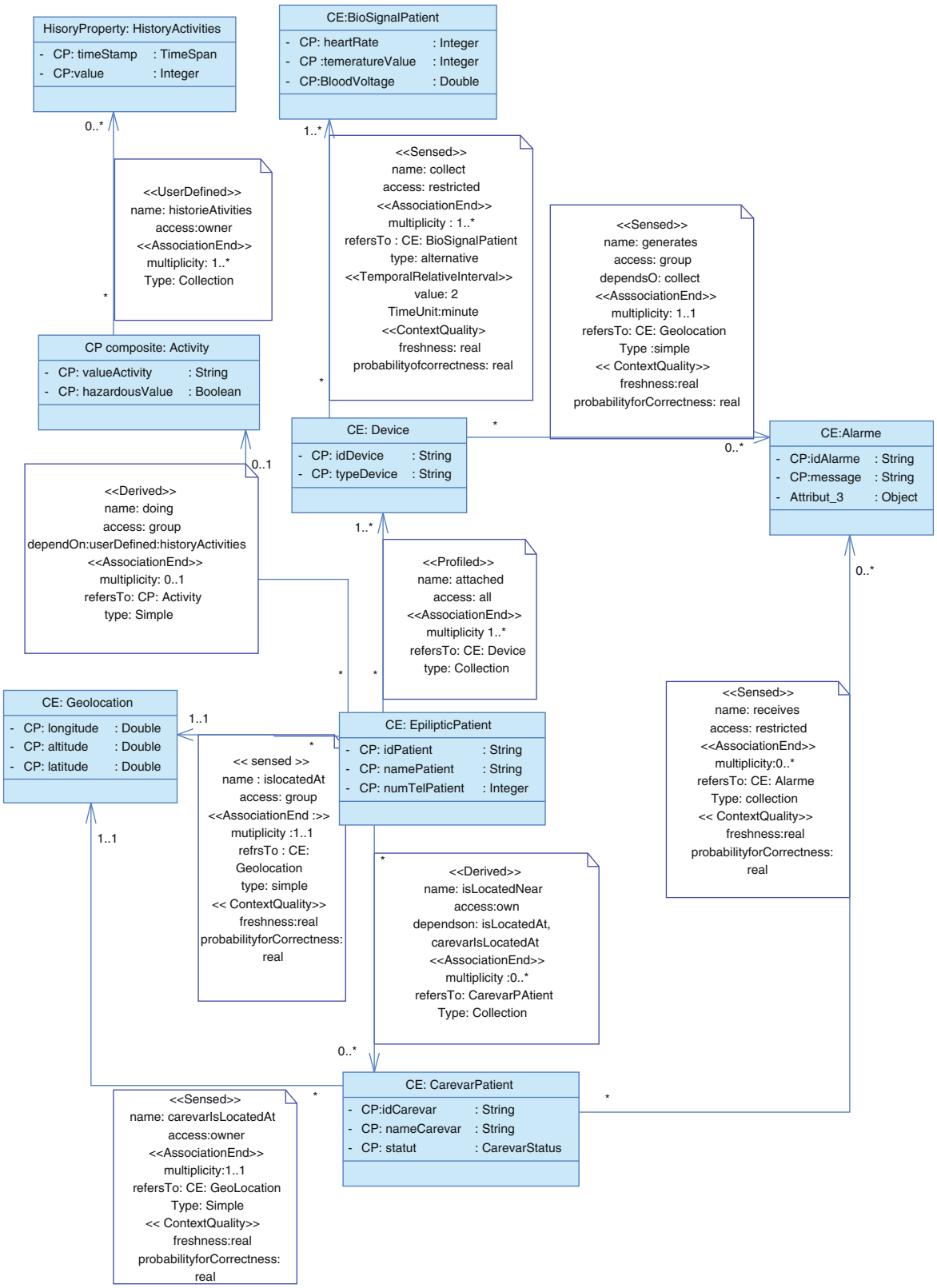


Fig. 3. Context model of the healthcare epilepsy system designed by the PowerAMC

6 Conclusions

In this paper we first presented and criticize context metamodels already proposed in the literature. Then we proposed a new generic context metamodel based on the weaknesses of the others work. It represents the historical values of each context element. In addition, our metamodel represents the temporal aspect and the quality of context to judge the reliability of information. The development of the proposed metamodel was the first step towards an approach-based on model for the development of context- aware adaptive applications. Further, we aim to refine the proposed conceptual metamodel and extend it to support the self adaptive process oriented context- aware applications by modeling the dynamic aspect of context.

References

1. Ayed D., Delanote D., Berbers Y.: MDD Approach for the Development of Context-Aware Applications. In: LNAI 4635, 6th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT'07), pp. 15-28. Springer-verlag, Heidelberg (2007)
2. Costa P.: Architectural support for context aware applications from context models to services platforms. Thesis in Computer Science, Enschede, Netherlands (2007).
3. Abowd G.D., Dey A.K., Brown P.J., Davies N., Smith M., Steggles P.: Towards a better understanding of context and context-awareness. In: HUC'99 proceedings of the 1st international symposium on Handheld and Ubiquitous Computing, pp. 304-307. Springer-Verlag, UK(1999)
4. Farias C.R.G, Leite M.M., Calvi C.Z, Pessoa R.M, Filho J.G.P.: A MOF metamodel for the development of context aware mobile applications. In: ACM Symposium on Applied Computing, pp. 947-952. ACM New York, USA (2007).
5. Hachani S., Chessa S.D., Front A.: A Generic Approach for Dynamic adaptation to the Context of CHI in French. In: 21st International Conference of the Association Francophone d'Interaction Homme-Machine, pp. 13-16. ACM New York, USA (2009).
6. Henricksen K., Indulska J., Rakotonirainy A.: Modeling Context Information in Pervasive Computing Systems. In: Pervasive'02 Proceedingsthe of the First International Conference on Pervasive Computing, pp. 167-180. Springer-Verlag, UK (2002).
7. Morfeo Project, http://forge.morfeo- project.org/wiki_en/index.php/Context_MetaModel.
8. OMG: Meta Object Facility (MOF) Core Specification. OMG Available Specification, Object Management Group (2006).
9. Simons C., Writz G.: CMP: A UML Context Modeling Profile for Mobile Distributed Systems. J. Visual Languages and Computing. 18, 420-439 (2007).
10. Vieira V., Tedesco P., Salgado A.C.: Using a Metamodel to design Structural and Behavioral Aspects in Context-sensitive Groupware. In: P 14th International Conference on Computer Supported Cooperative Work in Design, pp. 59-64. IEEE, China (2010).

Cost Effectiveness of Coverage-Guided Test-Suite Reduction for Safety-Relevant Systems

Susanne Kandl

This work has been partially funded by the ARTEMIS Joint Undertaking and the National Funding Agency of Austria for the project VeTeSS under the funding ID ARTEMIS-2011-1-295311.

1 Introduction

In [1] it says that *An essential fraction up to 50% of the development costs of a real-time computer system is devoted to ensure that the system is fit-for-purpose. In safety-critical applications that must be certified, this fraction is even higher*. One aspect of these costs is the size of the test suite (the set of test cases). Imagine a test suite consisting of 10^6 test cases and executing one test case requires in the average 1 second. Executing this test suite requires approximately 11.5 days of continuous testing (24 hours/day). Reducing this test suite by a factor of 30% reduces the test effort (regarding the time) to only 8 days for one complete test cycle. Especially for regression testing (the modified program is tested again and again with a given test suite to ensure that the modifications did not influence the intended program behavior) or for recertification of a system (renewing the certification after minor modifications) the system must be again exhaustively re-tested. The aim of test-suite reduction is to reduce or minimize the size of the test suite and thus contributing to a significant decrease in the test effort.

The test suite reduction problem was originally defined by Harrold et al. [2].

Given: A test suite TS, as set of requirements $r_1, r_2, \ldots, r_n$ that must be satisfied to provide the desired test coverage of the program, and subsets of $TS, TS_1, TS_2, \ldots, TS_n$, one associated with each of the r_i´s such that any one of the test cases t_j belonging to TS_i can be used to test r_i.

Problem: Find a representative set of test cases from TS that satisfies all r_i´s. An appropriate reduced test suite (a subset of the original test suite) must contain at least one test case from each subset TS_i. A maximum reduction is achieved by finding the smallest representative set of test cases.

The two main challenges occurring for test-suite reduction are: 1. Determining a minimal test suite is decidable but, as it is a combinatorial problem, the problem of identifying a minimal subset is NP-complete. Thus usually heuristics are applied to find a minimal test suite. 2. Reducing the test suite means removing some test cases. This has, in general, an impact on the fault-detection ability of the reduced test suite, so the quality of the reduced test suite is less than the original one.

The trade-off between the time required to execute, validate, and manage test suites, and the fault-detection capability of test suites, is central to any decision to employ test-suite reduction [3]. Many empirical studies have been carried out to determine the effect of reducing the size of a test suite on the fault-detection capability with rather contradictory results (minor loss in the fault-detection capability vs. a significant drop in the fault-detection capability). *Safety-relevant systems* have to fulfill high quality standards. Not detected faults or errors can cause fatal failures of the system with serious consequences. Therefore applying test-suite reduction is only reasonable as long as a possible reduction in the fault-detection capability is tolerable. We will define a cost function based on the costs for executing the test suite and costs for undetected faults (in the form of a penalty). This cost function will be evaluated for different test suites, defined by common coverage metrics (MCC (multiple condition coverage), MC/DC (modified condition/decision coverage), and DC (decision coverage)). We will apply this cost function to a safety-relevant use

S. Kandl (✉)
Institute of Computer Engineering, Vienna University of Technology, Vienna, Austria
e-mail: susanne@vmars.tuwien.ac.at

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_84,

case from the automotive domain in the context of ISO 26262 [4].[1]

The paper is organized as follows: First we give an overview of the state-of-the-art of test-suite reduction techniques. In the following section we describe the coverage-guided test-suite reduction. Then we introduce our cost function and describe our evaluation environment (including the case study, our settings for determining the fault-detection effectiveness, and the definition of different cost scenarios). Subsequently, we show our results for the evaluation and discuss the implication of these results. Finally, in the conclusion we give concrete recommendations for applying test-suite reduction to safety-relevant systems.

2 State-of-the-Art of Test-Suite Reduction and Related Work

First ideas for test-suite reduction were published by Leung and White [5]. The authors are using a dynamic analysis to reduce the number of test cases. The technique focuses on structural coverage. Harrold et al. [2] present a technique to select a representative set of test cases from a test suite that provides the same coverage as the entire test suite. The reduction is performed by a heuristic that selects a representative set of test cases as a subset of a test suite, but still providing the desired testing coverage with respect to data-flow testing (def-use pairs). For a given example a test-suite reduction of 33% is achieved. The approach is evaluated on a small selection of programs of relatively simple programs (the maximum number of original test cases is 80) and the average reduction ranges from 19% to 60%, 20% to 55% for implementation changes, and 0% to 40% for program enhancements. [3] gives a good overview on empirical studies of test-suite reduction. This paper examines the costs and benefits of test-suite reduction and focuses on the factors that influence these costs and benefits. The aim is to determine the trade-off between the time required to execute, validate, and manage test suites, and the fault-detection effectiveness of the test suites. Besides describing two existing studies ([6] and [7]), the authors present own data for two experiments. All experiments have in common: (1) Test-suite reduction can produce savings by reducing the size of the test-suite; those savings increase with the test-suite size. (2) Reduction in the fault-detection effectiveness increases as test-suite size increases. (3) Reduction techniques sustaining a specific coverage criterion result in reduced test suites with a higher fault-detection effectiveness than randomly reduced test suites; test suites that do not add coverage are not likely to detect additional faults.

Whereas the first two studies ([6] and [7]) indicate that the benefits of test-suite reduction outperforms the losses in fault-detection effectiveness (given by an acceptable reduction in the fault-detection capability around 5-7%), in the latter two experiments ([3]) the reduction of the fault-detection effectiveness occurred to be 50% or 100% (Siemens programs), or about 10%-20% (Space), so fault-detection effectiveness was severely compromised by test-suite reduction. [8] compare four typical test-suite reduction techniques to evaluate how effective they are in reducing the size of the test suite and how they perform regarding the execution time: Heuristic H (HH) [2], Heuristic GRE (HGRE) [9], Hybrid genetic algorithm (HGA) [10], and ILP (ILP) [11] with results for the time efficiency: HGA $\gg$ ILP $\sim$ HGRE $>$ HH. [12] presents an approach where test cases are created with a model checker and transformed in such a way that redundancy within the test suite is avoided by minimizing the length of the test cases. As test cases are not simply discarded, the impact on the fault sensitivity is minimal.

Jones and Harrold [13] introduce a test-suite reduction with the focus on the MC/DC-criterion, Staas et al. [14] discuss general considerations for the usage of code coverage metrics only as a quantitative measure, and in [15] the testing effort between decision coverage and MC/DC is compared. In [16] and [17] different code coverage metrics are compared regarding their effectiveness in finding errors, concluding that MC/DC is cost effective in relation to other criteria.

Reasons for the contradictory results regarding the decrease in the fault-detection capability of a reduced test suite may be:

1. Different algorithms for test-suite reduction were applied.
2. The programs differ in size and structure.
3. Different types of test suites are used: considering small test suites (the number of test cases is smaller than 5) we can assume also a reduced redundancy in the test suite, therefore by test-suite reduction the probability of omitting good test cases (test cases that are capable to reveal faults) is higher.
4. There are special test cases: in one sample 2 test cases were capable to detect all 15 faults of the program, in another sample 3 test cases of a test suite detected 29 out of 30 faults; if a strong test case (capable to reveal multiple faults) remains in the reduced test suite, the fault-detection effectiveness will stay the same, if a strong test case is removed by test-suite reduction the fault-detection effectiveness will drop significantly.
5. There are different types of faults: some of them are easy to detect (so there are also detected by the reduced test

[1] Please note that although the standard refers to *safety-related* systems we stick to the more popular term in automotive *safety-relevant* (also used by certification authorities, like TÜV SÜD).

suite), others are hard to detect (so most probably not detected by the reduced test suite); if a fault is detected by many test cases, the chance to be revealed after test-suite reduction is high.

Given the results of several empirical studies on the trade-off between savings by test-suite reduction and the corresponding loss in the fault-detection effectiveness, it is not possible to give a clear recommendation pro or contra test-suite reduction because employing test-suite reduction may, or may not, yield substantial losses in the fault-detection effectiveness of test suites. The presented works suggest that additional understanding of test-suite reduction is necessary, especially its potential cost-benefits and the factors that influence those cost-benefits. An analysis incorporating various factors that affect the cost effectiveness of test-suite reduction (see 1.-5.) is desirable.

In this paper we focus on the analysis of the cost effectiveness of test-suite reduction (determined by the trade-off between the testing effort and the fault-detection ability of the test suite) for *safety-relevant systems*. For non-safety-relevant systems a higher rate of undetected errors is acceptable. Safety-relevant systems have to ensure a high reliability, thus undetected errors are very expensive. We will analyze the cost effectiveness of coverage-guided test-suite reduction for a safety-relevant use case from automotive.

3 Coverage-Guided Test-Suite Reduction

Coverage metrics are a means to determine which parts of the program have been executed in the testing process. *Structural* code coverage metrics are defined on the control flow graph (CFG) of a program and can be used to evaluate which paths of the control flow graph have been covered or not. In coverage-guided test-suite reduction the resulting test suite is derived by applying a specific coverage metric. For each coverage metric we define the test suite as the minimal number of test cases achieving full coverage (regarding the coverage criterion) on the program.

Decision Coverage (DC): Decision coverage requires that every point of entry and exit in the program is invoked at least once, and every decision in the program takes all possible outcomes at least once. For decisions depending on a single condition, like `if (A==true) statement_1 else statement_2`, the two test cases for decision coverage *A=true* and *A=false* cover both branches of the decision and represent the whole input data space. For decisions depending on complex Boolean expressions (including more than one condition), like `if ((A==true) && (B==true)) statement_1 else statement_2`, a test suite for decision coverage may contain the test cases *(A=true, B=true)* and *(A=false, B=true)*. With these two test cases again both branches of the decision are covered, but the test suite does not cover the complete input data space as the test cases *(A=true, B=false)* and *(A=false, B=false)* are missing.

Modified Condition/Decision Coverage (MC/DC): MC/DC is introduced in DO-178B [18], discussed in detail in [19], and expanded with variations of the metric in [20], respectively. The metric is designed to test programs with decisions that depend on one or more conditions, like `if ((A==true) && (B==true)) statement_1 else statement_2`. For MC/DC we need a set of test cases to show that changing the value for each particular condition changes the outcome of the total decision independently from the values of the other conditions. For details please refer to [20]. The benefit of MC/DC is the linear growth of the number of test cases with an increasing number of conditions. For n conditions we need $n + 1$ test cases to achieve full MC/DC while sustaining a very high fault-detection ability. For the example above the test cases for MC/DC are *(A=false, B=true)*, *(A=true, B=false)*, and *(A=true, B=true)*.

Multiple Condition Coverage (MCC): Multiple condition coverage requires that every point of entry and exit in the program is invoked at least once, and all possible combinations of the outcomes of the conditions within each decision are taken at least once. This metric requires for n conditions 2^n test cases for programming languages with greedy evaluation (i.e., the Boolean expression is evaluated for all possible value assignments). For the example `if ((A==true) && (B==true)) statement_1 else statement_2`, the test suite achieving full MCC consists of the test cases *(A=false, B=false)*, *(A=false, B=true)*, *(A=true, B=false)*, and *(A=true, B=true)*. For a program with many decisions depending on complex Boolean expressions a huge number of test cases would be necessary for complete testing. Consider that for programming languages with short-circuit evaluation (i.e., value assignments that do no affect the outcome of the decision are not evaluated), the number of required test cases is less than 2^n [21]. For the given example the test cases for short-circuit evaluation are *(A=false, B=−)*, *(A=true, B=false)*, and *(A=true, B=true)* where "−" can be any value. So for the example the number of test cases for MCC is the same as for MC/DC. For expressions with more than 2 conditions (*A, B, C, D, etc.*) MCC requires in the majority of cases more test cases than MC/DC.

Correlation between the different test sets: Starting from an initial test suite (including randomly generated test cases supplemented by the necessary test cases to achieve full MCC, containing many redundant test cases) TS_{INI}, the test suites for the different coverage metrics relate to each other as $TS_{DC} \subseteq TS_{MCDC} \subseteq TS_{MCC} \subseteq TS_{INI}$.

4 Cost Function

The cost function we define for our evaluation depends on three different factors:

a. The costs for executing the test cases. In general, these costs are defined by the *time* needed to run the test suite containing n test cases. For simplicity, we assume the same cost factor C (e.g., the average value) for all the test cases (e.g., $C = 10$ for 10 seconds per test case). In reality the costs for the different test cases may vary. A test case consists of a specific number of test steps (iterations in the test-case execution). Short test cases are cheaper than long test cases (regarding the execution time).

b. The reward R for detected faults. Each detected fault is recompensed by a reward and minimizes the costs for testing, d denotes the number of detected faults.

c. The penalty P for undetected faults. For each undetected fault a penalty has to be considered in the cost function increasing the overall costs, u is the number of undetected faults.

Cost Function CF: $CF(n, d, u) = n \cdot C - d \cdot R + u \cdot P$

5 Test Setup

Our evaluation is carried out on a one component of a safety-relevant brake-by-wire application from an industrial partner. The program is given as C source code (generated from a Matlab Simulink model).

5.1 Integration of Faults

From the original case study we derive 150 different faulty versions (each faulty version contains one fault). The artificially introduced faults relate to variable names (e.g., original: `result = var_1/var_2`; fault: `result = var_1/var_1)`, value assignments (e.g., original: `threshold = 10`; fault: `threshold = 20)`, and operators (e.g., original: `A && B`; fault: `A  B`). For each test suite we monitor the number of detected faults and the number of undetected faults, respectively.

5.2 Test-Suite Reduction and Test Execution

Starting from a test suite generated by random test-case generation, we measured the coverage for this test suite. Although this test suite contains many redundant test cases (redundant regarding coverage), this test suite achieves less than 60% MCC on the program. Then we identified the missing test cases for MCC and extended the test suite by additional test cases to achieve full MCC. This initial test set TS_{INI} consists of around 1000 test cases including also many redundant test cases. Then we apply the coverage-guided test-suite reduction by removing all the test cases that have no impact on the coverage. By this we gain a test suite that sustains full MCC for the case study, but containing a *minimum* number of test cases for MCC. This test suite (TS_{MCC}) consists of around 100 test cases. In the next step we apply the coverage-guided test-suite reduction with respect to MC/DC, thus removing all test cases that do not affect the MC/DC-criterion. The resulting test suite TS_{MCDC} consists of 85 test cases. In a last step we reduce the test suite by the redundant test cases regarding DC. The final minimum test suite for achieving full DC TS_{DC} contains around 30 test cases.

Each faulty version is executed with the four different test suites TS_{INI}, TS_{MCC}, TS_{MCDC}, and TS_{DC}. For each test suite we monitor the number of detected faults and the number of undetected faults, respectively. Then we define different scenarios for the cost function, varying in the values of the parameters for the costs for the test execution and the penalty for undetected faults, respectively.

5.3 Cost Function

The behavior of the cost function depends significantly on the settings for the parameters C (costs for test-case execution), R (reward for detected faults), and P (penalty for undetected faults). As the concrete values for these parameters can differ strongly for different test environments, we define some exemplary settings for three different test scenarios.

Scenario 1: Assuming a non-safety-related test environment with average costs for the test-case execution, a default reward for detected faults, and a rather *low penalty* for undetected faults, we set $C = 10$, $R = 5$, and $P = 10$.

Scenario 2: Scenario 2 represents a safety-related test environment again with average costs for the test-case execution, a default reward for detected faults, and a *high penalty* for undetected faults, thus we define $C = 10$, $R = 5$, and $P = 50$.

Scenario 2a: Scenario 2a is a variant of scenario 2: A safety-relevant test environment, but this time with *high costs* for the test-case execution; for the reward and the penalty we keep the settings from scenario 2. Thus we define $C = 20$, $R = 5$, and $P = 50$.

6 Results and Discussion

6.1 Results

The results for our evaluation are shown in Table 1. In the table the number of test cases n, the number of detected faults d, the number of undetected faults u, the resulting costs for the different scenarios (CF_1, CF_2, CF_{2a}), and the execution time t (in minutes) are given for the different test suites TS_{INI}, TS_{MCC}, TS_{MCDC}, and TS_{DC}.

The cost function for the described test scenarios for the different test suites is graphically presented in Figure 1.

6.2 Discussion

We see that the costs for testing for safety-relevant systems are higher than for non-safety-relevant systems (no matter for which test suite). This is not surprising because it is a well-known fact that an increased confidence in the test result (what we aim for safety-relevant systems) needs a higher test effort and, as undetected faults cause more costs than in non-safety-relevant systems, they increase the overall costs.

We observe a fundamental decrease in the costs from the initial test suite to the reduced test suite for MCC (around 90% for the size of the test suite, and 96% for the test execution time). This can be argued by the huge overhead caused by many redundant test cases from the random test-case generation. Test-suite reduction should always be applied to an initial test suite of randomly generated test cases to gain the minimal test suite for MCC by removing the test cases that have no impact on the MCC-criterion. The resulting test suite TS_{MCC} covers the whole input data space of the program and guarantees that all *detectable* faults are identified.

Table 1 Summary of the Results

Test Suite	*n*	*d*	*u*	CF_1	CF_2	CF_{2a}	t
TS_{INI}	997	142	8	9340	9660	19630	31
TS_{MCC}	104	142	8	410	730	1770	1.1
TS_{MCDC}	85	137	13	295	815	1665	1.0
TS_{DC}	36	114	36	150	1590	1950	0.5

The evaluation shows that for a non-safety-relevant system (CF_1) the costs of TS_{DC} are the smallest compared to the other test suites. The costs of TS_{MCDC} are smaller than the costs of TS_{MCC} (around 30% less costs).

This changes significantly for a safety-relevant system (CF_2 and CF_{2a}). The costs for TS_{DC} increase (due to the higher penalty for undetected faults). The costs for TS_{MCC} and for TS_{MCDC} are very similar. Assuming lower costs for the test-case execution for a safety-relevant system (CF_2) the costs for TS_{MCC} are *less* than the costs for TS_{MCDC}, whereas assuming higher costs for the test-case execution (CF_{2a}) the costs for TS_{MCC} are a little bit *higher* than the costs for TS_{MCDC}.

Please, consider that some of the faults (8 out of 150) integrated into the case study occur to be latent faults (faults that have no impact on the behavior of the system and cannot be detected by testing).

Based on the described summary of the state-of-the-art of test-suite reduction and the provided results for the evaluation of the coverage-guided test-suite reduction in this paper, we can summarize with the following statements:

- Test-suite reduction should *always* be applied to a test suite with randomly generated test cases.
- The test-suite reduction should be guided by a coverage criterion, and *not* just by a random algorithm.

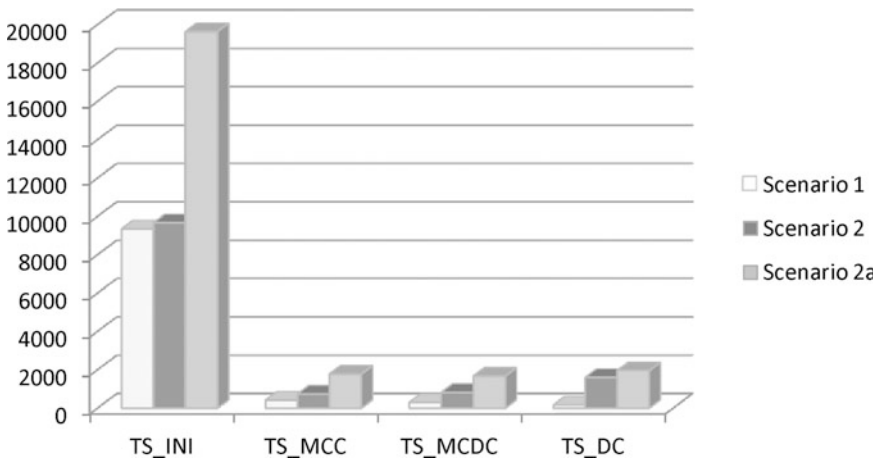


Fig. 1 Cost Function for the Different Scenarios

- The cost effectiveness of test-suite reduction depends on a variety of factors: the program under test, the type of faults, the quality of the test cases, the method of test-suite reduction, etc. and is strongly influenced by the assumed parameters for the costs for the test execution and the costs of undetected faults.
- A higher fault-detection ability requires a test suite of bigger size. There is no *minimal* test suite with a *maximum* fault-detection capability.
- For non-safety-relevant systems, test-suite reduction to reach a test suite achieving DC (decision coverage) appears to be cost effective.
- For safety-relevant systems, test-suite reduction resulting in a test set for MCC (multiple condition coverage) or MC/DC (modified condition/decision coverage) has a similar cost effectiveness. For low costs for the test execution test-suite reduction for MCC appears to be the most cost effective one. For higher costs for the test execution test-suite reduction for MC/DC seems to be a little bit more cost effective. Test-suite reduction for DC for safety-relevant systems is not at all cost effective (simply because of the high penalty for undetected faults).

7 Conclusion

Optimizing the test effort can only be achieved by optimizing the cost effectiveness. Regarding coverage-guided test-suite reduction we can conclude with the following recommendations: 1. Coverage should always be only a supplement (and not the main guard) in the assessment process of testing. 2. Test suites achieving a *weak* coverage criterion (weak in the meaning of a low fault-detection capability, like DC) appear to be cost effective for non-safety-relevant systems. 3. Cost effectiveness for test-suite reduction for safety-relevant systems is only given for test suites that sustain *strong* coverage criteria, like MCC or MC/DC. 4. Test-suite reduction is always cost effective applied to a test suite consisting of randomly generated test cases. 5. The cost effectiveness of test-suite reduction depends on the costs for test execution and the penalty for undetected faults. For a test environment with low test-execution costs, test-suite reduction for a test suite achieving MCC on the system appears to be the most cost effective strategy. For a test environment with higher test-execution costs, the test suite for MC/DC outperforms the MCC-test suite regarding its cost effectiveness.

References

1. H. Kopetz, *Real-Time Systems: Design Principles for Distributed Embedded Applications*, 2nd ed., ser. Series: Real-Time Systems Series. Springer, 2011.
2. M. J. Harrold, R. Gupta, and M. L. Soffa, "A methodology for controlling the size of a test suite," *ACM Trans. Softw. Eng. Methodol.*, vol. 2, no. 3, pp. 270–285, Jul. 1993. [Online]. Available: http://doi.acm.org/10.1145/152388.152391
3. G. Rothermel, M. J. Harrold, J. von Ronne, and C. Hong, "Empirical studies of test-suite reduction," *Journal of Software Testing, Verification, and Reliability*, vol. 12, pp. 219–249, 2002.
4. International Organization for Standardization, "ISO 26262: Road vehicles - functional safety," 2009.
5. H. Leung and L. White, "A study of regression testing," in *Proceedings of the 6th International Conference on Software Engineering*. USPDI, 1989.
6. W. E. Wong, J. R. Horgan, S. London, and A. P. Mathur, "Effect of test set minimization on fault detection effectiveness," in *Proceedings of the 17th international conference on Software engineering*, ser. ICSE '95. New York, NY, USA: ACM, 1995, pp. 41–50. [Online]. Available: http://doi.acm.org/10.1145/225014.225018
7. E. Wong, J. R. Horgan, A. P. Mathur, and A. Pasquini, "Test set size minimization and fault detection effectiveness: A case study in a space application," in *In Proceedings of the 21st Annual International Computer Software & Applications Conference*, 1997, pp. 522–528.
8. H. Zhong, L. Zhang, and H. Mei, "An experimental comparison of four test suite reduction techniques," in *Proceedings of the 28th international conference on Software engineering*, ser. ICSE '06. New York, NY, USA: ACM, 2006, pp. 636–640. [Online]. Available: http://doi.acm.org/10.1145/1134285.1134380
9. T. Chen and M. Lau, "A new heuristic for test suite reduction," *Information and Software Technology*, vol. 40, no. 56, pp. 347 – 354, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950584998000500
10. N. Mansour and K. El-Fakih, "Simulated annealing and genetic algorithms for optimal regression testing," *Journal of Software Maintenance*, vol. 11, no. 1, pp. 19–34, Jan. 1999. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1096-908X(199901/02)11:1<19::AID-SMR182>3.0.CO;2-M
11. J. Black, E. Melachrinoudis, and D. Kaeli, "Bi-criteria models for all-uses test suite reduction," in *ICSE 2004. Proceedings. 26th International Conference on Software Engineering, 2004.*, 2004, pp. 106–115.
12. G. Fraser and F. Wotawa, "Redundancy based test-suite reduction," in *Proceedings of the 10th international conference on Fundamental approaches to software engineering*, ser. FASE'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 291–305. [Online]. Available: http://dl.acm.org/citation.cfm?id=1759394.1759425
13. J. Jones and M. Harrold, "Test-suite reduction and prioritization for modified condition/decision coverage," in *Software Maintenance, 2001. Proceedings. IEEE International Conference on*, pp. 92–101.
14. M. Staats, G. Gay, M. Whalen, and M. Heimdahl, "On the danger of coverage directed test case generation," in *Proceedings of the 15th International Conference on Fundamental Approaches to Software Engineering*, ser. FASE'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 409–424. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-28872-2_28

15. Z. Szűgyi and Z. Porkoláb, "Necessary test cases for decision coverage and modified condition / decision coverage," *Department of Programming Languages and Compilers, Eötvös Loránd University*, 2002.
16. K. Kapoor and J. Bowen, "Experimental evaluation of the variation in effectiveness for DC, FPC and MC/DC test criteria," *Proceedings International Symposium on Empirical Software Engineering, ISESE 2003*, pp. 185–194, Sept.-1 Oct. 2003.
17. Y. T. Yu and M. L. Laub, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions," *Journal of Systems and Software*, vol. 79, no. Issue 5, pp. 577–590, May 2006.
18. RTCA Inc., "DO-178B: Software Considerations in Airborne Systems and Equipment Certification," Requirements and Technical Concepts for Aviation, Washington, DC, December 1992.
19. RTCA Inc., "DO-248B: Final Report for Clarification of DO-178B: Software Considerations in Ariborne Systems and Equipment Certification," Requirements and Technical Concepts for Aviation, Washington, DC, October 2001.
20. J. J. Chilenski, "An investigation of three forms of the modified condition decision coverage (MCDC) criterion," U.S.Department of Transportation, Federal Aviation Administration, DOT/FAA/AR-01/18, April 2001.
21. S. Kandl and S. Chandrashekar, "Reasonability of MC/DC for safety-relevant software implemented in programming languages with short-circuit evaluation," in *Proceedings of the 9th Workshop on Software Technologies for Future Embedded and Ubiquitous Systems (SEUS 2013)*. IEEE Proceedings, 2013, Paderborn, Deutschland.

Towards a Holistic Definition of System Engineering: Paradigm and Modeling Requirements

Hycham Aboutaleb and Bruno Monsuez

1 Introduction

Usually, the approach we follow in a project depends on how the results will be used. To optimize the design time, it is important to have a useful framework for analyzing complex systems and study their evolution. The use of such a framework requires an understanding of the boundaries of a given system, its components, its representation, and the evolution of its model and ways of representation. The complexity that emerges while designing and developing the system is usually the result of the multidimensionality of the system. To understand its behavior, a system is considered in the context of its environment, including interactions and interfaces. Indeed, the complexity of a system is often characterized, beyond the inherent complexity of components and their variety, by the complexity of the interaction network, from which emerges behaviors as intentional and unintentional, which may be harmful and difficult to predict and control.

With emergence of complex engineering systems, starting with defense systems after WW2, a new approach was required to handle the increasing complexity of these new complex engineering systems. The traditional reductionist approach " divide and conquer " was no longer valid to address this issue. It was necessary to define a holisitic approach: every part in a system is related to every other to form a coherent whole. Such an approach analyzes better the emerging behavior as well as the interdependecies. Using a holistic approach induces considering the system as a unified whole. Besides, the ability to model and design a system is limited by the capacities of the used tools for that purpose, and consequently, if any, their underlying modeling languages.

This paper introduces a new System Engineering paradigm, then it associates it with MBSE concepts. Then it identifies the main issues in system representation and modeling, and specifies the requirements a system model shall meet. Finally it specifies the requirements a modeling tool shall fulfill to be used efficiently in System Engineering.

2 System Engineering Paradigm

Unprecedented levels of complexity have emerged from contemporary engineering systems. While the organic and functional aspects remain at the core of the systems engineering method, there is an urgent need to more effectively address additional aspects that are correlated to the functional spaces of the system of interest. Dimensions to be taken into account will vary from one system to another depending on the system perimeter. These dimensions could be, for example, physical dimensions, temporal dimensions, logical dimensions, and so on. Therefore, it is necessary to clearly analyze the system-to-design context since a software system will not have the same environment as a network or a mechanical system. However, since a system function is performed by the system itself, various dimensions of such a system are related to each other. As different aspects become too complex for the mind to easily understand or operate with, different approaches are possible in order to better understand a complex system. Four concepts have been taken into account in this method: abstraction level, decomposition level, view, engineering perspective.(Figure 1)

While abstraction level allows the observer to have a holistic view of a system but in respect to different aspects, the level of decomposition partitions the problem space and allows a localized understanding of the different dimensions

H. Aboutaleb (✉)
PhD Student Computer and System Engineering Department, ENSTA ParisTech, 91120 Palaiseau, France
e-mail: hycham.abou-taleb@ensta-paristech.fr

B. Monsuez
Director of Computer and System Engineering Department, ENSTA ParisTech, 91120 Palaiseau, France

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_85,

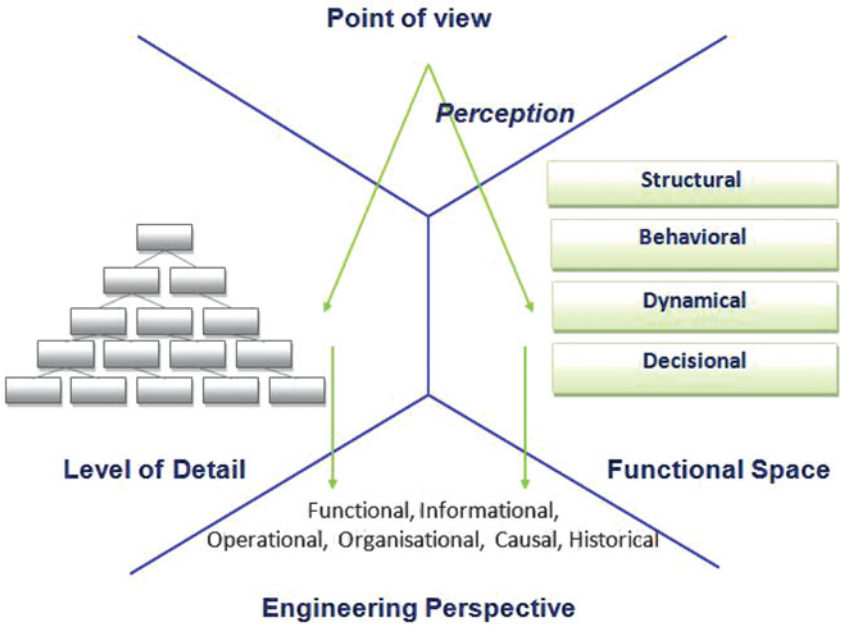


Fig. 1 Proposed Framework

of a system. As each person understands a given problem in his/her particular manner, it is of common sense that we can analyze a system from different points of view that are perfectly coherent with each other see. This approach is function-centered since it depends on the characterization of a system according to its functional spaces.

- Abstraction: view of the system that is relative to both the level of detail through decomposition and the type of information captured, but one does not need to consider these layers to understand a general phenomenon or one that is possible only in certain conditions. The functional spaces are the abstraction levels.
- Decomposition: isolate system components for a detailed analysis, given that all information of the context of the analyzed element is regarded.
- Perception: the point of view of each actor that limits or filters the available information, it allows building different models or representations of the system.
- Engineering Perspectives: what is needed to be taken into account to design the system. There are technical aspects (which are the technical processes) as well as non-technical ones.

  Non-technical engineering perspectives are [1] and [2]:
- Functional, what is to be done, the tasks.
- Behavioural, when each task is to be performed.
- Informational, which data are used, how does it flow.
- Operational, how is the work done, the tools.
- Organisational, who performs the work.
- Causal, why the process is performed, its objectives.
- Historical, what happened during the performance.

3 Model-Based System Engineering

Many Model-Based System Engineering (MBSE) methodologies emerged after the introduction of system engineering in the industry. In order to better understand the key features of different methodologies, it is important to establish a terminology for better understanding these methodologies (Figure 2) [3]:

- A *process*: sequence of tasks aiming to achieve a particular objective. Process defines what is to be done without defining how each activity has to be performed.
- A *method*: specifies *how* to perform each task.
- A *tool*: helps to accomplish of *how*. It usually supports a language that helps applying the method.
- A *methodology*: is defined as a collection of related processes, methods, and tools. In model driven context, MBSE can be defined as a collection of process, tools and methods help to harmonize system engineering discipline.
- An *environment*: consists of external conditions, systems, or factors that have an influence on systems, actors. The purpose of environment is to put in practice the use of tools and methods of a project.

Thus, the ability to model and design a system is limited by the capacities of the used tools for that purpose, and consequently, if any, their underlying modeling languages. Without a holistic approach, the cost of model construction and the effort required to integrate various system models may present critical concerns that might be reflected in the resulting system design [4]

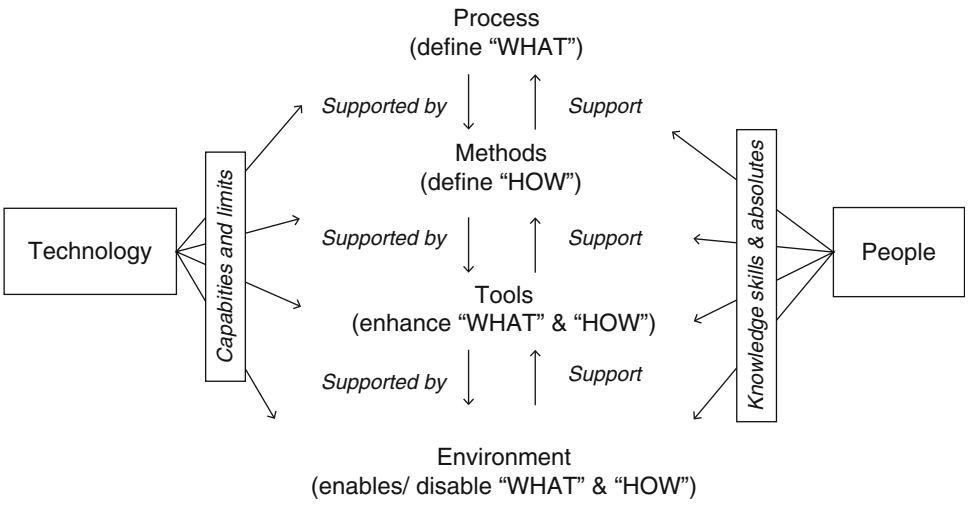


Fig. 2 System Engineering Methodology Terminology

4 System Representation and Modeling

Model-based development has been adopted more or less in development of complex systems today. To understand this trend, it is necessary to focus on the properties of complex systems to design and to the needs of the stakeholders involved in the development of these complex systems. A model has a clear purpose: to help designing the system of interest. Modelers must exclude all factors not relevant to the problem to ensure the project scope is feasible and the results timely. The value of the modeling process begins early on, in the problem definition phase. The modeling process helps focus diagnosis on the system of interest.

4.1 Modeling issues

When developing complex systems, two main problems arise:

- The need to address all the aspects of the system of interest (to design and develop). [2]
- The need to share the knowledge between people involved in the process. [5]

To match these needs, model-based system engineering is necessary. However, the need of a model-based approach induces new issues:

- trustworthiness of the model: how close the model is to the reality?
- understandability of the model: is the model perceived and understood the same way by people?
- usefulness of the model: does the model help to get the desired results?

Trustworthiness of system model Given the limited cognitive capabilities of humans, we use models of the properties of the system and its context/environment that are of relevance and interest, and disregard details considered irrelevant for the system design and development. A model is thus a deliberate simplification of reality with the objective of explaining a set of selected properties of the real system that is relevant for the purpose of its development. This model starts first with a mental process to capture relevant information, then the information captured is expressed through means to be communicated. This information is the minimum information necessary to have a satisfactory understanding of the perceived real system and environment.[6]

Understandability of system model The understandability of a model depends on how an individual perceives the model that he/she is going to use. Two people share the same *mental model* if they have similar descriptions, explanations, and predictions of the system of interest. Specifically, models allow people to similarily predict and explain the behavior of the system of interest, to recognize and remember relationships among its components and with its environment, and to construct expectations for what is likely to occur next.

Usefulness of system model To help ensure the utility of shared mental models, a distinction is often drawn among different types of mental models, normally based on their underlying content. In order to be useful, a model shall facilitate accomplishing a task and allow each individual to work effectively as a member of the team [7]. Thus, a model would be considered effective if team performance is increased. According to [8], a team performance is related to the *taskwork mental model similarity*, the *teamwork mental model similarity*, the *taskwork mental model perceived accuracy*, and the *teamwork mental model perceived accuracy*. Moreover, it shall allow engineers to reuse and share past solutions. This has an additional advantage: inexperienced engineers benefit from the work of more experienced ones and are able to work at their quality levels.

4.2 Model Complexity and Hierarchy

Since systems are inherently complex, it is necessary to fully understand a system without reaching human mind limitations by handling this complexity. This system *real* complexity is indeed reflected in the corresponding system model. In fact, the system *perceived* complexity is the model complexity. To obtain a model that is *trustworthy*, *understandable*, and *useful*, it is necessary to architecture the complexity. As it is described in [9], there is a form of organized complexity in systems.

To handle large amounts of data, it is often useful to have a classification or an order. One effective way to classify a set of elements is to use a hierarchical organization of this set of elements, introducing sometimes a new order relations among the elements. With the hierarchy, in addition to be able to handle elements together, it becomes possible to handle subsets of elements together. There are two ways how to organize hierarchically a set: grouping and encapsulation.

- It is possible to group items based on similar properties or characteristics.
- It is possible to encapsulate many elements within a single element of a higher level and then consider only the properties of this element when an analysis is performed.

Therefore, to handle complexity of the real system, its model should be the result of a simplification strategy consisting in:

- Conceptual chunking: refers to the formation of a higher-level concept that captures the essence of the problem-at-hand and reduces the complexity by omitting irrelevant detail and reducing its dimensionality [10].
- Segmentation: refers to the decomposition of a complex system into into smaller parts that can be studied in isolation, in order that the capacity limitations of the human mind are avoided.

Consequently, we can indentify two types of models hierarchies.

On one hand, there is the *generalization*, i.e. hierarchy of types.

The word *type* refers generally to a representation that gather main properties of objects that have common characteristics [11].

One type allows to group elements with common characteristics. The mechanism of subtyping induces a hierarchy: an entity type T_2, derived from type T_1has at least all the properties of an entity type T_1.

On the other hand, there is *aggregation*.

The word *aggregation* refers generally to a representation that gathers elements into another higher-level element to hide them when necessary.

The higher-level element that encapsulates its contained elements has properties that are the emerging properties at this level due to the contained elements. Other names like nested hierarchy or container hierarchy are also common Encapsulation decreases the complexity of the system model [12].

Finally, the hierarchy has an additional advantage: depending on the selected level, it is possible to observe different points of view.

4.3 Modeling Requirements

As presented in the previous section, the modeling approach must be powerful enough to express all relevant properties of a system. In order to enable the modeling of systems with the characteristics mentioned in the previous section, the modeling approach needs to meet the following requirements [13],[14],[15],[16]:

- Analogy: A model shall be analogous to existing models or at least to a *conceptual* model to avoid conceptual *clash*. If there is such analogy, it should be pointed out to help establishing links to the existing conceptual model of a user and facilitates reasoning: human reasoning is less based on an application of formal laws of logic than on memory retrieval and analogy.
- Utility: A model shall serve a useful well-defined purpose.
- Stability: A model shall be usable uniformly in many different contexts without any qualification or modification.
- Projection: A model shall support diverse and integrated views on the system under development. By this way, different aspects of the system can be independently analyzed and specified.
- Modularity: A model shall have a hierarchical organization of its composing elements. Elements might be aggreagated and encapsulated in a higher-level element to facilitate analysis and eventual reuse.
- Compositionality: A model shall allow deducing the properties of a system from the properties of its subsystems. It is essential for the reuse of existing components, and enables to incrementally build systems out of modularly specified parts.
- Abstraction: A model shall enable capturing properties of a system that are needed to understand one of its aspects without paying attention to the details or the other aspects of the system.
- Refinement: A model shall start with high-granular descriptions and allow to incrementally refining them into more detailed ones. A refined model shall guarantee all the properties of the abstract model.

5 System Modeling Language

To model a complex socio-technical system, it is necessary to address all the aspects. Usually, models focus on structure and behavior. It is likely that both model categories need to be combined and used at different points in the system development process.

Based on literature and industrial experience, a set of modeling language requirements have been identified:

1. Simplicity: The language shall be simple, with few basic concepts [17],[18]. It should be straight-forward to determine what a model element represents.
2. Visual, graphical language: The language shall allow a visual, graphical depiction of the model, giving both an overview of the whole system and details about its parts [18].
3. Visual flexibility: The language shall allow to modify the graphical properties of the elements represented in the model if necessary (for example to highlight elements) [4]
4. Semantic flexibility: Model Semantics might evolve. The meaning of a model element should not always be fixed.[19]
5. Domain and user specificity: Modeling language shall be domain/user specific: it shall map to the conceptual world of those performing the work [2],[20],[21]. Domain specific concepts should be used [22].
6. Semantic Preciseness: Since models have many users and uses, they shall be precise and unambiguous. Any ambiguity will lead to errors, confusion and consequently to increased cost [23].
7. Customizability: The language shall allow contextualisation and personalisation for both semantic and graphical aspects: predefined rules may be bypassed when the situation requires it. [2]
8. Compositionality: Modeling languages shall allow composition of models from parts [24].
9. Multiple views: Modeling language shall allow multiple perspectives and interpretations to coexist and evolve [25].
10. Integrability: Different viewpoints on each concept shall be represented together, not as disintegrated model views [26].
11. Extensibility: Modeling language shall be extensible so that users can its own metamodel from an existing one [21].
12. Textual property: Modeling language shall have an integrated user interface that integrate textual attributes [2].
13. Versioning: Modeling language shall allow capturing history of events and model changes [2][27].
14. Completeness check property: Modeling language shall be able to identify and interprete incomplete models [2].

6 Conclusion

System complexity is usually due to the recursive intricacy and the interactions between the subsystems. However, human behavior makes a system far more complex and complicated due to the perception: stakeholders usually do not have the holistic view that enables understanding of the system and taking into account all the factors and elements that can are related to the design process. To handle the complexity of a complex system design, we proposed a better understanding by defining a system engineering paradigm. We identified the several levels a system can have, and define the functional spaces a system usually has. It is expected that this approach could be used for any system.

In this paper, we also defined the requirements a system model needs to meet to be trustworthy, useful and understandable. One of the most important issues addresses was the model complexity. To handle the complexity, it is necessary to architecture the model. Hierarchy is the most intuitive way to address this issue. Two main types of hierarchy have been defined in that purpose. Besides, a set of modeling requirements have been defined. They have then been refined to get modeling language requirements. These requirements are then used to choose the tools that are to be used in System Engineering, or as user needs when designing such tools. Literature review and industrial experience indicate that system engineers are still in need of a modeling language that is simple and intuitive to support many tasks in system engineering and architectural reasoning.

References

1. Stefan Jablonski: Mobile: A modular workflow model and architecture. In Proceedings of Conference on Dynamic Modelling and Information Systems, Noordijkerhout, Netherlands, (1994)
2. Hvard D. Jorgensen: Interactive Process Models. PhD thesis, Norwegian University of Science and Technology Trondheim, Norway, (2004)
3. Martin, James N.: Systems Engineering Guidebook: A Process for Developing Systems and Products. CRC Press, Inc.: Boca Raton, FL, (1996)
4. Hsueh-Yung Benjamin Koo: A Meta-language for Systems Architecting. PhD thesis, MIT, USA, (2005)
5. Mark Sean Avnet: Socio-Cognitive Analysis of Engineering Systems Design:Shared Knowledge, Process, and Product. PhD thesis, MIT, USA, (2009)

6. Ali Mostashari: Stakeholder-Assisted Modeling and Policy Design Process for Engineering Systems. PhD thesis, MIT, USA, (2005)
7. J.A. Cannon-Bowers, E. Salas, and S.A. Converse: Shared mental models in team decision making. In Individual and Group Decision Making, pages 221246. N.J. Castellan Jr, (1993)
8. Kevin Forsberg and Harold Mooz: The relationship of systems engineeringto the project cycle. Journal of Applied Psychology, 85(2):273283, 2000.
9. Herbert A. Simon: The architecture of complexity. Proceedings of the American Philosophical Society, 106(6):467482, (1962)
10. Graeme S. Halford, Rosemary Baker, Julie E. McCredden, and John D. Bain: How many variables can humans process? Psychological Science, 16(1):7076, (2005)
11. Luca Cardelli and Peter Wegner: On understanding types, data abstraction,and polymorphism. ACM Comput. Surv., 17 (4):471522, (1985)
12. Valerie Ahl and T. F. H. Allen: Hierarchy Theory - A Vision, Vocabulary, and Epistemology. Columbia University Press, (1996)
13. Alexander Harhurin, Judith Hartmann, and Daniel Ratiu: Motivation and formal foundations of a comprehensive modeling theory for embedded systems. Technical report, Technical University of Munich, (2009)
14. G. S. Halford, J. Wiles, M. S. Humphreys, and W. H. Wilson: Parallel distributed processing approaches to creative reasoning: Tensor models of memory and analogy. In Proceedings of the AAAI Spring Symposium, Palo Alto, California, USA, (1993)
15. Hermann Kopetz: The complexity challenge in embedded system design. In ISORC, pages 312, (2008)
16. Marc Bouissou: Gestion de la complexit dans les tudes quantitatives de sret de fonctionnement de systmes. Lavoisier, (2008)
17. A. Agostini and G. DeMichelis: A light workflow management system using simple process models. Computer Supported Cooperative Work, 9(3-4):335363, (2000)
18. M.L. Jaccheri, G.P. Picco, and P. Lago: Eliciting software process models with the e3 language. ACM Transactions on Software Engineering and Methodology, 7(4):368410, (1998)
19. Mario Bunge: Philosophy of Science - Vol.1 From Problem to Theory. Transaction Publishers, (1998)
20. R. E. Grinter: Workflow systems: Occasions for success and failure. Computer Supported Cooperative Work, 9(2):189214, (2000)
21. H. D. Rombach and M. Verlage: Directions in software process research. Advances in Computers, 41:163, (1995)
22. Juha pekka Tolvanen and Kalle Lyytinen: Flexible method adaptation in case: The metamodelling approach. Scandinavian Journal of Information systems, 5:5177, (1993)
23. Matti J Kinnunen: Complexity measures for system architecture models. Master of science in engineering and management, Massachusetts Institute of Technology, (2006)
24. B. C. Warboys, D. Balasubramaniam, R.M. Greenwood, G. N. C. Kirby, K. Mayes, R. Morrison, and D. S. Munro: Collaboration and composition: Issues for a second generation process language. In Proceedings of the 7th European Software Engineering Conference, pages 7590. Springer-Verlag, (1999)
25. J. C. Grundy, J. G. Hosking, and W. B. Mugridge: Inconsistency management for multiple-view software development environments. IEEE Transactions on Software Engineering, 24 (11):5177, (1998)
26. Andreas L. Opdahl and Guttorm Sindre: Facet modelling: An approach to flexible and integrated conceptual modelling. Information Systems, 22(5):291323, (1997)
27. G. Cugola: Tolerating deviations in process support systems via flexible enactment of process models. IEEE Transactions on Software Engineering, 24(11):9821001, (1998)

Migration from Legacy Systems to SOA Applications: A Survey and an Evaluation

Sukanya Suwisuthikasem and M.H. Samadzadeh

1 Introduction

As computer software grows in power, users demand ever more powerful and reliable programs, resulting in ever larger and seemingly more complex software systems. Thus a major challenge in large-scale software development is managing the complexity encountered during the construction of new applications that is partly attributable to frequent customizations due to requirement changes.

This complexity can potentially be reduced if applications could be implemented based upon an open standard-based interface and communication protocol. From this, all applications can be accessed more efficiently and easily, thus enabling businesses to leverage their existing software systems. In addition, since business requirements are typically subject to frequent changes, the demand on existing systems to evolve is inevitable. Consequently, there is a need to have an efficient method to support the software evolution process. Service Oriented Architecture (SOA), a new approach to developing software systems, has been eventually invented to fulfill this demand.

SOA has gained significant popularity for achieving business goals and implementing business processes in a flexible manner. SOA is becoming a mainstream approach for software development. Abrams and Schulte [1] indicated that during 2007, more than 50 % of large, newly-developed systems and business processes were designed and developed based on the Service Oriented Architecture paradigm. Lublinsky [2] suggested there are three primary reasons that businesses are interested in SOA. First, by adopting SOA they can achieve better alignment between business and Information Technology (IT). Second, SOA enables them to construct more flexible and responsive IT infrastructure. And last, that SOA can simplify the implementation of data integration among a business' applications. Based on this argument, in order to sustain their competitiveness to be leaders in the market, businesses need to transform their legacy systems to SOA applications toward providing more efficient services to their customers.

Generally, migration from legacy systems to SOA applications is carried out manually by domain experts with subject-matter knowledge with some training in software development or with the assistance of software developers [3, 4]. The problem of migrating (i.e., transforming/adapting/retargeting) large-scale legacy software systems to a modern environment, to new hardware platforms, and to new run-time support is a major issue facing the software industry. The focus of this work was to investigate the existing migration approaches and capture the significant features of each approach so as to give the guideline for businesses to choose a tailor-made migration approach suiting them.

2 General Migration Processes

2.1 Legacy System Assessment

Sommerville [6] presented four strategic options for software evolution as illustrated below.

As shown in Fig. 1, there are four clusters for legacy systems as described below.

- Low Quality, Low Business Value: These systems should be discarded since it is costly and unproductive for businesses to keep them.
- Low Quality, High Business Value: These systems are still productive but costly to maintain. Therefore they should be transformed to a system with a new SOA architectural style.
- High Quality, Low Business Value: Although they are inexpensive to maintain, these systems should be discarded as they are unproductive.

S. Suwisuthikasem (✉) • M.H. Samadzadeh
Computer Science Department, Oklahoma State University, Stillwater, USA
e-mail: suwisut@cs.okstate.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_86,

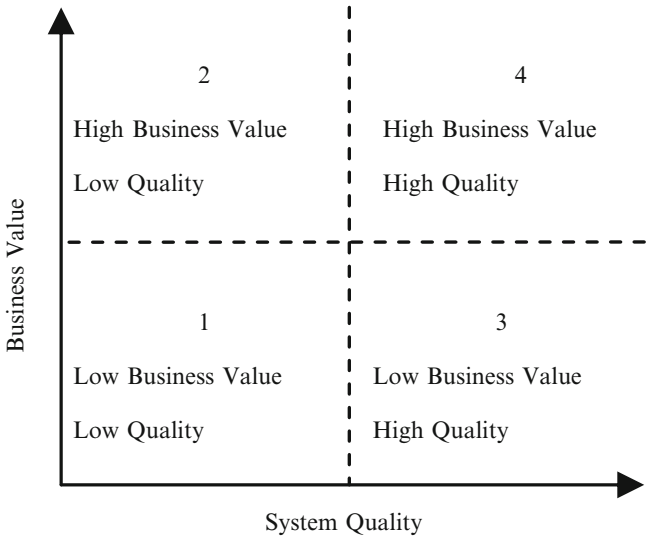


Fig. 1 Legacy System Assessment (Source: Software Engineering, 9th edition, p.253)

- High Quality, High Business Value: It is cost-effective to maintain these systems. So these systems should be kept without making any changes the system.

Several criteria need to be defined and quantified so as to measure the quality of an SOA system. This can be performed by interviewing some domain experts or using a questionnaire. Fortunately, there are a lot of researchers working in this area by defining a number of criteria to assess legacy systems. Four basic attributes of a legacy system, namely Business Value, Decomposability, Obsolescence, and Deterioration, and guidelines to measure each attribute were introduced by Cimitile et al. [25]. Ransom et al. [26] presented a method to assess legacy systems namely technical, business, and organizational aspects. This method could help businesses to assign a value to each assessment characteristic and also to interpret these values. They indicated that their method can be tailored to specific organizations.

2.2 Feasibility Analysis

Khadka et al. [12] proposed the seviciFi method using method engineering to determine the economical and technical feasibility of the migration based on the legacy system's charactersitics and the requirements of the target SOA application. Erradi et al. [13] proposed a decision framework to help organizations to select the optimal combination of legacy modernization options. Aly and Amir [14] presented a decision making tool using decision theory and the weighted sum methodology to generate the most optimal strategy to be used in modernizing legacy systems. They also proposed an automated decision making process for choosing a migration strategy using a combination of the approaches proposed in [23] and [13]. Aversano and Tortorella [27] proposed a strategy to help businesses define the evolution requirement to SOA based on characteristics of their legacy systems. They specified nine steps in software evolution: analyze the organization, reconstruct processes, identify and analyze the processes, identify technology, formulate evolution requirements, assess the legacy software system, define software system evolution requirements, reengineer the processes, and perform the system evolution. The authors reported the first stage of testing their method with two different systems: a bank system and a public administration system. According to their result from the first stage, this method is certainly applicable however it needs additional information and it also needs to be supported by a specific integrated and Web-based software environment.

2.3 Migration

Migration refers to any approach that can be applied to a legacy system in its entirety in the process of transforming it to SOA architecture. This section discussed how legacy code components are identified, decomposed, and extracted using several techniques. The User interfaces of legacy systems are reengineered to be SOA-based system compatible. Migration strategies incorporate both top-down and bottom-up approaches and aim to produce a system with an improved SOA-compatible design.

Aly and Amir [14] proposed a method which automatically generates a modular software structure from a given source code by using spectral clustering. First, undirected graphs are generated based on the existing dependencies among the components and then spectral clustering is performed to generate the component structure of the target system. To evaluate this method, they applied it to CoCoME, a small and well-documented software system. They compared the resulting software structure to the one created by an expert. They reported that the result were similar. However, the component structure resulting from their approach cannot be represented by any architectural style. In their case study, the legacy system is a three-tier architecture but the result is a single-layer architecture.

Chen et al. [8] and Millham [11] presented a service oriented reengineering approach using feature analysis to extract candidate services from legacy systems. Feature analysis addresses the understanding of features in software systems and defines mechanisms for carrying a feature from the problem domain into the solution domain. In [8], the specified feature analysis activities are: identifying system features, constructing a feature model to organize the defined features, and identifying their implementation in the legacy system through feature-location techniques. Based on a feature model, certain service identification and packaging processes are performed that result in service delegation.

In a research conducted by Millham [11], data and control dependencies among the component files of a legacy system are analyzed and then clustered into groups. Cuadrado et al. [9] described a case study of the evolution of an existing legacy system towards a more maintainable SOA system. To define the specific evolution plan, the architecture of the legacy system was recovered. This approach was applied to a medical imaging system evolving it into an SOA-based application. Matos and Heckel [3] proposed the new methodology for migration based on source code analysis for identifying the contribution of code fragments to architectural elements and a graph transformation approach for architectural migration. Alahmari et al. [10] introduced a framework to identify optimal services from legacy code with the appropriate level of granularity, by focusing on the significance of the classification of service types, to define service properties.

Reddy et al. [15] proposed guidelines for evaluating the suitability of existing assets by identifying the core principles of SOA, namely cohesion, reusability, discoverability, loose coupling, abstraction, formal contract, composability, and statelessness. They argued that organizations can use their guidelines to improve the quality of their migrations by considering their defined metrics and guidelines. Stroulia et al. [16] described the overall process for legacy system migration to a Web-based system using the CelLEST method. This method addresses the issue of migration based on understanding and modeling the users' interaction with the legacy system's interface. There are three main steps in this method. The first step is to model the behavior of the old system using a state transition diagram. The second step is to find the users' tasks as frequently-occurring interaction patterns to recover the specifications of the application's functions. The last step is to construct the new user interface allowing the legacy functions to be accessible over the Web.

Aversano et al. [17] presented a migration project aiming to integrate a COBOL system into a Web-enabled system. The legacy system was dissected into user interfaces and server (application logic and database) components. The user interfaces were migrated into a Web browser shell using Microsoft Active Server Pages (ASP) and VBScript. All server components were wrapped with dynamic load libraries written in Microfocus Object COBOL, loaded into Microsoft Internet Information Server, and accessed via the ASP pages.

Werth et al. [18] introduced Business Service Management as an interdisciplinary approach for business-driven deployment of SOA. The main purpose of this work was to represent a business' characteristics and requirements toward IT as business processes. Bhallamudi and Tilley [19] presented the Evolution Process Framework for SOA (a mechanism for analysis of existing SOA migration projects) to learn about factors such as technology selection, migration approach utilized, legacy system type, and SOA governance that influence the success or failure of each project.

Table 1 Relationship between Design Properties and Quality Attributes [24]

	Effectiveness	Understandability	Feasibility	Reusability
Coupling		↓	↓	↓
Cohesion	↑	↑		↑
Complexity		↓		
Design Size		↓		
Service Granularity	↑	↑	↑	↑
Parameter Granularity	↑	↑	↑	
Consumability		↑		↑

Mohagheghi and Sæther [20] applied the model-driven approach to construct a methodology and a tool for transforming a legacy system into a service oriented application. O'Brien et al. [21] used architecture reconstruction in the process of migration. To accomplish this, dependencies among components in the legacy system were identified. Based on this information, an essential step in making decisions regarding migration of legacy components to services was devised. They also suggested that using architecture reconstruction techniques in conjunction with other analytical methods could provide an essential set of analytical methods for decision making.

Marchetto and Ricca [22] proposed a stepwise approach based on testing to migrate Java application into SOA-based application. This approach, which is a hybrid top-down and bottom-up approach, was applied to four Java applications. Lewis et al. [5], [23] described their Service Oriented Migration and Reuse Technique (SMART), a technique helping businesses to analyze legacy capabilities for use as services in an SOA. Their technique considers the specific interactions that will be required by the target SOA and any changes that must be made to the legacy components. They described a wide range of information about legacy functionalities, the target SOA, and the potential services that were aggregated to produce a service migration strategy.

2.4 Evaluation

Shim et al. [24] suggested that, in order to evaluate the quality of an SOA application, a quality assessment model is needed that defines the desired quality attributes and measures them. From this, design problems can be detected and resolved before the development of the system. The relationship between design properties and quality attributes is described in Table 1. Shim et al. [24] also

Table 2 Methods/Techniques Used in the Migration Process

Legacy System Assessment			
Method/Technique	Tools	Pros.	Cons.
Standard Decision Framework [25]	–	– Suitable for any kind of legacy system	– Needs a lot of documentation – Needs experts – Can be affected by errors or biases
Assessment Method [26]	–	– Can be tailored for any specific kind of legacy system – Can be iterated to reduce inaccuracy of the assessment – Provides practical advice and guidance to businesses	– Not yet fully tested and evaluated
Feasibility Analysis			
Method/Technique	Tools	Pros.	Cons.
– SMART [23]	– SMIG	– Helps businesses to analyze and determine if a legacy system can be exposed as services	– Needs experts – Need lots of documentation and user feedback
– Method engineering [12]	– serviciFi	– Helps businesses select supporting technology based on migration feasibility	– Processes are performed manually – Not support for large projects
– Decision theory – Weighted sum methodology – Gap analysis [14]	– Decision Making tool	– Can be customized for specific organizations – High degree of automation	–
– Decision framework [13]	–	– Helps businesses to create the most beneficial and cost-effective migration approach meeting a business' requirement	– Low degree of automation – Cannot guarantee the QoS of the target SOA system
Migration			
Method/Technique	Tools	Pros.	Cons.
– Code Analysis – Pattern Matching – Graph Transformation [8]	– ATX (L-CARE) [35] – EMF [36]	– High degree of automation	– Ongoing project – Support a specific language
– Feature Analysis – Slicing Technique [8], [11]	– WSW [29] – Captain Feature [33] – FEAT[34]	– High degree of automation	– Not suitable for large projects – Supports a specific language
– Undirected Graph – Spectral Clustering [7]	– UML model	– High degree of automation – Supports large projects – No need for documentation	– Needs a method to identify meaningful clusters, otherwise the result will not be correct
– Architecture Recovery [5], [9]	– MOOSE [37] – Rigi [38] – QAR [30] – Jude[31] – Omondo UML Studio [32] – Eclipse TpTP [28]	– High degree of automation	– Not supporting large projects – Specific to Java applications
– Domain Analysis – Wrapping [11]	– UML model – TAGDUR [39] – CORBA/IDL parser cc	– High degree of automation	– Specific type of target SOA: Web-Based system

defined three metric groups that can be directly applied to design components. The metrics in the first group, service internal metrics, use service internal elements such as service name, operations provided by the service, and characteristics of the messages defined in the service. The second group of metrics is service external metrics that use information from services they are connected to. Metrics in this group are used to measure the characteristics of the consumer and producer services that are either directly or indirectly connected to a given service. The last group is system metrics. The metrics in this group are used to measure the characteristics of the entire system in general.

In Table 1, an up/down arrow means increase/decrease of the attribute vis-à-vis an increase/decrease of the respective property.

3 Existing Migration Techniques/Tools

From Section 2, we can summarize the techniques and tools that support the migration as follows.

The information summarized and tabulated in Table 2 can be used to choose from among the techniques and tools available for each step of the migration process.

4 Conclusions

In this paper, several approaches for migrating legacy systems to SOA architecture are investigated and the main processes and tools of each approach are captured and analyzed. The different approaches are compared and contrasted based on their key features and their target legacy system types. Using this comprehensive comparative analysis, businesses can create tailor-made approaches that could best fit their needs and satisfy their requirements.

References

1. Charles Abrams, Roy W. Schulte: Service Oriented Architecture Overview and Guide to SOA Research. Gartner Research (2008)
2. Boris Lublinsky: Defining SOA as an Architectural Style, IBM, http://www.ibm.com/developerworks/architecture/library/ar-soastyle/
3. Carlos Matos, Reiko Heckel: Migrating Legacy Systems to Service Oriented Architectures. In: Doctoral Symposium at the International Conference on Graph Transformation (ICGT 2008), Vol. 16, pp.1-15 (2008)
4. Gerardo Canfora, Anna Rita Fasolino, Gianni Frattolillo, Porfirio Tramontana: Migrating Interactive Legacy Systems to Web Services. In: 10th European Conference on Software Maintenance and Reengineering, pp. 27-36. Bari, Italy (2006)
5. Grace Lewis, Edwin Morris, Dennis Smith: Analyzing the Reuse Potential of Migrating Legacy Components to a Service Oriented Architecture. In: 10th European Conference on Software Maintenance and Reengineering, pp. 15-23. Bari, Italy (2006)
6. Ian Sommerville: Software Engineering, 9th Edition. Pearson Education Inc., Essex, England and Addison-Wesley Publishers. Boston, MA (2011)
7. Constanze Deiters, Andreas Rausch, Mirco Schindler: Using Spectral Clustering to Automate Identification and Optimization of Component Structures. In: 2nd International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE), pp. 14-20. San Francisco, CA (2013)
8. Feng Chen, Shaoyun Li, Hongji Yang, Ching-Huey Wang, William Cheng-Chung Chu: Feature Analysis for Service Oriented Reengineering. In: 12th Asia-Pacific Software Engineering Conference: APSEC '05, pp. 201-208. Taipei, Taiwan (2005)
9. F. Cuadrado, B. Garcia, J. C. Dueas, H. A. Parada: A Case Study on Software Evolution Towards Service Oriented Architecture. In: 22nd Int. Conf. on Advanced Information Networking and Applications: AINAW 2008, pp. 1399-1404. Okinawa, Japan (2008)
10. Saad Alahmari, Ed Zaluska, David De Roure: A Service Identification Framework for Legacy System Migration into SOA. In: 7th International Conference on Services Computing, pp. 614-617. Miami, FL (2010)
11. Richard Millham: Migration of a Legacy Procedural System to Service Oriented Computing Using Feature Analysis. In: International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS), pp. 538-543. Krakow, Poland (2010)
12. Ravi Khadka, Gijs Reijnders, Amir Saeidi, Slinger Jansen, Jurriaan Hage: A Method Engineering based Legacy to SOA Migration Method. In: 27th IEEE International Conference on Software Maintenance, pp.163-172 (ICSM) (2011)
13. Abdelkarim Erradi, Sriram Anand, Naveen Kulkarni: Evaluation of Strategies for Integrating Legacy Applications as Services in a Service Oriented Architecture. In: IEEE Int. Conf. on Services Computing (SCC'06), pp. 257-260. Chicago, IL (2006)
14. Sherif G. Aly, Rafik Amir: Automated Selection of Legacy Systems SOA Modernization Strategies Using Decision Theory. International Journal of Software Engineering and Its Applications, Vol. 3, No. 4, pp. 65-86 (2009)
15. Vinay Kumar Reddy, Alpana Dubey, Sala Lakshmanan, Srihari Sukumaran, Rajendra Sisodia: Evaluating legacy assets in the context of migration to SOA. In: 10th IEEE International Symposium on High Perfor-mance Distributed Computing, pp.51–63. Springer Science + Business Media, LLC (2008)
16. E. Stroulia, M. El-Ramly, P. G. Sorenson: From Legacy to Web Through Interaction Modeling. In: 18th Int. Conf. on SW Maintenance, pp. 320-329. Montreal, Canada (2002)
17. Lerina Aversano, Gerardo Canfora, Aniello Cimitile, Andrea De Lucia: Migrating Legacy Systems to the Web: An Experience Report. In: 5th European Conference on Software Maintenance and Reengineering, pp. 148-157. Lisbon, Portugal (2001)
18. Dirk Werth, Katrina Leyking, Florian Dreifus, Jörg Ziemann, Andreas Martin: Managing SOA Through Business Services: A Business-Oriented Approach to Service Oriented Architectures. In: 4th International Conference on Service-Oriented Computing: ICSOC 2006, LNCS 4652, pp. 3-13. Chicago, IL (2007)
19. Pushparani Bhallamudi, Scott Tilley: SOA Migration Case Studies and Lessons Learned. In: IEEE Int. Systems Conference (SysCon), pp. 123-128. Montreal, Canada (2011)
20. Parastoo Mohagheghi, Thor Sæther: Software Engineering Challenges for Migration to the Service Cloud Paradigm: Ongoing Work in the REMICS Project. In: IEEE World Congress on Services, pp. 506-514. Washington, DC (2011)
21. Liam O'Brien, Dennis Smith, Grace Lewis: Supporting Migration to Services Using Software Architecture Reconstruction. In: 13th IEEE International Workshop on Software Technology and Engineering Practice, pp. 81-91. Budapest, Hungary (2005)
22. Alessandro Marchetto, Filippo Ricca: From objects to services: toward a stepwise migration approach for Java applications. In: International Journal of Software Tools Technology Transfer. Springer-Verlag (2009)
23. Grace Lewis, Edwin Morris, Dennis Smith: The Service Oriented Migration and Reuse Technique (SMART). In: 13th IEEE International Workshop on Software Technology and Engineering Practice, pp. 222-229. Budapest, Hungary (2005)
24. Bingu Shim, Siho Choue, Suntae Kim, Sooyong Park: A Design Quality Model for SOA. In: 15th Asia-Pacific SE Conference, pp. 304-410. Beijing, China (2008)
25. Aniello Cimitile, Anna Rita Fasolino, Filippo Lanubile: Legacy Systems Assessment to Support Decision Making. In: IEEE Workshop on Empirical Studies of Software Maintenance (WESS '97), pp.145-150. Bari, Italy (1997)
26. Jane Ransom, Ian Sommerville, Ian Warren: A Method for Assessing Legacy Systems for Evolution. In: 2th Euromicro Conference on Software Maintenance and Reengineering, pp. 128–134. Florence, Italy (1998)

27. Lerina Aversano, Maria Tortorella: An assessment strategy for identifying legacy system evolution requirements in eBusiness context. Journal of Software Maintenance and Evolution: Research and Practice, pp. 255-276. John Wiley & Sons, Ltd. (2004)
28. Eclipse TPTP (Test and Performance Tools Project), an Eclipse top-level project, http://www.eclipse.org/tptp
29. H. Guo, C. Guo, F. Chen, H. Yang: Wrapping Client-Server Application to Web Services for Internet Computing. In: 6th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'05). Dalian, China (2005)
30. Arciniegas, J.L: Contribution to Quality-driven Evolutionary Software Development Process for Service Oriented Architecture. Ph. D. Thesis, Polytechnic Uni of Madrid (2006)
31. Jude (Java and UML Developer Environment), a Java UML modeling tool, http://jude.change-vision.com
32. Omondo Eclipse UML Studio, an Eclipse plug-in for UML modeling, http://www.omondo.com
33. K. Czarnecki, U. W. Eisenecker: Generative Programming, Addison Wesley (2000)
34. M. P. Robillard, G. C. Murphy: FEAT: A Tool for Locating, Describing, and Analyzing Concerns in Source Code. In: 25th Int. Conf. on SE. Oregon, Portland (2003)
35. The migration specialists, http://www.atxsoftware.com/
36. Eclipse. Eclipse Modeling Framework (EMF), http://www.eclipse.org/emf/
37. Ducasse, S., Lanza, M., Tichelaar, S.: Moose: an Extensible Language-Independent Environment for Reengineering Object-Oriented Systems. In: 2nd International Symposium on Constructing Software Engineering Tools: CoSET'00 (2000)
38. Kazman, R. O'Brien, L., Verhoef, C: Architecture Reconstruction Guidelines, 2nd Edition, CMU/SEI-2002-TR-034 (2002)
39. Richard Millham, Jianjun Pu, Hongji Yang: TAGDUR: A Tool for Producing UML Sequence, Deployment, and Component Diagrams Through Reengineering of Legacy Systems. In: 8th IASTED Int. Conf. on SE and Application: SEA (2004)
40. Chapter 5: Introduction to CORBA IDL, http://documentation.progress.com/output/Iona /orbix/gen3/33/html/orbix33java_pguide/IDL.html. IONA Technologies PLC (2000)

An Approach to Schedule Production using the Reservation Tables

Sergiu Zaporojan, Vasile Moraru, and Adrian Groza

1 Introduction

The reengineering and improvement of modern manufacturing enterprises is an extremely complex process because it involves a combination of technological, human, machine, and organizational issues. In order to solve such a problem, managers need models of the enterprise at the shop floor level, as well as at organizational and business levels.

Particularly, the problem of finding optimal schedules is the crucial one in real productions systems. Furthermore, research should even more focus on solving real-world scheduling problems. It is very important and necessary to provide models, methods and algorithmic tools which help managers of enterprises to construct and maintain (online!) optimal production lines. Users of the decision making support systems want to find the best suited solution. In order to achieve this goal a variety of supporting decision making algorithms are used to advice manager an optimal production scheme. Decision making is related to the problem of selecting the optimal solution from all accessible schedules.

In this paper we focus on the problems of scheduling of production where specific requirements have to be considered like production of glass-coated microwire and products based on it. Scheduling of such production systems represents a complex problem because of many specific issues. The paper discusses the possibility to use the method of reservation tables, which is well known in the computer pipelining, in the area of production planning and control. The main purpose of the paper is to show that the method of reservation tables can be the way to develop feasible schedules in computation times short enough to be accepted in real-world decision support systems.

The paper is organized as follows. Firstly, a brief review of the problem domain is presented. This includes some basic knowledge related to joint balancing and scheduling problem (section 2). Secondly, we put in discussion the manufacturing process of microwire and products based on it. The first subsection of section 3 contains a brief introduction to the method of reservation tables. Next subsection describes the process of microwire production and resources involved. Some technological, human, and organizational issues are pointed out. After that, a possible tentative approach to solve the problems of optimal scheduling in a microwire production system is formulated. Finally, the conclusions are drawn in section 4.

2 Background

The requirements for optimal organization of the enterprise activity are actual and fit into the idea of reengineering. This idea is centered on all processes in the modern enterprise. Reengineering is a radical redesign of a business or production process to achieve a considerable improvement in performance indicators (cost, quality, productivity, etc.). The idea of reengineering is actual one as the information technology is constantly changing.

Production scheduling can be defined as the allocation of available production resources over time to best satisfy some set of criteria [1–3]. Generally speaking, the scheduling problems are NP-hard. It is probably impossible to secure optimal solutions using fast algorithms. However, some problems are in fact “easy”, in the sense that they are solvable to optimality by fast algorithms. It is well known, that for NP-hard problems an approximation is constructed for the whole problem. When building a specific approximation, different methods and techniques may be used together such as exact polynomial methods, iterative approaches,

S. Zaporojan (✉) • V. Moraru
Technical University of Moldova, Chisinau, Republic of Moldova
e-mail: zaporojan@mail.utm.md; moraru@mail.utm.md

A. Groza
Technical University of Cluj-Napoca, Cluj-Napoca, Romania
e-mail: Adrian.Groza@cs.utcluj.ro

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_87,

relaxation methods, genetic algorithms. Scheduling problems are found in a lot of different applications.

In general, scheduling deals with the temporal assignment of known activities to accessible and limited resources. Usually, in a production system a set of constraints has to be regarded. The generation of a normative schedule under the production constraints is a hard problem. However, in the real-world production system another serious problem should be tackled. It is about adaptation of an existing normative schedule to the changing scheduling environment. In other words, a real schedule must be maintained under the information about changes and turbulences, which include events from the shop floor like resource breakdown or maintenance of the equipment.

Decision support for an optimal business in the framework of an industrial enterprise is not always possible without software products that are based on mathematical models of combinatorial optimization. It is the case of the manufacturing operations within a production line. The assembly production line consists of a finite number of workstations that are running individual operations (tasks) to manufacture a product. The decision problem of optimally partitioning the assembly tasks among the stations with respect to some objective is known as the assembly line balancing problem (ALBP).

Different aspects of balancing and scheduling problems are deeply discussed in the literature [3–10]. An excellent classification of ALBP is given in [6]. Anyway, the problem of optimal production scheduling remains very actual and important.

3 Scheduling Based on Reservation Tables

3.1 Method of the reservation tables and scheduling

An assembly line in a manufacturing plant is somewhat similar to the use of a computer pipeline. An assembly line takes advantage of the fact that a product goes through various stages of production. By laying the production process out in an assembly line, products at various stages can be worked on simultaneously. This process is referred to as pipelining.

In computers, pipelining is an implementation technique whereby multiple instructions (operations) are overlapped in execution. It takes advantage of parallelism that exists among the actions needed to execute an operation.

A pipeline is like an auto assembly line, where there are many steps, each contributing to the construction of the car. Each step works in parallel with the other ones. In a computer pipeline, each step performs a part of an operation. Each of these steps represents a pipe stage. Similar to the assembly line, different stages are performing different parts of different operations in parallel [11].

In an assembly line, throughput is defined as the number of completed units per hour and is determined by how often a unit exits the assembly line. The throughput of a computer pipeline is determined by how often an operation exits the pipeline.

The designer of the assembly line wants to balance the time for each step in the production process. On the other hand, one of the pipeline designer's goals is to balance the length of each pipeline stage. Another well known problem is the scheduling of production where manufacturing operations have to be assigned to limited resources like stations. The designers of pipelined computers are also faced with scheduling, which deals with the temporal assignment of operations to pipeline stages.

The key idea of a pipeline computer is to overlap pipeline operations as much as possible. The approach is known as the method of reservation tables [12, 13]. This technique permits to construct controllers which admit new operations (tasks) to the pipeline in a manner that can sustain the maximum throughput of the pipeline, while guaranteeing that two or more operations do not collide within the pipeline.

A reservation table gives the necessary timing information about a concrete operation (task). To anticipate future collisions, one needs to develop a two-dimensional timing diagram that shows the flow of data through the pipeline unit. Each row of the table represents a stage and each column of the table represents a time step. The rows are labeled according to the structure of the operation.

The problem then becomes one of determining how to control a pipeline with the given reservation tables. To simplify the problem, a single function pipeline it is better to be considered. Now, if an operation is launched into the pipeline, at what future times can we launch another identical operation? The answer is contained in the so-called collision vector, and which is derived from the original reservation table. The collision vector is a binary vector and contains the collision information. Position j contains a bit which indicates whether or not a new operation can be launched j time units after an operation has been initiated. A binary 0 indicates that no collision will occur and a 1 prevents on future collision.

To understand more deeply the principles and importance of collision vectors, the so-called reduced state-diagrams must be analyzed. Designers can find the maximum rate that can be sustained for the pipeline by examining the cycles in the reduced state-diagram [12, 13]. We apply some principles of the method of reservation tables in the next subsection.

3.2 The microwire production system: technological and human issues

The aim of materials processing in the metallurgic industry consists in developing and reaching of some new materials, with improved properties and performances, and finding of some new processing methods. The microwire technology and its production represent an important way to develop new advanced materials and products based on them.

A microwire consists of an inner metallic nucleus covered by a Pyrex-like coating. Such microwires show magnetic properties of great technological interest like magnetic bistability, soft magnetic and memory shape properties. The above properties are quite useful for a variety of sensor applications. The investigation into technology and physical properties of glass-coated microwires is presently attracting much attention because of their use in sensor devices and fiber-based products.

From another point of view, the modern equipments uses more and more powerful high-frequency radiators which, on the one hand, widely expand the technical opportunities of industry and communication, but on the other hand are dangerous to the life and health of people and the environment. Therefore, the problem of protection against such radiations is vital. Application of both woven and non-woven materials containing microwire allows creating a flexible protective material easy to be used. The microwire can be also built in glass and plastic which allows obtaining shielding material with practically full transparency. We have indicated just a few applications of glass-coated microwire.

Let us briefly describe the process of microwire casting [14]. Glass-coated microwires are manufactured by means of the Taylor-Ulitovsky technique. A rod of the alloy of desired composition is put into a Pyrex-like glass tube and placed within a high frequency inductor heater. The alloy is heated up to its melting point, forming a droplet. While the metal melts, the portion of the glass tube adjacent to the melting metal softens, enveloping the metal droplet.

A glass capillary is then drawn from the softened glass portion and wound on a rotating bobbin. At suitable drawing conditions, the molten metal fills the glass capillary and a microwire is thus formed where the metal core is completely coated by a glass shell. The amount of glass used in the process is balanced by the continuous feeding of the glass tube through the inductor zone, whereas the formation of the metallic core is restricted by the initial quantity of the alloy droplet. The process of casting is carried out at a temperature that will melt the alloy and soften the glass tube. The final microwire structure is formed by water-cooling to obtain a metallic core in amorphous or non crystalline state.

The microstructure of a microwire (and hence, its properties) depends mainly on the cooling rate, which can be controlled by a cooling mechanism when the metal-filled capillary enters into a stream of cooling water. After passing through the cooling water of the crystallizer, the microwire comes to spool on the receiving mechanism.

The geometrical characteristics of the microwire depend on the physical properties of both the glass and alloy composition, the diameter of the initial glass tube, and the parameters of the heating inductor. The diameter of a microwire produced by Taylor-Ulitovsky method has both upper and lower limits depending on the speed of casting. Typical limits for the metallic core diameter are between 1 and 50 microns, while the thickness of the coating is in the range of 2 and 15 microns. It should be noted, that even during the stationary casting process, there is some variation in diameter of the metallic nucleus and in the glass coating thickness along the wire length. Depending on the required diameter, the precision can vary from 5 % for wires in the range between 5 and 10 microns, to 10 % for wires in the range between 10 and 30 microns. On the other hand, it may be adequate to have a system that is capable of casting to a precision of rather better than 5 % within quite wide limits. Currently used plants are not capable to meet the latter requirements.

Having reviewed the process of casting, let us examine more carefully the industrial production machine. An industrial production machine for the fabrication of glass-coated microwire consists of some special blocks and mechanisms. At the level of mechanisms we can distinguish three of them. Two are intended for moving down the rod alloy and the glass tube. Another mechanism is the receiving one.

Next, the alloy of desired composition is put into the glass tube and placed within the inductive heater. To heat the alloy up to its melting point, a high frequency generator must be there. To maintain the molten drop at an optimum position over the inductive heater, it is necessary to control the pressure inside the glass tube. Another level of the machine is dealing with measurement and sensor devices. The most important sensor is the meter of microwire resistance. Finally, at the top level of the machine a control block must be present.

Basically, the technological resources of the microwire production system involve a set of casting machines configured so as to produce a desired microwire, which represents a component of a final product. The numerous human operators are present on the production floor. After casting, the microwire must be tested (fig.1) for quality on measurement equipment. The last one contains an electro-mechanical rewinding system which is equipped with a sensor connected to a processing electronic device. Depending on the application, there are a few additional specialized measurement devices.

From the above, we can observe a few technological, human, and organizational issues. The main technological

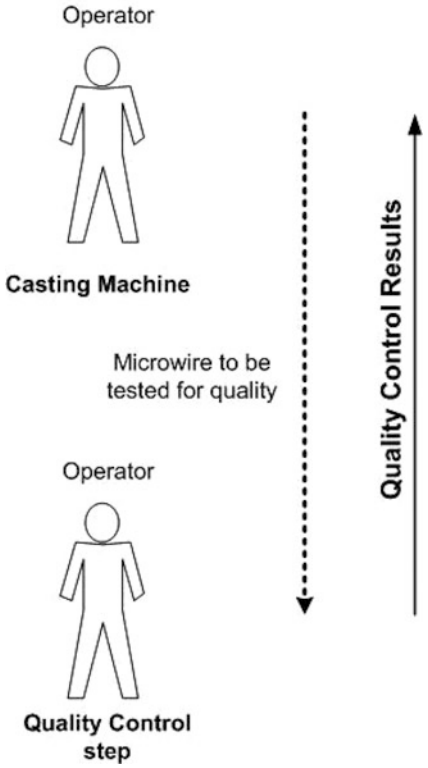


Fig. 1 The interaction between human operators

problem is to maintain the optimum thermal and flow conditions of the process, in order to cast the microwire of a desired stable diameter. Some important factors, such as microwire vibrations and defects of the quality of glass and alloy compositions may significantly disturb the casting process.

At issue here is the cost of manufacturing. Microwire requires testing to make sure that each new bobbin of microwire represents a good copy of an original bobbin. The quality assurance is very complex, especially for some types of microwire. When the microwire does not meet the requirements for quality, this leads to unscheduled loss of production. Breakdown of a casting machine affects the production system, causing a new cycle of casting and testing for quality. This introduces idle times for the next stages of the production line.

A brute force approach is to replicate the casting machine and measurement equipment of the line to allow the manufacturing system to use two or multiple wire production streams. Despite some essential drawbacks, this strategy can improve performance of the manufacturing system.

In a real production system, a reconfiguration becomes necessary whenever there is a substantial change in the structure of the production program. In a reconfiguration, each station can be recognized in the form of allotted resources and a location in the workshop. What it is very important, that not only the machinery, but also the operators of the production line are assigned to a certain station. They are especially trained to carry out specific work content, so that a change may be associated with training costs. Therefore, it is desirable to maintain the reconfigured line as close as possible to the previous one. This aspect must be considered by the decision support system.

Figure 2 presents a simplified view on the production line of microwire-based products. The production line consists of five interconnected stages. The first stage performs the operation of loading the casting machine (given by second stage) with glass tube, rod of desired alloy, and an empty bobbin. The third stage represents the quality control equipment. The fully tested wire is then processing according to the production program. Lastly, required batches of final wire-based products are provided.

Unfortunately, the model in figure 2 is unrealistic one because of technological and human issues shortly described above.

Another relative simple line is shown in figure 3. What it is important, is that this model is fully realistic. As it can be seen from the figure, there is a critical path within the line. This path includes the loading and casting stages, as well as the stage of quality control. Another very important observation is that feedbacks are present within the critical path. This fact makes the problem of balancing and scheduling of such lines to be very complex.

Figure 4 shows a possible diagram for the production line. It should be mentioned, that this diagram is one from a huge of cases. In order to obtain useful information for the process of scheduling, we decided to apply the method of reservation tables.

The timing diagram is reflected in the Table 1. Unlike the computer pipeline, here we must consider some constraints regarding to the overlapping of operations within the line. We use the notion of joint operations. A pair of two given operations is tightly jointed if no overlapping is permitted between them. This is the case of “load – casting” chain.

On the other hand, we can have a pair of operations which may be conditionally treated as a joint operation. Then, two given operations are loosely jointed. This is the case of “casting – quality control” chain.

The completed reservation table considers that operations are tightly jointed. In this case, a feasible cycle of length 10 (50 minutes) can be observed (table 2). However, from the reservation table follows that the minimal achievable latency (MAL) equals 7 time units. So, there are may be cycles of length 7 (35 minutes), but it is not sure.

If we want to construct the reduced state-diagram, we need the value of the collision vector. The collision binary

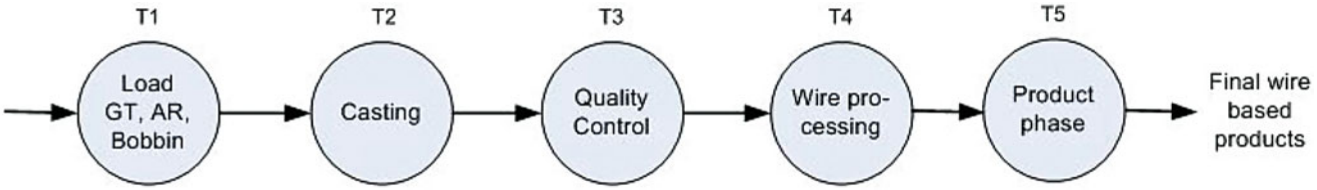


Fig. 2 A simplified model of a microwire-based production line

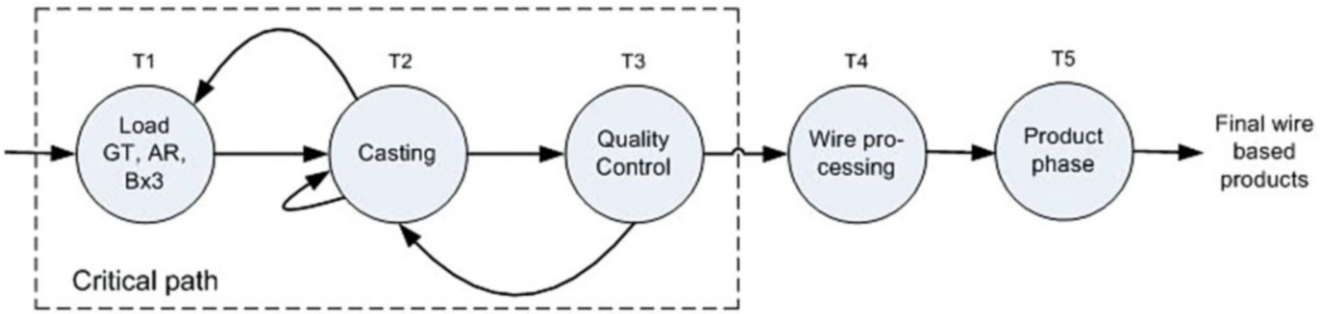


Fig. 3 A realistic production line

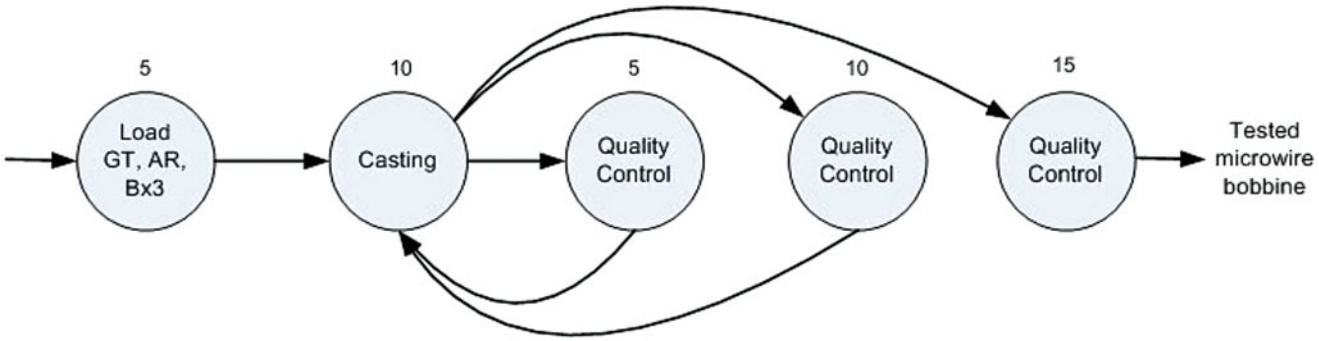


Fig. 4 A diagram of the production line

Table 1 The reservation table

	0	1	2	3	4	5	6	7	8	9	10	11	12
Load	x												
Casting		x	x		x	x			x	x			
QControl				x			x	x			x	x	x

Table 2 Derivation of collision vector

	0	1	2	3	4	5	6	7	8	9													
Load	x										**y**										**z**		
Casting		x	x		x	x			x	x		y	y		y	y			y	y		z	z
QControl				x			x	x			x	x	x	y			y	y			y	y	y

vector is {1111111111000}. The reduced state-diagram is trivial with this collision vector.

Let us suppose that the constraints for "casting – quality control" chain can be relaxed. If this chain is considered to be loosely jointed, a new reservation table will be obtained. It is easy to show, that there is a feasible cycle of length 7 (35 minutes). The utilization of resources is very high in the last case.

4 Conclusions

The proposed approach it seems to be an acceptable one. So, we suppose that the method of reservation tables should be adapted and developed to schedule nontrivial production systems, especially when multi-mixed model lines work in small batches. Combining with other methods, the proposed approach may offer feasible schedules in computation times short enough to be accepted in real-world decision support systems. Lastly, user-friendly decision support system is to be developed, which must be flexible enough to be successfully applied to the microwire production planning.

Acknowledgments This paper was supported by the Microwire Industrial Technologies company (MFTI Ltd, Chisinau, Republic of Moldova), which works in the field of production and research of glass-coated microwires.

References

1. Graves, S.: A Review of Productions Scheduling. Operations Research. 29, 646-676 (1981)
2. Sauer, J.: Knowledge-Based Systems Techniques and Applications in Scheduling. In: Leondes, T. (Ed.) Knowledge-based systems: Techniques and Applications. Academic Press, San Diego (2000)
3. Scholl, A., Boysen, N., Fliedner, M.: The Assembly Line Balancing and Scheduling Problem with Sequence-Dependent Setup Times. Operations Research Spectrum. 35(1), 291-320 (2013)
4. Uddin, M.., Lastra, J.: Assembly Line Balancing and Sequencing. In: Grzechca, W. (Ed.) Assembly Line – Theory and Practice. InTech (2011)
5. Grzechca, W.: Station Estimation in Assembly Line Balancing Problem. In: Grzech, A., Swiatek, P., Brzostowski, K. (eds) Applications of Systems Science. EXIT, Warsaw (2010)
6. Boysen, N., Fliedner, M., Scholl, A.: A Classification of Assembly Line Balancing Problems. Jenaer Schriften zur Wirtschaftswissenschaft 12/06, University of Jena (2006)
7. Boysen, N., Fliedner, M., Scholl, A.: Assembly Line Balancing: Which Model to Use when. Jenaer Schriften zur Wirtschaftswissenschaft 23/06, University of Jena (2006)
8. Kriengkorakot, N., Pianthong, N.: The Assembly Line Balancing Problem: Review articles. KKU Engineering Journal. 34(2), 134-140 (2007)
9. Chong, K., Omar, M., Bakar, N.: Solving Assembly Line Balancing Problem using Genetic Algorithm with Heuristics-Treated Initial Population. In: Proceedings of the World Congress on Engineering, London (2008)
10. Yokoyama, K., Morikawa, K., Takahashi, K.: A Modified Multi-Agent System for Simple Assembly Line Balancing. In: 8th International Conference of Modeling and Simulation: Evaluation and optimization of innovative production systems of goods and services, Hammanet, Tunisia (2010)
11. Hennessy, J., Patterson, D.: Computer Architecture: A Quantitative Approach. Morgan Kaufmann, San Francisco (2003)
12. Kogge, P.: The Architecture of Pipelined Computers. McGraw-Hill, New York (1981)
13. Stone, H.: High-Performance Computer Architecture. Addison-Wesley, New York (1993)
14. Zaporojan, S., Plotnic, C., Calmicov, I., Larin, V.: A knowledge-based approach for microwire casting plant control. In: Jozefczyk, J., and Orski, D. (Eds) Knowledge-Based Intelligent System Advancements: systemic and cybernetic approaches. Information Science Reference, Hershey New York (2011)

Applying System of Systems Engineering Approach to Build Complex Cyber Physical Systems

Lichen Zhang

1 Introduction

The scale and complexity of advanced cyber physical systems (CPSs) are steadily growing. Cyber-physical systems are changing the way we interact with the physical world. Complex cyber physical rely heavily on the interplay of dozens of individual sub-systems. CCPSs aim to dramatically improve the autonomous capabilities of a collection of individual sub-systems. When each of the system in the collection becomes entirely independent, the CPSs are typical systems of systems [1].

Thus, due to their increased size and complexity relative to real time systems, complex cyber physical systems present numerous developmental challenges. The long-term viability of complex cyber physical systems confronts these challenges through the development of new specification, modeling, design, composition, verification, and validation techniques.

System of Systems (SoS) approach [2] is heavily applied in defense applications, but is increasingly being used to non-defense domains such as air and auto transportation, healthcare, global communication networks, search and rescue, space exploration and many other System of Systems application domains. In this paper, we propose a new paradigm for specifying and modeling cyber physical systems based on system-of-systems approach. In this paper, we extend AADL in modeling dynamic continuous aspect and spatial aspect, and integrate AADL with Modelica, and formal methods to specify and model cyber physical systems based on systems of systems approach. We specify and model cyber part of cyber physical systems with, and model physical part of cyber physical systems with Modelica. We apply formal specification method in requirement analysis process in order to ensure that the software requirements model satisfies required system function and performance goals and constraints, including safety. The effectiveness of the approach is demonstrated with a case study of Vehicular Ad-hoc NETwork.

L. Zhang (✉)
Faculty of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510090, China

Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai 200062, China
e-mail: zhanglichen1962@163.com

2 System of Systems Engineering (SoSE)

Currently, numerous "System of Systems" (SoS) definitions exist. The specifics of each definition vary somewhat, but most agree that a system of systems arises when a set of needs are met through a combination of several systems. Each system can operate independently, but each also must interact effectively with other systems to meet the specified needs [3].

The systems of systems have following characteristics as shown in Fig. 1 [4]:

Autonomy—The ability of a system as part of SoS to make independent choices.

Belonging—Constituent systems have the right and ability to choose to belong to SoS.

Connectivity—The ability to stay connected to other constituent Systems.

Diversity—Evidence of visible heterogeneity.

Emergence—Formation of new properties as a result of developmental or evolutionary process.

System-of-Systems Engineering (SoSE) is a set of developing processes, tools, and methods for designing, redesigning and deploying solutions to System-of-Systems challenges. System of systems engineering "deals with planning, analyzing, organizing, and integrating the capabilities of a mix of existing and new systems into a SoS capability greater than the sum of the capabilities of the constituent parts". [6]. System-of-Systems Engineering (SoSE)focuses on integrating multiple complex systems. Integration into a metasystem, or systems of systems, may involve existing systems, newly designed systems, or a hybrid mixture. What is new is the formation of the

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_88,

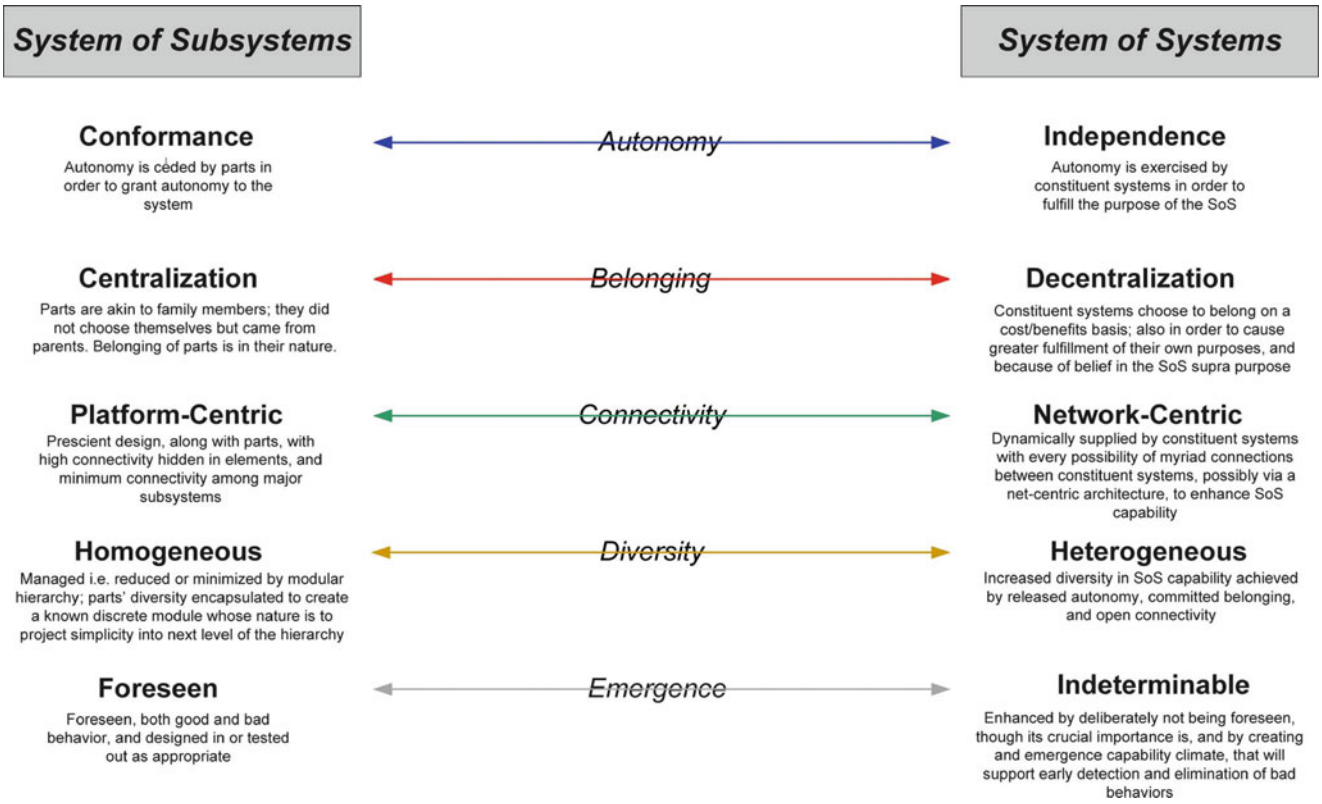


Fig. 1 Distinguishing characteristics of SoS and their opposing forces [5].

metasystem comprised of a set of integrated complex systems brought about by an emerging need or mission. This represents significant departure from traditional systems engineering single system focus [7]

Establishing requirements and measures for a system of systems is one of the most important activities for SoSE. SoSE is based on a different paradigm than systems engineering, therefore, what has served systems engineering well should be met with a healthy skepticism prior to wholesale incorporation into the SoSE domain. The nature of the SoSE problem domain suggests that requirements are simultaneously loose and tight. requirements in the SoSE problem domain can and should shift to compensate for further levels of resolution in the complex system problem domain. Requirement resolution should increase with additional understanding of the complex SoS problem domain and emergent conditions. [8, 9].

Current specification and modeling methods for a system-of-systems typically consist of little more than a "sticks-and-circles" diagram. The "circles" represent the various systems that comprise the system-of-systems while the "sticks" are means of information transfer, a messaging protocol, and, perhaps, a translator box to translate the messaging format from one system to another, armed with this sophomoric view of the system-of systems. Traditionally, these methods failed to achieve an interoperable and integrated system-of-systems [10].

There are many benefits that can be realized when applying a model based approach compared to a more document-centric approach. These include [11]:

- Reduced development time, due to increased automation, consistency checking and traceability analysis.
- Enhanced analysis, due to the ability to automate trade studies, impact analyses and to perform simulations.
- Increased re-uses, when used in the same context, then elements of a MBSE system may be re-used.

Models is an abstraction in which the complexity of a real system is reduced by selecting only the most significant system effects in order to understand the behaviours of the system [12]. Models could be horizontally integrated in the sense of combining different types of models to represent different discipline based effects [13]. Currently, this is only beginning to address the complexity that will be required for effective prediction. Of course, modelling approaches for SoS exist. A good model should incorporate judicious trade-off between realism and simplicity. [14, 15].

The object-oriented paradigm offers a new system-of-systems requirements and design methodology that provides for both minimizing accidental complexity and controlling essential complexity through the use of decentralized control flow, minimal messaging between classes, implicit case analysis, and information-hiding mechanisms. [16].

3 The Proposed Method for Specification and Modeling of Cyber Physical System Based on Systems of Systems Approach

One class of systems of systems with the additional challenge of integrating different system types are cyber-physical systems (CPS)[17]. Specifying and modeling a set of CPSs as systems of systems is beneficial where sub-systems exhibit traits of managerial and/or operational independence. With some groups of cyber-physical sub-systems, a subset of sub-systems may be useful outside the context of the entire collection. Similarly, a subset of sub-systems is able to satisfy some useful goal outside of the entire collection. Moreover, evolutionary development and emergent behavior are both common traits of cyber-physical sub-systems exhibiting behaviors of self-organisation. In order to model such characteristics using systems of systems approach, subsets of sub-systems can be modeled as constituents of an overall system. This approach is useful in that it can be applied to give indication of structural approach alternatives, and emphasizes possible interaction behavior between constituents.

In this paper, we proposed a design methodology for cyber physical systems based on Systems of System approach as shown in Fig. 2. First, The requirements of cyber physical systems are specified as either systems-of-systems requirements which are properties of the overall system-of systems that are described using the capabilities of the constituent systems, or constituent system-level requirements which are allocated to particular constituent system(s). These types of requirements are constantly changing and this makes the partitioning very difficult. This means the traditional concept of static, signed-of requirements is not suitable for systems-of-systems. New requirements engineering processes, management methods, techniques and tools that can dynamically respond to unstable, fragmented, continually changing requirements are needed. The proposed methods and tools should not only be able to deal with problems that are associated with influence requirements, cascades effects, epidemics but also be able to handle problems associated with partitioning a large-scale system-of-systems into multiple autonomous independently evolving constituent systems.

Second, in this paper we proposes the description of CPS architecture based on the research and understanding of CPS in multi-dimension, multi-domain, multi-view. We first divide CPS into three dimensions: physical world dimension, communication dimension and cyber computation dimension.

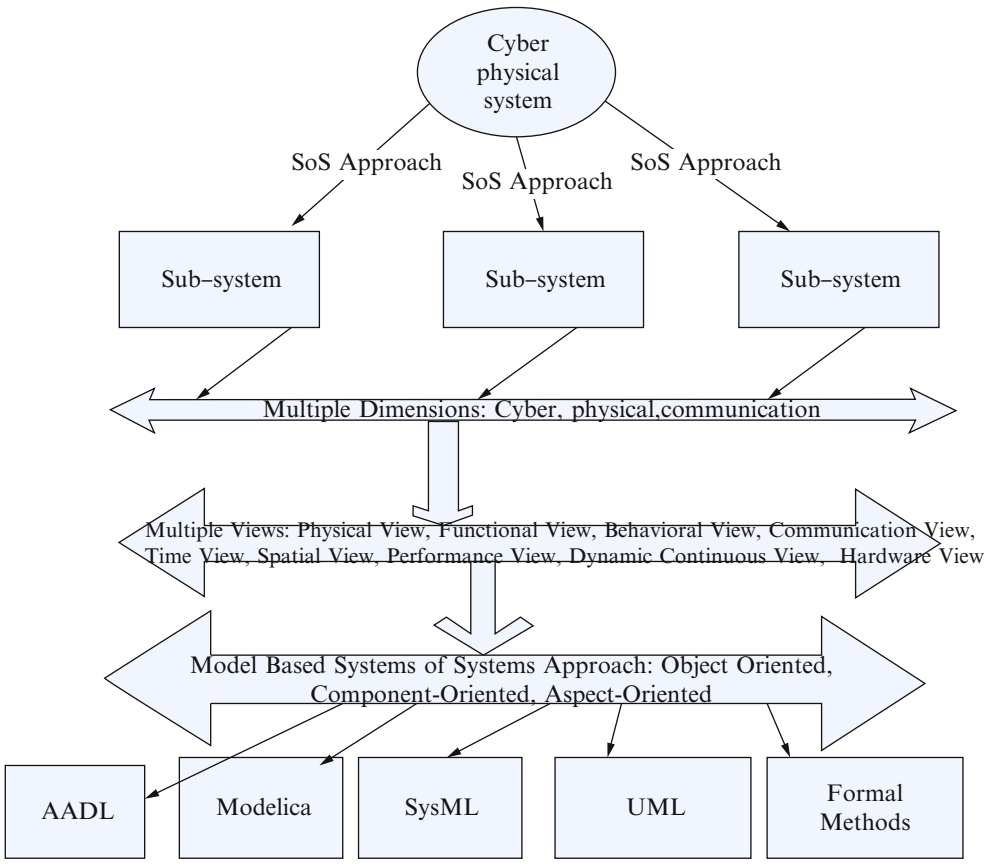


Fig 2 The proposed methodology for cyber physical system development base on Systems of Systems approach

Third, in this paper, we propose a view oriented approach to integrated different models to specify and model cyber physical systems as a whole Using multi-view modeling, it is possible to describe different aspects of cyber physical system being designed and thus support system development process phases such as requirements analysis, design, implementation, verification, validation and integration. Especially, multi-view modeling approach supports mathematical modeling with equations since equations specify behavior of physical world of aerospace cyber physical system. In the specification, analysis, design and implement of cyber physical systems, multi-views on the system to be developed are often used. These views typically composed of models in different formalism. Different views usually are suitable to various partial aspects of the complex aerospace cyber physical systems in a multi-view approach, individual views are simpler than a single model representing all aspects of the system. As such, multi-view modeling, like modular, hierarchical modeling, simplifies system development. Most importantly, it becomes possible for individual experts on different aspects of a system develop to work in isolation on individual, possibly domain specific views without being encumbered with other aspects. These individual experts can work mostly independently, thereby considerably speeding up the development process.

Finally, in this paper, we propose an approach to support specification and modeling cyber physical systems based on SOS engineering in the well established modeling language AADL [18 [19]]. The proposed SoS engineering specification and modeling method of cyber physical systems consists of a sequence of steps:

1). Development of system of systems scenarios and operational architectures
2). Identification of system of systems threads
3). Representation of operational architectures in AADL
4). Identification of system of systems design parameters and factor levels
5). Transformation of AADL format representation into executable models
6). Application of design of experiments
7). Simulation runs and analysis of results

In this paper, we integrate AADL [20] with Modelica [21], UML [22], and formal methods to specify automotive cyber physical systems using systems of systems approach. We specify cyber part of cyber physical systems with AADL and UML, and model physical part of cyber physical systems with Modelica. Modelica is an object-oriented equation based programming language, oriented towards computational applications with high complexity requiring high performance. Integrating the descriptive power of AADL models with the analytic and computational power of Modelica models provides a capability that is significantly greater than provided by AADL or Modelica individually. AADL and Modelica are two complementary languages supported by two active communities. By integrating AADL and Modelica, we combine the very expressive, formal language for differential algebraic equations and discrete events of Modelica with the very expressive AADL constructs for requirements, structural decomposition, logical behavior and corresponding cross-cutting constructs.

In this paper, we apply formal specification method in requirement analysis process in order to ensure that the software requirements model satisfies required system function and performance goals and constraints, including safety.

4 Case Study: Specification and Modeling of VANET by the Proposed Method

Automotive Cyber-Physical Systems (ACPS) have attracted a significant amount of interest in the past few decades. These networks are also known as Vehicular Ad Hoc Networks (VANET) [23]. VANE is a typical Systems of Systems. VANET has become an active area of research, standardization, and development because it has tremendous potential to improve vehicle and road safety, traffic efficiency, and convenience as well as comfort to both drivers and passengers. [24]

Fig. 3 [25] shows the system architecture of VANET from the perspective of the different components and domains as well as their interactions.

In this paper, we specify the requirements of Vehicular Ad-hoc NETwork (VANET) base on AADL [30], Modelica and formal methods..

The VANET and the sub-systems is modeled by AADL as shown in Fig. 4.

Vehicle station system is one of the sub-systems of VANE, Fig. 5 represents Vehicle station system model in AADL.

The engine is the subsystem of vehicle station system, and The engine is physical system. We model the engine using Modelica. The vehicle dynamics library of Modelica is used to model vehicle. Fig. 6 represents the engine model in Modelica.

We suppose that the street is unidirectional. Cars run on the street. No cars can run with another side by side but the moment of overtaking another car. Cars can accelerate and brake. Control system collects the information of cars and sends the command to cars via signal system described by CSP [27]. Signal system has four parts: CarReceive, CarSend, CtrlReceive and CtrlSend. The CarSend part collects car's data. CtrlReceive part receives data from CarSend part. CtrlSend part sends command from control

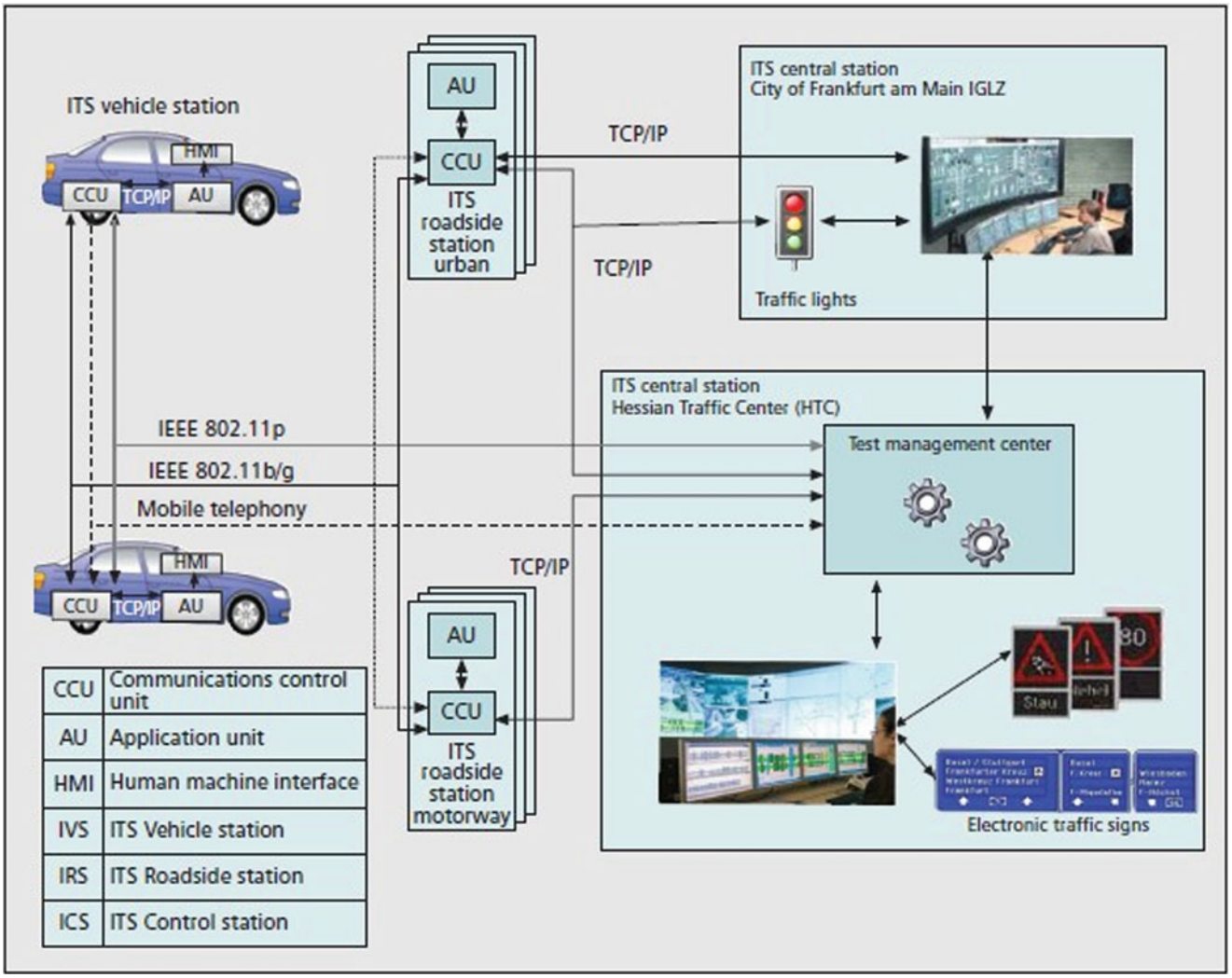


Fig. 3 The Architecture of VANET

system. CarReceive part receives command from CtrlSend part. We introduce the variables needed: x is the position of the car, x is a scalar; v is the velocity of a car and a vector; a is the acceleration of a car and a vector; ID is the identification number of a car, which is assigned by the car's manufacturers and different from each other; state is the current state of a car.Control system has five scenes to deal with. There is a standard time interval t and a length l. There is a length variable l. If two cars length is longer than l and the car can outstrip the one just in front of it within t, change the state of this car to ACCELERATION. If the distance between two cars is longer than I, keep the cars speed v to its current value and s to 0. If the distance between two cars is shorter than I and the car can't outstrip the one just in front of it within t, change the state of this car to DECELERATION. We specify the control system by DL [28], as in (1).

$$
\begin{aligned}
ControlSystem \equiv & (?x_1 - x_2 < l \& s_1 - s_2 > x_1 - x_2; state = ACCELERATION; \\
& a = b; v\prime = a; x\prime = v) \\
& \cup \\
& (?x_1 - x_2 < l \& s_1 - s_2 < x_1 - x_2; state = DECELERATION; \\
& a = b_2; v\prime = a; x\prime = v; ?v > 0) \\
& \cup \\
& (?x_1 - x_2 \geqslant l; state = CONSTANTSPEED; a = 0;) \\
& \cup \\
& (?state = SETOUT; a = b; v\prime = a; x\prime = v) \\
& \cup \\
& (?state = PARKING; a = -b; v = v - at; v >= 0)
\end{aligned} \tag{1}
$$

The Signal System works in this way: process Car sends its data to process of CarSend; process CarSend sends this data to CtrlReceive; CtrlReceive sends this data to ControlSystem. After computing on this data, ControlSystem sends its command for the car, which is also data like: a, v, state and x. The sequence is:

ControlSystem sends its command to process CtrlSend; CtrlSend sends its command to process CarReceive; CarReceive sends this data to Car. We models the processing of a car with CSP, which accepts a call and sends the information out, as in (2). We model the delivering the command that the length is 5 to the car, as in (3). We specify the receiving the command that the length is 5 from control system, as in (4). We models the control system, as in (5). We model the receiving the commands of control system as in (6). We specify the receiving information of car, as in (7).

The formulas are below:

$$Car = left?call \rightarrow right!x \rightarrow right!a \rightarrow right!v \rightarrow right!ID \rightarrow right!state \rightarrow Car \tag{2}$$

$$\begin{aligned} CarSend = & P\langle\rangle \\ & where \\ & \quad P_s = right!s \rightarrow P\langle\rangle if \#s = 5 \\ & \quad P_s = left?x \rightarrow P_{s^\frown\langle x\rangle} \end{aligned} \tag{3}$$

And, x means the variable from Car, that is, x, a, v, ID and state.

$$\begin{aligned} CarReceive = & P\langle\rangle where \\ & P\langle\rangle = left!s \rightarrow P\{s\} \\ & P\langle x\rangle = right!x \rightarrow P\langle\rangle \\ & P\langle x\rangle^{smallfrown} s = right!x \rightarrow P_s \\ & left = \{s | s \in right^* \# s = 5\} \end{aligned} \tag{4}$$

$$\begin{aligned} ControlSystem = & left?x \rightarrow left?a \rightarrow left?v \rightarrow left?ID \rightarrow left?state \rightarrow \\ & right!call \rightarrow right!s \rightarrow ControlSystem \\ & where \#s = 5 \end{aligned} \tag{5}$$

$$CtrlSend = left?call \rightarrow left?s \rightarrow right!call \rightarrow right!sCtrlSend \tag{6}$$

$$CtrlReceive = left?call \rightarrow left?s \rightarrow right!call \rightarrow CtrlReceive \tag{7}$$

5 Conclusion

Due to their increased size and complexity relative to real time systems, complex cyber physical systems present numerous developmental challenges. The long-term viability of complex cyber physical systems confronts these challenges through the development of new specification, modeling, design, composition, verification, and validation techniques. In this paper, we proposed a new paradigm for specifying and modeling cyber physical systems based on system-of-systems approach. In this paper, we extended AADL in modeling dynamic continuous aspect and spatial aspect, and integrated AADL with Modelica, and formal methods to specify and model cyber physical systems based on systems of systems approach. We specified and model cyber part of cyber physical systems with, and modeled physical part of cyber physical systems with Modelica. We applied formal specification method in requirement analysis process in order to ensure that the software requirements model satisfies required system function and performance goals and constraints, including safety. The effectiveness of the approach was demonstrated with a case study of Vehicular Ad-hoc NETwork.

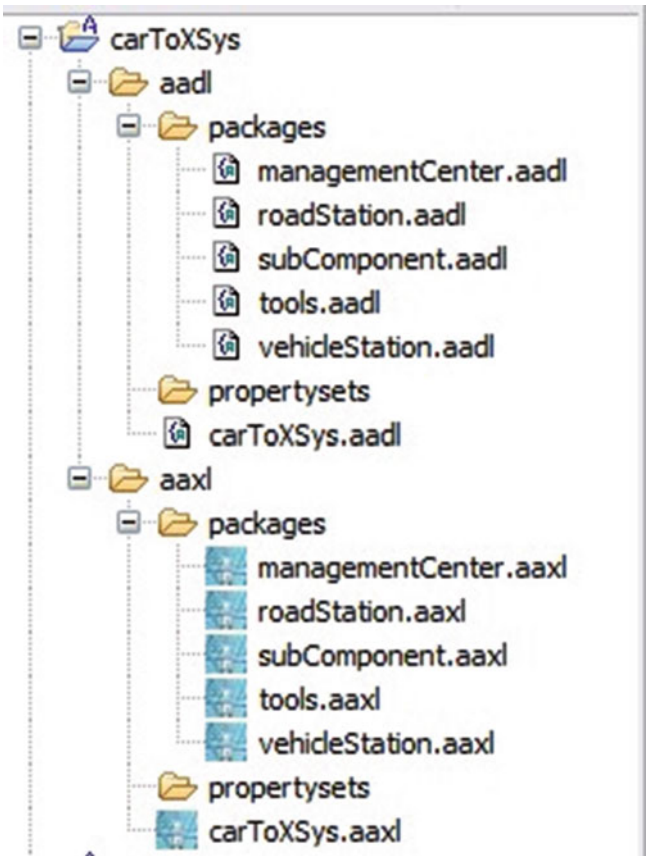


Fig. 4 The VANET and the sub-systems model in AADL

Five characteristics of Systems of systems are named autonomy, belonging, connectivity, diversity, and emergence, whose degrees of strength determines the foundation of any SoS. Future work will focus on specifying and modeling these characteristics of Systems of Systems, particularly, we shall study the five characteristics specification and modeling method of cyber physical systems.

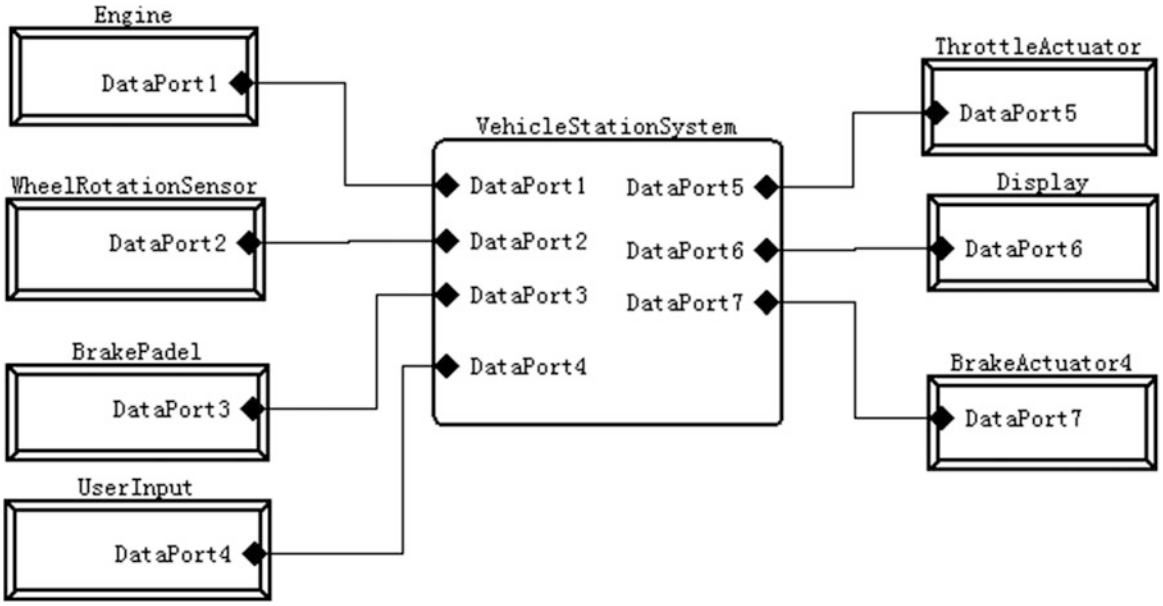


Fig. 5 Vehicle Station system model in AADL

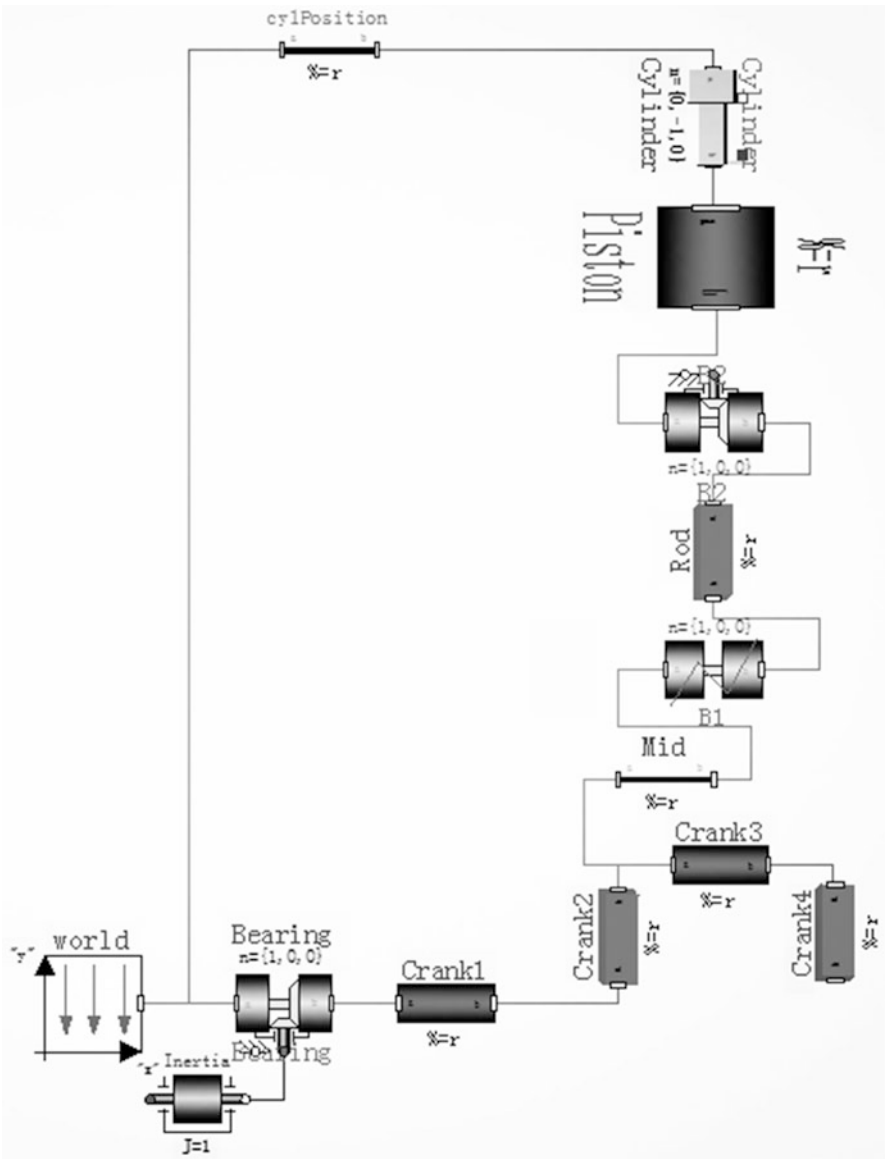


Fig. 6 The Modelica model of engine

Acknowledgments This work is supported by the national natural science foundation of China under grant (No.61370082, No.61173046), natural science foundation of Guangdong province under grant (No.S2011010004905). This work is also supported by Shanghai Knowledge Service Platform Project (No.ZF1213)

References

1. D. Luzeaux & JR Ruault. Systems of Systems, ISTE Ltd and John Wiley & Sons Inc, 2010
2. Ireland, V. and Ooi-Sanches, Y. Recognising further concepts from complex systems research in SoSE, the Proceedings of 2013 8th International Conference on System of Systems Engineering (SoSE),. Hawaii, USA, p 93-98, Date 2-6 June 2013
3. D.A. DeLaurentis and W.A. Crossley, "A Taxonomy Bases Perspective for System of Systems Design Methods", *IEEE International Conference on Systems, Man and Cybernetics*, 2005
4. Alex Gorod, Brian Sauser, and John Boardman. System-of-Systems Engineering Management: A Review of Modern History and a Path Forward, IEEE SYSTEMS JOURNAL, VOL. 2, NO. 4,p 484-499, DECEMBER 2008
5. B. Sauser, J. Boardman, and A. Gorod, "SoS management," in *System of systems engineering—Innovations for the 21st Century*, M. Jamshidi, Ed. Hoboken, NJ: Wiley, 2008.
6. 'Systems Engineering Guide for Systems of Systems' Office of the Under Secretary of Defense, USA DoD, August 2008
7. C. Keating, R. Rogers, R. Unal, D. Dryer, A. Sousa-Poza, R. Safford, W. Peterson, and G. Rabadi, "System of systems engineering," *EMJ – Engineering Management Journal*, vol. 15, pp. 36-45, 2003.
8. Keating CB, Padilla JJ, Adams K. System of systems engineering requirements: challenges and guidelines. Eng Manag J 20:24–31
9. Keating CB, Katina PF.System of systems engineering: prospects and challenges for the emerging field. Int J Syst Syst Eng 2:234–256, 2011
10. Dale S. Caffall1 and James B. Michael2. A New Paradigm for Requirements Specification and Analysis of System-of-Systems, RISSEF 2002, LNCS 2941, pp. 108–121, 2004.
11. Jon Holt, Simon Perry, Mike Brownsword, Daniela Cancila, Stefan Hallerstede, Finn Overgaard Hansen: Model-based requirements engineering for system of systems. SoSE 2012: 561-566
12. Fitzgerald JS, Larsen PG, Woodcock JCP. Foundations for Model-based Engineering of Systems of Systems. In: Complex Systems Design & Management. 2014, Paris: Springer International Publishing
13. Fitzgerald JS, Bryans JW, Payne RJ. A Formal Model-based Approach to Engineering Systems-of-Systems. In: Collaborative Networks in the Internet of Services. 2012, Bournemouth, UK: Springe
14. Trans-Atlantic Research and Education Agenda in Systems of Systems (T-AREA-SoS). Loughborough University © 2013. https://www.tareasos.eu/docs/pb/SoA_V3.pdf
15. TAREA-PU-WP5-R-LU-26, Issue 2, Release: 15th August 2013, Loughborough University © 2013. https://www.tareasos.eu/docs/pb/SRA_Issue2.pdf
16. Frederic Loiret, Romain Rouvoy, Lionel Seinturier. Software Engineering of Component-Based1Systems-of-Systems: A Reference Framework. 14th International ACM SIGSOFT Symposium on Component-Based Software Engineering, (CBSE'11) (2011) 61-65
17. M Mansfield, J Fitzgerald. Modelling Systems of Cyber-Physical Systems. wiki.overturetool.
18. SAE AS-2C. Architecture Analysis & Design Language. SAE International Document AS5506B(2012) Revision 2.1 of the SAE AADL standard, Sept 2012.
19. Feiler P H, Gluch D P, Hudak J J. The architecture analysis & design language (AADL): An introduction[R]. CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST, 2006
20. Dionisio de Niz and Peter H. Feiler. Aspects in the industry standard AADL. AOM '07 Proceedings of the 10th international workshop on Aspect-oriented modeling.P15 – 20
21. Modelica - a unified object-oriented language for physical systems modelling. Language specification. Technical report, Modelica Association, 2002.
22. Selic, B., Rumbaugh, J.: Using UML for modeling complex real-time systems.(1998). [Online]. Available: http://www.objectime.com/uml
23. Ghassan M. T. Abdalla, Mosa Ali Abu-Rgheff, Sidi Mohammed Senouci, "*Current Trends in Vehicular Ad Hoc Networks*", Ubiquitous Computing and Communication Journal, p.p.1-9, 2007
24. Sherali Zeadally et al. Vehicular ad hoc networks (VANETS): status, results, and challenges, Telecommunication Systems, Volume 50, Issue 4, pp 217-241, August 2012
25. Hagen Stubing,Adam Opel Gmbh Marc Bechler.simTD: A Car-to-X System Architecture for Field Operational Tests[J]. IEEE Communications Magazine. 2010,48(5):148-154.
26. Hudak J J, Feiler P H. Developing aadl models for control systems: A practitioner's guide[J]. 2007.
27. Reed G M, Roseoe A W. A timed model for communicating sequential processes[C]//Pro ICALP'86. Lecture Notes in Computer Science. Berlin:Springer, 1986
28. A. Platzer. Logical Analysis of Hybrid Systems: Proving Theorems for Complex Dynamics. Springer, 2010. 426 p. ISBN 978-3-642-14508

Model Integration and Model Transformation Approach for Multi-Paradigm Cyber Physical System Development

Lichen Lichen

1 Introduction

Cyber-Physical Systems (CPS) are next generation of engineered systems in which physical systems and cyber systems not only are converged, but also computing, communication, and control technologies are tightly integrated [1]. Cyber physical system is hard to develop because developers need to consider functional properties, non-functional properties, such as timeliness, energy, memory, safety and reliability, dynamic continuous properties, spatial requirements and the interaction with physical world [2]. The lack of specification and modeling methods and techniques has a very important impact, as the increasing complexity of cyber-physical systems built today pushes traditional development processes to their limits [3]. In these development processes the different aspects and disciplines of mechanics, electrics, and software usually act isolated from each other, which inhibits taking advantage of the full potential of integrated solutions. One approach to overcome this separation of engineering disciplines is an integrated abstract model that serves as a common language for specification and analysis of the cyber physical system. One of the fundamental challenges in research related to cyber-physical system is accurate modeling and representation of these systems. The main difficulty lies in developing an integrated model that represents both cyber and physical aspects with high fidelity. In cyber-physical system, the physical world aspect of the system resides in the continuous and continual domain. Thus, on the physical world side of cyber-physical systems we must contend with not only real-time requirements but also requirements related to the continuous and continual nature of cyber physical systems [4].

Modeling a system using different multi- multi-disciplinary leads to the concept of model transformation [5]. Model transformations are mappings of one or more models into one or more target models. The models are graphical or textual. Model transformation is a central concept in model-driven development approaches and integration development approaches in cyber physical systems, as it provides a mechanism for automating the manipulation of models. Model transformations provide a mechanism for automatically creating or updating target models based on information contained in existing source models [6].

In this paper, we propose an integrated approach to develop cyber physical systems based on multi-dimensions, multi-views and multi-paradigm. This model-integrated development approach addresses the development needs of cyber physical systems through the pervasive use of models. We present the model transformation methods of cyber physical systems; we propose an approach to transform the models of among AADL [7], Modelica [8], SysML [9] and formal methods, clarify the transformation principles and to illustrate the important synergies resulting from the integration between these modeling languages.

2 Integration Approach to Specify and Model Cyber Physical Systems

Cyber physical systems are becoming increasingly complex to design because of distributed and networked large applications and highly integrated products encompassing various engineering domains such as mechanical, electrical and chemical domains. To manage this complexity, integrated development approach can be applied; integrated development approach is an interdisciplinary approach to creating and verifying an integrated set of system solutions to satisfy system development needs.

L. Lichen (✉)
Faculty of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510090, China

Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai 200062, China
e-mail: zhanglichen1962@163.com

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_89,

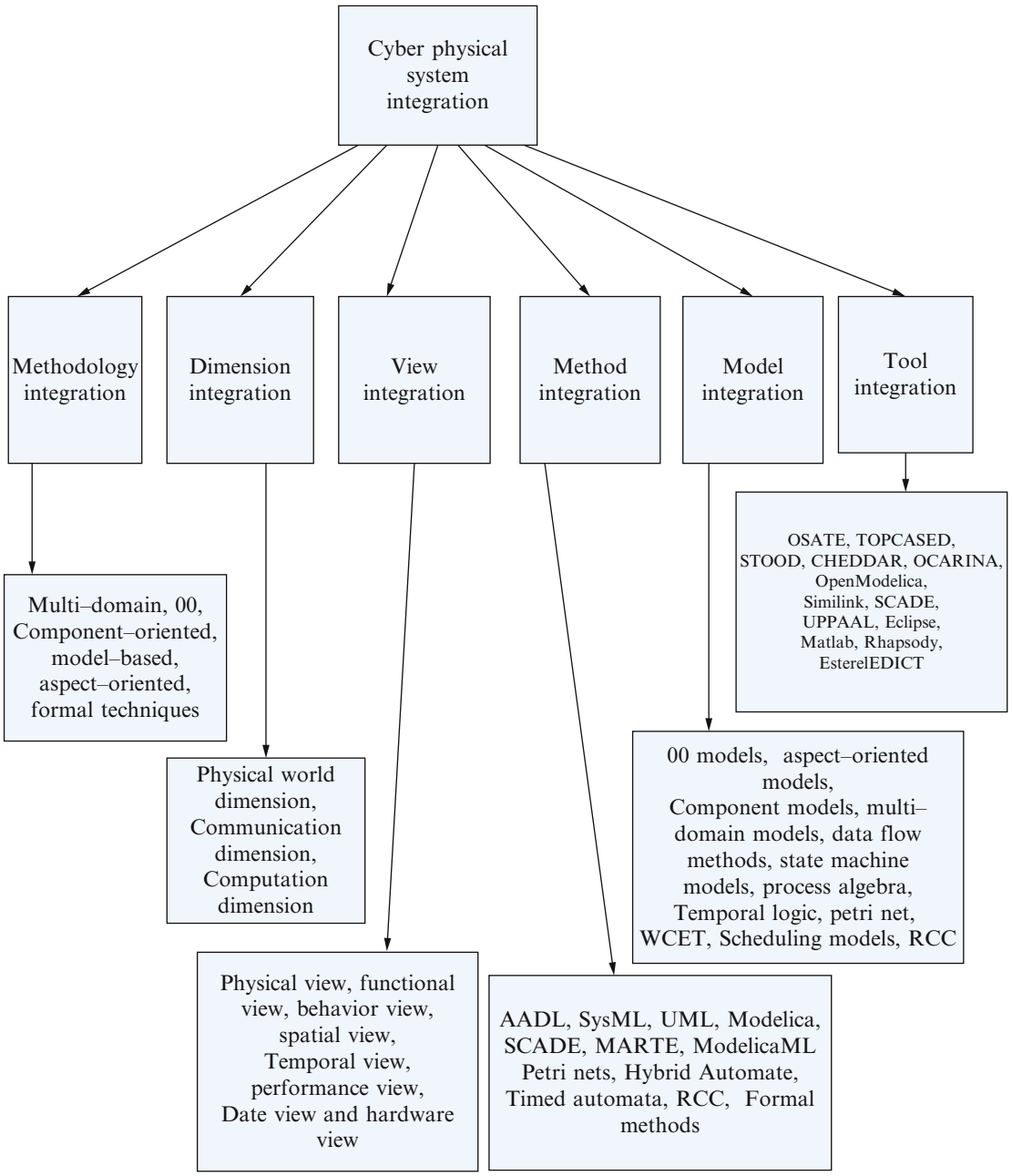


Fig.1 The multi-disciplinary, multi-Domains, multi-dimensions, multi-View, multi-paradigm integrated approach to develop CPS

Integration based development enables development cycle compression by reuse of existing methodologies, methods, models, tools and techniques [10], encapsulated in integratable and customizable models and components that can be rapidly used in a design. Components and Models in cyber physical systems are heterogeneous, span multiple domains (physical – thermal, mechanical, electrical, fluid, .. and cyber–software, computing platforms), and require multiple models to soundly represent physical aspect, the requirements, architectures, behavior, spatio-temporal constraints , and interfaces, at multiple levels of abstractions. We advocate a multi-disciplinary, multi-dimension, multi-domain and multi-paradigm approach to the development of cyber physical systems, based on the identification of complementary dimensions and views on the system as starting point for modeling and design as shown in Fig.1. The integrated approach allows creating multi-model multi-domain representation of cyber physical system models and components that are composable by system development.

The objective of our work is to propose an integration method, which supports a single design methodology that

covers all phases of the life cycle, ensuring that whole requirements of cyber-physical systems will be met, even on network and distributed architectures **as** target. A key issue in current networked cyber physical system development is the desire to integrate various objects oriented analysis and design methods and tools, aspect-oriented development methods and tools, multi-domain physical modeling methods and tools, and formal methods that address different aspects of the development process of cyber-physical systems. Typically, different authors, each of whom has chosen to focus on a specific part of the overall problem, produce these methods and tools. Users of these methods and tools want them to work together and fully support the user's design and development process. Integrated approach is intended to provide an approach that support the entire system development life cycle. Specification, modeling and design method integration of cyber physical systems actually involves many aspects of integration and different levels:

- The integration of physical world dimension, communication dimension and computation dimension
- The integrated object-oriented methodology, multi-domain methodology, aspect-oriented methodology and formal techniques
- The integration of different design views
- The integration of the methods used to specify and implement systems requirements
- The integration of tools that support these methods
- The integration of physical components and cyber components
- The integration of different representations
- The integration of the multiple specification fragments produced by applying these methods and tools
- Integration between informal specification methods and formal specification methods is desired

3 Model Transformation Methods for Cyber Physical Systems

A model can tell what something does (specification) as well as how the function is accomplished (implementation). These aspects should be separated in modeling. It is important to get what correct before investigating much time in the how [11].

Kleppe et al. [12] give the following definition of model transformation. A transformation is the automatic generation of a target model from a source model [13], according to a transformation definition. A transformation definition is a set of transformation rules that together describe how a model in the source language can be transformed into a model in the target language. A transformation rule is a description of how one or more constructs in the source language can be transformed into one or more constructs in the target language. A horizontal transformation is a transformation where the source and target models reside at the same abstraction level. A vertical transformation is a transformation where the source and target models reside at different abstraction levels [14].

A model is a simplified representation of a system and a model helps to obtain a better understanding of the system. Models are often expressed in dedicated domain-specific languages or general purpose modeling languages or multi domain languages such as Modelica. A metamodel of a model X describes the structure that model X must follow to be valid. A metamodel can be compared to a grammar in language design. Precisely defined metamodels are a prerequisite for model transformations [15].

We propose a model transformation approach for cyber physical systems as shown in Fig.2.

OMG SysML™ [16] is a general-purpose systems modeling language that enables systems engineers to create and manage models of engineered systems using well-defined, graphical constructs. SysML is capable of representing the specification, analysis, design, verification and validation of engineered systems. Modelica [17] is an object-oriented language for describing differential algebraic equation (DAE) systems combined with discrete events. Such models are ideally suited for representing the flow of energy, materials, signals, or other continuous interactions between system components. By integrating SysML and Modelica, we combine the very expressive, formal language for differential algebraic equations and discrete events of Modelica with the very expressive SysML constructs for requirements, structural decomposition, logical behavior and corresponding cross-cutting constructs [18]. Fig.3 represents the different component connection model of the aircraft controller in Modelicaml [19] which is the extension of SysML to make model transformation between SysML and Modelica.

The different component connection model of the aircraft controller in Modelica is as follows:

```
within ModelicaMLModel.Simulations;
model ControllerConnectedPI
    ControlColumnController;
    RuddelPanelController;
    Calculator;
    Interface;
    ModelicaMLModel.Simulations.Components.
ControlColumnController ControlColumnController;
    ModelicaMLModel.Simulations.Components.
RudderPanelController RudderPanelController;
    ModelicaMLModel.Simulations.Components.Calcula-
tor Calculator;
    ModelicaMLModel.Simulations.Interfaces.
ReadSignalOut Sensors;
```

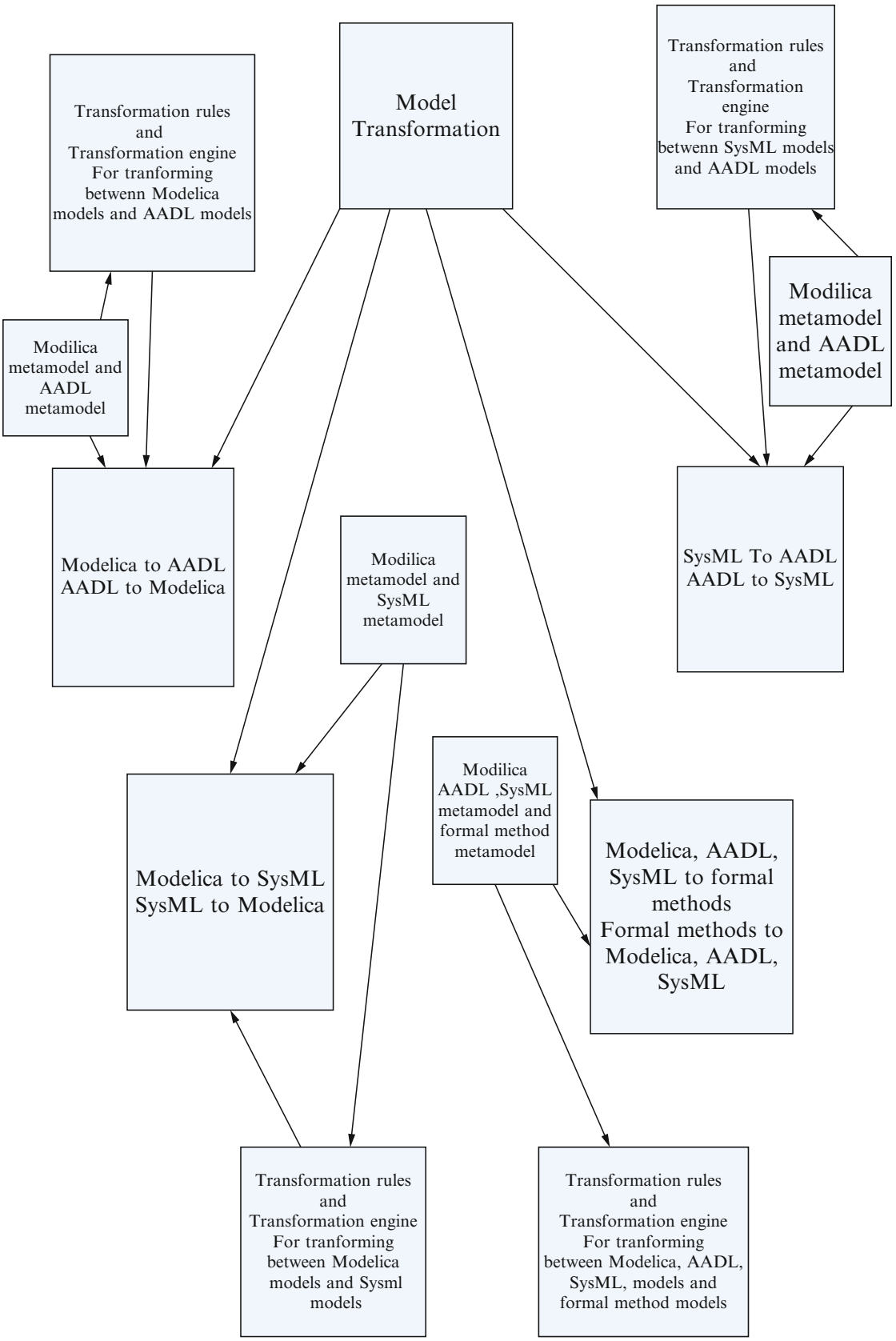


Fig.2 model transformation approach for cyber physical systems

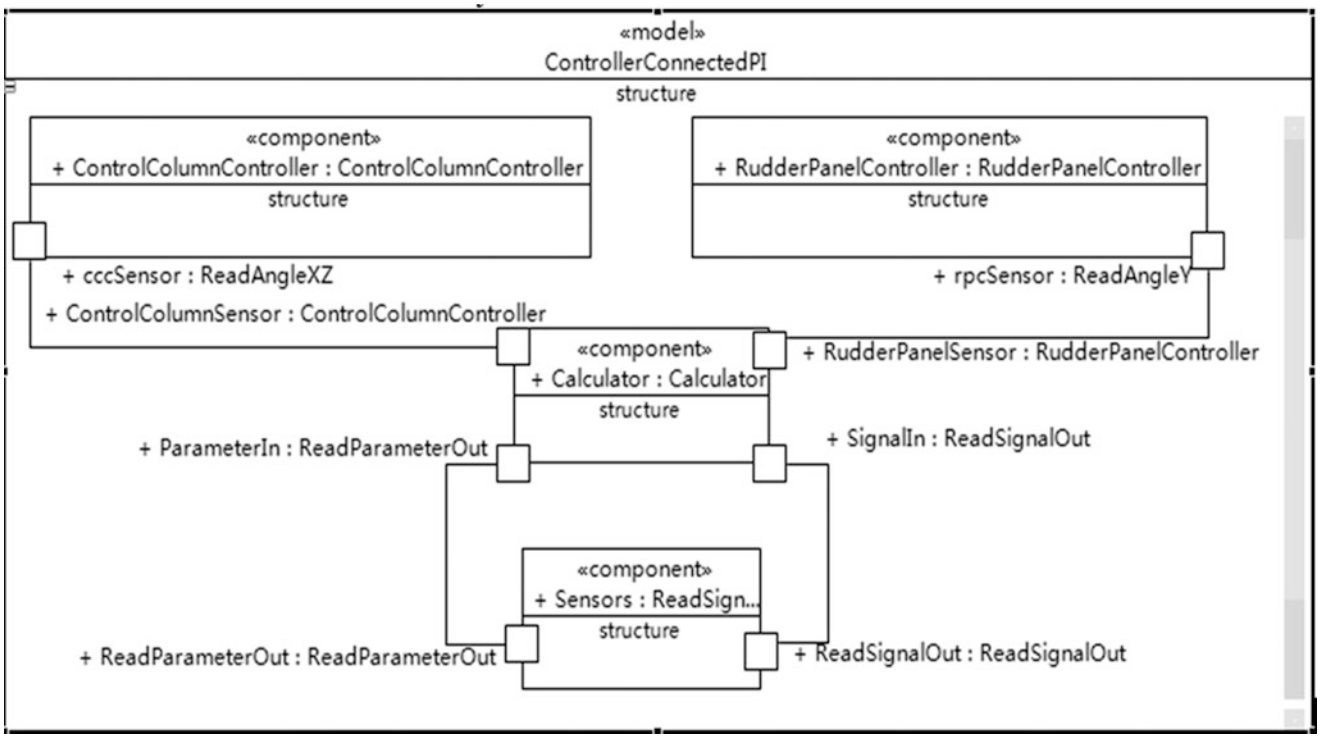


Fig.3 Controller Connnection Diagram

Equation

connect(Calculator.ControlColumnSensor. cccSensor, ControlColumnController.cccSensor);

connect(Calculator.RudderPanelSensor.rpcSensor, RudderPanelController.rpcSensor);

connect(Calculator.ParameterIn, Sensors. ReadParameterOut);

connect(Calculator.SignalIn, Sensors. ReadSignalOut);

end ControllerConnectedPI;

The Architecture Analysis & Design Language (AADL) [20] is an architecture description language standardized by SAE. AADL is used to model the software and hardware architecture of an embedded, real-time system. Due to its emphasis on the embedded domain, AADL contains constructs for modeling both software and hardware components . This architecture model can then be used either as a design documentation, for analyses (such as schedulability and flow control) or for code generation.The main components in AADL are divided into three parts: software components, hardware components and composite components. Software components include data, thread, thread group, process and subprogram. Hardware components include processor, memory, bus and device. Composite components includesystem.

AADL and Modelica both are object oriented modeling languages, AADL components can be projected into Modelica classes. After the analyzing of AADL components and Modelica classes, a transformation table is listed in Table 1. The keywords in AADL are different from those in Modelica. So, we need to transform the keywords in AADL into Modelica.

Table 1 Transformation rules between Modelica models and AADL models

Modelica Models	AADLModels
package	package
class, model	system, device, memory, processor
nested class	subcomponent
connector	port, portgroup
equation	Property extension
variables	parameter
connect()	connection
function	parameter, subprogram
constant	Property extension
initial value	Property extension
basic datatye, record, array, Modelica.SIunits	datatype

We extend AADL property [20] to transform the model. ica model into AADL model:

Property Modelica_property is

Equation : aadltring applies to (device/system/memory/processor);

Const : aadltring applies to (device/system/memory/processor);

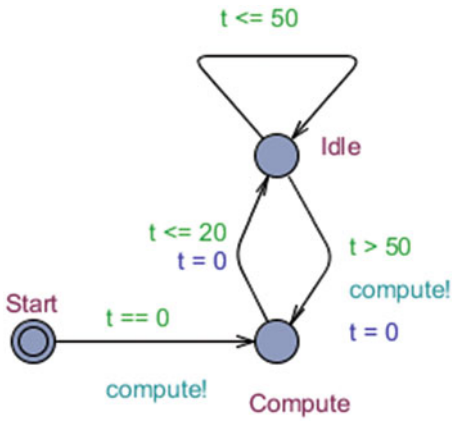


Fig.4 time automata of mode compute in UPPAAL

Const_value : aadltring applies to (device/system/memory/processor);

end Modelica_property;

UPPAAL [21] is a formal technique for modeling, analysis, validation and verification of real-time and embedded systems. Uppaal uses a network of timed automata extended with discrete variables to model, analyze and verify software systems. We can transform AADL models to UPPAAL models, and analyze and verify AADL with UPPAAL.

The behavior of flight management computation is expressed by the behavior annex of AADL as follows:

System implementation FMC.imp

Annex behavior{**

States

Start:initial state; Compute, Idle:state;

Transitions

Start-[t == 0]-> Compute;

Compute-[t < = 20] -> Idle;

Idel-[t < = 50]-> Idle;

Idle-[t > 50] -> Compute(t = 0);

**};

End FMC.imp

We transform above AADL model in to UPPAAL Model as shown in Fig. 4

4 Conclusion

Cyber physical systems are becoming increasingly complex to design because of distributed and networked large applications and highly integrated products encompassing various engineering domains such as mechanical, electrical and chemical domains. To manage this complexity, integrated development approach is applied in this paper; integrated development approach is an interdisciplinary approach to creating and verifying an integrated set of system solutions to satisfy system development needs. Our integration based development enables development cycle compression by reuse of existing methodologies, methods, models, tools and techniques, encapsulated in integratable and customizable models and components that can be rapidly used in a design. Components and models in cyber physical systems are heterogeneous, span multiple domains (physical – thermal, mechanical, electrical, fluid, .. and cyber–software, computing platforms), and require multiple models to soundly represent physical aspect, the requirements, architectures, behavior, spatio-temporal constraints , and interfaces, at multiple levels of abstractions.

In this paper, we propose an integrated approach to develop cyber physical systems based on multi-disciplinary, multi-domain, multi-dimensions, multi-views and multi-paradigm. This model-integrated development approach addresses the development needs of cyber physical systems through the pervasive use of models and aims to deal with complexity due to the convergence of different domains and technologies in cyber physical systems. This proposed approach requires models for very different domains able to work together. Modeling a system using different multi-multi-disciplinary leads to the concept of model transformation. Model transformations are mappings of one or more models into one or more target models. Model transformation is a central concept in model-driven development approaches and integration development approaches in cyber physical systems, as it provides a mechanism for automating the manipulation of models. We present the model transformation methods of cyber physical systems, we propose an approach to transform the models of among AADL, Modelica, SysML and formal methods, clarify the transformation principles and to illustrate the important synergies resulting from the integration between these modeling languages.

Since the application domains of model transformation technology are very diverse. Therefore, there is no unique answer to the question which approach to model transformation is the best. In future work, we will concentrate on transformation of AADL, Modelica and SysML models into the models of formal methods. Model-driven development should be applied to model transformations, with verification of the correctness and consistency of the transformations being carried out as an integral part of such development.

Acknowledgment This work is supported by the national natural science foundation of China under grant (No.61370082, No.61173046), natural science foundation of Guangdong province under grant (No.S2011010004905). This work is also supported by Shanghai Knowledge Service Platform Project (No.ZF1213)

References

1. David Broman, Edward A. Lee, Stavros Tripakis, and Martin Törngren. Viewpoints, Formalisms, Languages, and Tools for Cyber-Physical Systems, in Proceedings of the 6th International Workshop on Multi-Paradigm Modeling (MPM'12), Innsbruck, Austria, October, 2012.
2. Manfred Broy: Cyber Physical Systems (Part 2). it - Information Technology 55(1): 3-4 (2013
3. Manfred Broy: Challenges in modeling cyber-physical systems. IPSN 2013: 5-6
4. John Eidson, Edward A. Lee, Slobodan Matic, Sanjit A. Seshia, Jia Zou. Distributed Real-Time Software for Cyber-Physical Systems, *Proceedings of the IEEE (special issue on CPS)*, 100(1):45-59, January 2012
5. M. Biehl. Literature Study on Model Transformations. Technical Report ISRN/KTH/MMK/R-10/07-SE, Royal Institute of Technology, July 2010.
6. K. Czarnecki and S. Helsen. Feature-based survey of model transformation approaches. IBM Systems Journal, special issue on Model-Driven Software Development, 45(3):621–645, 2006.
7. Feiler P H, Lewis B, Vestal S, et al. An overview of the SAE architecture analysis & design language (AADL) standard: a basis for model-based architecture-driven embedded systems engineering[M].Architecture Description Languages. Springer US, 2005: 3-15.
8. Modelica Association. Modelica: A Unified Object- Oriented Language for Physical Systems Modeling: Language Specification Version 3.0, Sept 2007.www.modelica.org
9. Johnson, T. A., C. J. J. Paredis and R. M. Burkhart. "Integrating Models and Simulations of Continuous Dynamics into SysML." *6th International Modelica Conference*, Bielefeld, Germany, March 3-4, Modelica Association, 135-145, 2008
10. Xiping Song. Systematic Integration of Design Methods, IEEE Software, Volume 14 Issue 2, March 1997, Page 107-117
11. Atif Aftab Ahmed Jilani et al. Model Transformations in Model Driven Architecture, Universal Journal of Computer Science and Engineering Technology. 1 (1), 50-54, Oct. 2010
12. Kleppe, Warmer, J., Bast., W.: MDA Explained, The Model-Driven Architecture: Practice and Promise. Addison Wesley (2003)
13. Jean Bezivin et al. Model Transformations? Transformation Models!, MoDELS 2006, LNCS 4199, pp. 440–453, 200
14. S. Sendall and W. Kozaczynski, Model Transformation: the heart and soul of MSD, Software, IEEE 2003; 20(5): 42-45
15. Kevin Lano et al. Comparative Evaluation of Model Transformation Specification Approaches, Int J Software Informatics, Volume 6, Issue 2 (2012), pp. 233{269
16. *OMG Systems Modeling Language (OMG SysML), v1.2*. OMG, Needham, MA.
17. Fritzson P, Engelson V. Modelica—A unified object-oriented language for system modeling and simulation[M]//ECOOP'98—Object-Oriented Programming. Springer Berlin Heidelberg, 1998: 67-90.
18. OMG SE DSIG SysML-Modelica Working Group. *SysML-Modelica transformation specification*.http://www.omgwiki.org/OMGSysML/doku.php?id=sysml-modelica:sysml_and_modelica_integration, 2009.
19. Wladimir Schamai,EADS Innovation Works (Hamburg, Germany), Linkoping University (Linkoping, Sweden). ModelicaML: Getting Started. Issue 1.1, November 2009
20. Feiler P H, Gluch D P, Hudak J J. The architecture analysis & design language (AADL): An introduction[R]. CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST, 2006.
21. Behrmann G,David A,Larsen K G.A turorial on UPPAAL[C]// Proceedings of the 4th Int'l School on Formal Methods for the Design of Computer, Communication, and Software Systems. Bertinoro: Springer ~ Verlag,2004:200-236

2D Multi-Slice and 3D k-Space Simulations using a 3D Quadric Head Phantom with MRI Properties

H. Michael Gach

1 Introduction

In 1974, the 2D Shepp-Logan phantom consisting of various ellipses was introduced to simulate a single slice of the human cerebrum for the analysis of back-projection reconstructions typical of computer tomography (CT) [1]. In 1980, L.A. Shepp developed a 3D version of the Shepp-Logan phantom for CT and MRI projection simulations. Quadric phantoms consisting of multiple ellipsoids or ellipses are useful for analyzing MRI acquisitions since the closed form Fourier transforms of an ellipsoid and ellipse are known and the analytic k-space can be easily calculated [2–4]. For 2D multi-slice acquisitions using a 3D quadric phantom, the intersection of the image plane with the ellipsoids must be calculated prior to the conversion to the analytic representation of k-space for each 2D slice, or the generation of 1D projections.

2D and 3D quadric (i.e., ellipse and ellipsoid) phantoms continue to have utility because: 1) The computations of the quadric phantoms are much simpler and faster than digital brains [5], permitting rapid prototyping and troubleshooting; 2) The analytic k-space functions avoid sampling artifacts associated with digital brains; 3) Not all users and applications require the resolution and detail of digital brains; 4) The geometric phantoms are more versatile for simulating different image contrasts and orientations; and 5) One can compare multi-slice 2D and 3D k-space acquisitions.

A generalized matrix-based solution is provided herein for finding the specifications of the ellipse resulting from the intersection of a plane and ellipsoid of arbitrary orientations, sizes, and centers [6–10]. Examples of the solution are demonstrated for various 2D slices through a 3D Shepp phantom that was constructed with MRI properties for k-space simulations. MRI contrast weighting and magnetic susceptibility were incorporated into the model.

2 Theory

Intersection of Plane and Ellipsoid: We define three coordinate systems: 1) The canonical (or absolute) coordinate system of the MRI scanner xyz (henceforth $\mathbf{x}$); 2) The coordinate system for each imaging plane $x'y'z'$ (henceforth $\mathbf{x}'$); and 3) The axis-aligned coordinate system of each ellipsoid or ellipse $x''y''z''$ (henceforth $\mathbf{x}''$). The rotation from $\mathbf{x}'$ to $\mathbf{x}$ is given by:

$$\mathbf{x} = \mathbf{R}\mathbf{x}' \tag{1}$$

where $\mathbf{R}$ is the rotation matrix. Based on the Digital Imaging and Communications in Medicine (DICOM) convention, the canonical x-axis runs from the human subject's right to left, the y-axis runs from anterior to posterior, and the z-axis runs from inferior to superior (foot to head). For successive rotations (e.g., oblique slices), we used:

$$\mathbf{R}_p = \mathbf{R}_x(\theta)\mathbf{R}_y(\psi)\mathbf{R}_z(\phi) \tag{2}$$

where the rotation matrices are defined herein as:

$$R_x(\theta)=\begin{bmatrix}1 & 0 & 0\\ 0 & \cos(\theta) & \sin(\theta)\\ 0 & -\sin(\theta) & \cos(\theta)\end{bmatrix}, R_y(\psi)=\begin{bmatrix}\cos(\psi) & 0 & -\sin(\psi)\\ 0 & 1 & 0\\ \sin(\psi) & 0 & \cos(\psi)\end{bmatrix},$$
$$R_z(\phi)=\begin{bmatrix}\cos(\phi) & \sin(\phi) & 0\\ -\sin(\phi) & \cos(\phi) & 0\\ 0 & 0 & 1\end{bmatrix} \tag{3}$$

Since we will be dealing with 4x4 conic matrices herein, $\mathbf{R}$ becomes:

H.M. Gach (✉)
Departments of Radiology and Bioengineering,
University of Pittsburgh, Pittsburgh, Pennsylvania, USA
e-mail: gach@pitt.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_90,

$$\mathbf{R}_{4x4} = \begin{bmatrix} \mathbf{R}_{3x3} & \mathbf{0}_{3x1} \\ \mathbf{0}_{1x3} & 1 \end{bmatrix} \tag{4}$$

consisting of the 3x3 minor of the rotation matrix **R**, and 1x3 and 3x1 vectors of zeroes. The coordinate vectors are also expanded to 4-vectors, e.g.,:

$$\mathbf{x} = [\,x \quad y \quad z \quad 1\,] \tag{5}$$

The matrix representation of a plane in the coordinate system **x** is given by:

$$\boldsymbol{\alpha}^{\mathbf{T}}(\mathbf{x} - \mathbf{x}_p) = 0 \tag{6}$$

Herein, $\boldsymbol{\alpha}^{\mathbf{T}}$ represents a 1x4 vector whose first three elements are the direction cosines of the normal ***n*** to the plane and its fourth element is 0, while $\mathbf{x_p}$ represents a point lying on the plane that represents a translation of the plane from the origin of **x**. Typically, we acquire images of the brain in one of three orthogonal planes (whose normal is represented by $\boldsymbol{\alpha_o}^{\mathbf{T}}$): sagittal with $\boldsymbol{\alpha_o}^{\mathbf{T}} = [1,0,0,0]$, coronal with $\boldsymbol{\alpha_o}^{\mathbf{T}} = [0,1,0,0]$, or axial with $\boldsymbol{\alpha_o}^{\mathbf{T}} = [0,0,1,0]$ but we can also construct oblique slices of the phantom.

Transforming into the coordinate system of the imaging plane, $\mathbf{x}'$, we have:

$$\begin{aligned} \boldsymbol{\alpha}^{\mathbf{T}}(\mathbf{x} - \mathbf{x}_p) &= \boldsymbol{\alpha}^{\mathbf{T}}(\mathbf{R}_p\mathbf{x}' - \mathbf{x}_p) = \boldsymbol{\alpha}^{\mathbf{T}}(\mathbf{R}_p\mathbf{x}' + \mathbf{b}) \\ &= \boldsymbol{\alpha}^{\mathbf{T}}\mathbf{A}_p\mathbf{x}' = 0 \end{aligned} \tag{7}$$

where $\mathbf{A}_\mathrm{p}$ is given by the 4x4 matrix:

$$\mathbf{A}_p = \begin{bmatrix} \mathbf{R}_{p3x3} & \mathbf{b}_{3x1} \\ \mathbf{0}_{1x3} & 1 \end{bmatrix} \tag{8}$$

consisting of the 3x3 minor of the rotation matrix $\mathbf{R_p}$, and the 3x1 minor of the translation vector $\mathbf{b} = \mathbf{x_p}$. In this notation, $\mathbf{x}'$ is a 4x1 vector.

Given the axis-aligned coordinate system $\boldsymbol{x''y''z''}$ (henceforth $\mathbf{x}''$), the matrix equation of an ellipsoid is:

$$\begin{aligned} \mathbf{x}''^{T}\mathbf{Q}_{3D}\mathbf{x}'' &= \left(\mathbf{A}_e^{-1}\mathbf{x}\right)^T\mathbf{Q}_{3D}\left(\mathbf{A}_e^{-1}\mathbf{x}\right) \\ &= \mathbf{x}^T\left(\left(\mathbf{A}_e^{-1}\right)^T\mathbf{Q}_{3D}\mathbf{A}_e^{-1}\right)\mathbf{x} = \mathbf{x}^T\mathbf{Q}_{3DA}\mathbf{x} = 0 \end{aligned} \tag{9}$$

where $\mathbf{Q_{3D}}$ and $\mathbf{Q_{3DA}}$ are 4x4 symmetric positive definite matrices (i.e., $Q_{ij} = Q_{ji}$) representing the ellipsoids before and after rotation and translation to **x**, where **x** and $\mathbf{x}''$ are 4-vectors, and $\mathbf{A_e}^{-1}$ is the 4x4 inverse affine transformation matrix:

$$\mathbf{A}_e^{-1} = \begin{bmatrix} \mathbf{R}_{e3x3}^T & -\mathbf{R}_{e3x3}^T\mathbf{b}_{3x1} \\ \mathbf{0}_{1x3} & 1 \end{bmatrix} \tag{10}$$

$\mathbf{x}^{\mathrm{T}}\mathbf{Q_{3DA}}\mathbf{x}$ represents the quadric equation:

$$\begin{aligned} &Q_{11}x^2 + Q_{22}y^2 + Q_{33}z^2 + 2Q_{12}xy + 2Q_{13}xz + \\ &2Q_{23}yz + 2Q_{14}x + 2Q_{24}y + 2Q_{34}z + Q_{44} = 0 \end{aligned} \tag{11}$$

Rotations of the ellipsoid affect the minor matrix consisting of the first 6 terms while translations affect the last 4 terms.

Now that we have an expression for each ellipsoid in **x**, we need an expression for each ellipsoid in the coordinate system of the imaging plane $\mathbf{x}'$ by applying the direction cosines $\mathbf{R_p}$:

$$\begin{aligned} \mathbf{x}^T\mathbf{Q}_{3DRT}\mathbf{x} &= \left(\mathbf{R}_p\mathbf{x}'\right)^T\mathbf{Q}_{3DRT}\left(\mathbf{R}_p\mathbf{x}'\right) \\ &= \mathbf{x}'^T\left(\mathbf{R}_p{}^T\mathbf{Q}_{3DRT}\mathbf{R}_p\right)\mathbf{x}' = \mathbf{x}'^T\mathbf{Q}_F\mathbf{x}' = 0 \end{aligned} \tag{12}$$

The coefficients for the ellipsoid $\mathbf{Q_F}$ in $\mathbf{x}'$ are Q_{ijF}, with $\mathbf{x}'^{\mathrm{T}}\mathbf{Q_F}\mathbf{x}'$ representing the quadric equation:

$$\begin{aligned} &Q_{11F}x'^2 + Q_{22F}y'^2 + Q_{33F}z'^2 + 2Q_{12F}x'y' + 2Q_{13F}x'z' + \\ &2Q_{23F}y'z' + 2Q_{14F}x' + 2Q_{24F}y' + 2Q_{34F}z' + Q_{44F} = 0 \end{aligned} \tag{13}$$

Next, we find the intersection of the plane and ellipsoid in $\mathbf{x}'$ by substituting the expression for the plane into that of the ellipsoid. The matrix representation for the ellipse is similar to that of the ellipsoid, except the dimensions of the vectors and matrices are reduced by 1. For the ellipse, $\mathbf{Q_{2D}}$ is a 3x3 symmetric positive definite matrix (i.e., $q_{ij} = q_{ji}$) representing the coefficients of a conic section $\mathbf{u}'^{\mathrm{T}}\mathbf{Q_{2D}}\mathbf{u}'$:

$$q_{11}u'^2 + q_{22}v'^2 + 2q_{12}u'v' + 2q_{13}u' + 2q_{23}v' + q_{33} = 0 \tag{14}$$

We substitute the coefficients from $\mathbf{Q_F}$ (21) into $\mathbf{Q_{2D}}$ (23) to find the resulting intersection ellipse. Substituting $u' \rightarrow y'$, $v' \rightarrow z'$, for sagittal or oblique sagittal ($x' = k$), we obtain:

$$\begin{aligned} &Q_{22F}y'^2 + Q_{33F}z'^2 + 2Q_{23F}y'z' + 2(Q_{24F} + kQ_{12F})y' + \\ &2(Q_{34F} + kQ_{13F})z' + \left(Q_{44F} + 2kQ_{14F} + k^2Q_{11F}\right) = 0 \end{aligned} \tag{15}$$

For coronal or oblique coronal ($y' = k$) and substituting $u' \rightarrow x'$, $v' \rightarrow z'$, we obtain:

$$\begin{aligned} &Q_{11F}x'^2 + Q_{33F}z'^2 + 2Q_{13F}x'z' + 2(Q_{14F} + kQ_{12F})x' + \\ &2(Q_{34F} + kQ_{23F})z' + \left(Q_{44F} + 2kQ_{24F} + k^2Q_{22F}\right) = 0 \end{aligned} \tag{16}$$

For axial or oblique axial ($z' = k$) and substituting $u' \rightarrow x'$, $v' \rightarrow y'$, we obtain:

Table 1 3D Phantom Specifications

No.	Tissue Primary, Complementary	Ellipsoid Center x_0,y_0,z_0	Semi Axis Lengths a_1,a_2,a_3	Rotation Angles Θ,ψ,φ
1	Air, None	0,0,0	1.50,1.50,1.50	0,0,0
2	Scalp, Air	0,0,0	0.75,0.95,1.28	0,0,0
3	Fat, Scalp	0,0,0	0.73,0.93,1.26	0,0,0
4	Bone, Fat	0,0,0	0.72,0.92,1.24	0,0,0
5	Marrow, Bone	0,0,0	0.705,0.905,1.23	0,0,0
6	Bone, Marrow	0,0,0	0.68,0.88,1.20	0,0,0
7	CSF, Bone	0,0,0	0.67,0.87,1.19	0,0,0
8	GM, CSF	0,0,0	0.65,0.85,1.18	0,0,0
9	WM, GM	0,0,0.45	0.51,0.55,0.51	0,0,0
10	WM, GM	0,0.50,-0.30	0.40,0.20,0.20	0,0,0
11	CSF, WM	0.22,0.10,0.38	0.11,0.31,0.10	-10^0,0,18^0
12	CSF, WM	-0.22,0.10,0.38	0.16,0.41,0.10	-10^0,0, -18^0
13	Left eye, GM	0.26,-0.60,0	0.13,0.13,0.13	0,0,0
14	Right eye, GM	-0.26,-0.60,0	0.13,0.13,0.13	0,0,0
15	Nose (Skin), Air	0,-1.14,-0.20	0.10,0.25,0.15	57^0,0, 0
16	Nasal Cavity (Air), GM	0,-0.63,-0.22	0.20,0.20,0.17	-20^0,0, 0
17	Mouth (Air), GM	0,-0.13,-0.67	0.47,0.47,0.45	0,0,0
18	Left ear (Skin), Air	0.79,0.14,-0.19	0.06,0.32,0.32	$0,5^0,6^0$
19	Right ear (Skin), Air	-0.79,0.14,-0.19	0.06,0.32,0.32	$0,-5^0,174^0$
20	L. Aud. Canal (Air), GM	0.33,0.14,-0.19	0.30,0.15,0.15	$0,10^0,20^0$
21	R. Aud. Canal (Air), GM	-0.33,0.14,-0.19	0.30,0.15,0.15	$0,-10^0,-20^0$
22	Right tumor, GM	-0.08,0.61,0.38	0.05,0.02,0.02	0,0,0
23	Center tumor, GM	0,0.61,0.38	0.02,0.02,0.05	0,0,0
24	Left tumor, GM	0.06,0.61,0.38	0.02,0.05,0.02	0,0,0
25	Tumor, WM	0.02,0,0.38	0.05,0.05,0.05	0,0,0
26	Tumor, WM	0.02,0.20,0.38	0.05,0.05,0.05	0,0,0
27	Tumor, WM	0,-0.28,0.38	0.20,0.20,0.23	0,0,0
28	Blood clot, GM	0.40,0.40,0.38	0.03,0.03,0.20	$0,-168^0,-155^0$

White Matter: WM, Gray Matter: GM, Cerebrospinal fluid (CSF), Left: L., Right: R.

$$Q_{11F}x'^2 + Q_{22F}y'^2 + 2Q_{12F}x'y' + 2(Q_{14F} + kQ_{13F})x' + 2(Q_{24F} + kQ_{23F})y' + \left(Q_{44F} + 2kQ_{34F} + k^2Q_{33F}\right) = 0 \tag{17}$$

If the determinant of $\mathbf{Q_{2D}}$ is zero, then the conic is degenerate. Assuming the conic is not degenerate, we can verify the conic is an ellipse since the determinant of the minor $\mathbf{Q_{11}}$ must be positive (i.e., $|\mathbf{Q_{11}}| > 0$) where:

$$\mathbf{Q}_{11} = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \tag{18}$$

The center of the ellipse in $\mathbf{x}'$ is calculated by setting the first two rows of $\mathbf{Q_{2D}}$ to zero and solving for u' and v'. The solution is given by:

$$u'_c = \frac{q_{22}q_{13} - q_{23}q_{12}}{q_{12}^2 - q_{11}q_{22}}, v'_c = \frac{q_{23}q_{11} - q_{13}q_{12}}{q_{12}^2 - q_{11}q_{22}} \tag{19}$$

The length of the major and minor semi-axes (i.e., b_i) of the ellipse are eigenvalues of $\mathbf{Q_{2D}}$ but can be acquired using the determinant of $\mathbf{Q_{2D}}$ and its minor $\mathbf{Q}_{11}$. First, one obtains the eigenvalues $\boldsymbol{\lambda}$ of $\mathbf{Q_{11}}$ and then:

$$b_i = \sqrt{\frac{-|\mathbf{Q_{11}}|}{|\mathbf{Q}_{2D}|\lambda_i}} \tag{20}$$

The direction cosines of the major and minor semi-axes of the ellipse are provided by the eigenvectors of $\mathbf{Q}_{11}$. If the conic is degenerate and the determinant of $\mathbf{Q_{11}}$ is positive then the conic is empty (e.g., the plane and ellipsoid do not intersect).

3 Methods

The 3D Shepp head phantom was modified and expanded from 17 to 28 primary and 27 complementary ellipsoids for the MRI simulations and placed into the DICOM convention (Table 1). The dimensions of the ellipsoids are unit-less. For the purpose of this demonstration, we assumed the slices lie in parallel planes with a fixed distance (spacing) between

Table 2 Tissue MR Properties

Tissue	T_1 (s) A	T_1 (s) C	T_2 (s)	Magnetic Susceptibility χ (x10^{-6})	Proton Density ρ
Air	0	0	0	0.36	0
Scalp/Skin	0.32	0.14	0.07	-8.44	0.80
Fat	0.32	0.14	0.08	-8.44	0.98
Bone	0.07	0.14	<0.005	-11.31	0.12
Marrow	0.53	0.09	0.08	-9.77	0.56
CSF, Eye	4.20	0	1.99	-9.05	0.98, 0.99
Gray Matter	0.86	0.38	0.10	-9.05	0.75
White Matter	0.58	0.38	0.08	-9.05	0.62
Tumor	0.93	0.22	0.10	-9.05	0.95
Blood Clot	1.35	0.34	0.20	-9.05	0.85

them and that the image slices are infinitesimally thin. Although the latter assumption is typically used for medical imaging simulations it does not represent the reality of a medical image that represents the integral of the contents of a slice of finite thickness. Such an integral or an average may be performed after decomposing an image plane of finite thickness into multiple component planes.

The T_1, T_2, and susceptibility properties of the tissues are shown in Table 2. The T_1 is dependent on the static magnetic field strength (B_0) according to [11]:

$$T_1(B_0) \approx A \cdot (B_0)^C \tag{21}$$

where the values of A and C in Table 2 assume B_0 is in units of Tesla. We assumed that the T_1 of CSF and the T_2 of the tissues do not vary with field strength.

The complementary ellipsoids must be subtracted from the primary ellipsoids to avoid superposition of the ellipsoid properties. The complementary ellipsoids have the same geometric specification as the primary ellipsoids, but have the relaxation properties and proton densities of the larger enclosing ellipsoids. The subtraction does not resolve partial overlaps. Therefore, we adjusted the ellipsoid sizes and centers to avoid partial overlaps. Overlaps were tested by calculating the eigenvalues and eigenvectors of the matrix product $\mathbf{Q_{F1}}^{-1}\mathbf{Q_{F2}}$ where $\mathbf{Q_{F1}}$ and $\mathbf{Q_{F2}}$ are the quadric matrices for the two ellipsoids under test. If the eigenvalues and eigenvectors are complex then the ellipsoids partially overlap [12]. If the eigenvectors and eigenvalues are real, then the ellipsoids either have no intersection, touch, or one of the ellipsoids is fully contained within the other. This technique can also be used to detect if ellipses partially overlap by replacing the ellipsoid matrices with the ellipse matrices (e.g., $\mathbf{Q_{2D}}$).

The matrix equations, phantom, and k-space simulator were implemented in IDL Version 8 (ITT Visual Information Systems, Boulder, Co). The analytic k-space representation of the 2D phantom was the sum of the ellipses' 2D FTs [2]:

$$o_{2D}\left(\vec{k}\right) = \sum_m \frac{w_m \rho_m b_{1,m} b_{2,m} J_1(2\pi\kappa_m)}{\kappa_m} e^{-2\pi i \vec{k} \bullet \vec{r}_{0m}} - \sum_n \frac{w_n \rho_n b_{1,n} b_{2,n} J_1(2\pi\kappa_n)}{\kappa_n} e^{-2\pi i \vec{k} \bullet \vec{r}_{0n}} \tag{22}$$

where m and n are the indices for the intersection ellipses resulting from the primary and complementary ellipsoids, respectively; $b_{1,2}$ represents the semi-axis lengths of each intersection ellipse; J_1 is the Bessel function of the first kind; and w_{mn} represents the contrast weighting (e.g., from T_1, T_2, or T_2*).

$$\kappa_p^2 = \left(\vec{b}_{1,p} \bullet \vec{k}\right)^2 + \left(\vec{b}_{2,p} \bullet \vec{k}\right)^2 \tag{23}$$

with the semi-axes in vector form.

Demonstration images presented herein were created using T_1 and T_2 weighting based on a fast low-angle shot (FLASH) sequence. At steady-state, the combined weighting is [13]:

$$\omega_i = \frac{\sin(\Theta)\left(1 - e^{-TR/T_{1i}}\right)}{1 - \cos(\Theta) e^{-TR/T_{1i}}} e^{-TE/T_{2i}} \tag{24}$$

where Θ is the flip angle, TR is the repeat time, and TE is the echo time. For a conventional spin-echo sequence at steady-state, the weighting is [13]:

$$\omega_i = \frac{\sin(\Theta)\left(1 - 2e^{-(TR-TE/2)/T_{1i}} + e^{-TR/T_{1i}}\right)}{1 + \cos(\Theta) e^{-TR/T_{1i}}} e^{-TE/T_{2i}} \tag{25}$$

More complicated weightings and artifacts can be incorporated into the phantom through matrix operations [14].

Magnetic susceptibility effects were modeled based on the susceptibility-convolution and spin dephasing methods described by Eqs. 19 and 24 in Yoder et al. [14]. We calculated the B_0 map from the 3D susceptibility image. The analytic k-space representation of the 3D susceptibility image was derived from the primary and complementary ellipsoid 3D FTs using [2]:

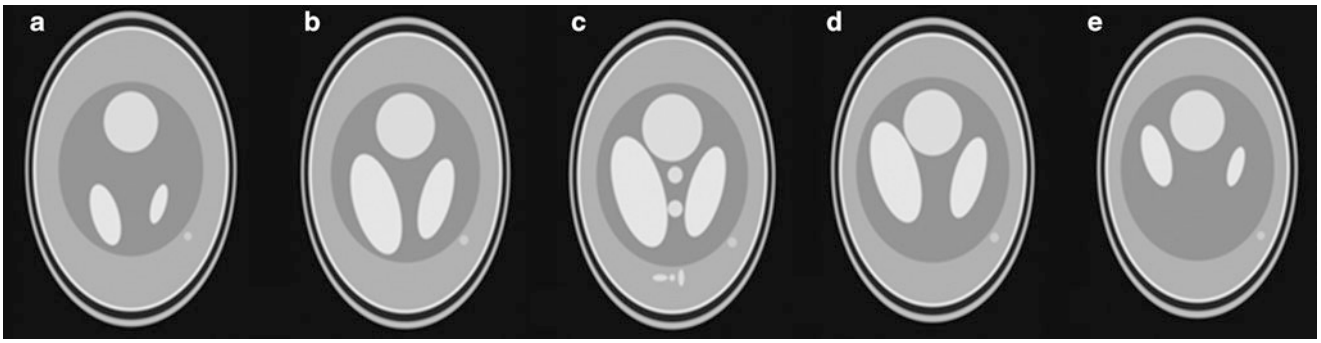


Fig. 1 Axial 2D proton density (spin-echo) slices through the phantom (Matrix: 256x256, TR: 14 s, TE: 0.005 s, Flip angle: 90^0, B_0: 3 T) with the nominal Shepp-Logan phantom (c). The imaging volume was centered at x = y = 0 with z = 0.28 (a), 0.33 (b), 0.38 (c), 0.43 (d), 0.48 (e). The slices were spaced 0.05 units apart.

$$o_{3D}\left(\vec{k}\right) = \sum_m \frac{\chi_m 4\pi(a_{1,m}a_{2,m}a_{3,m})}{(2\pi\kappa_m)^2} \left(\frac{\sin 2\pi\kappa_m}{2\pi\kappa_m} - \cos 2\pi\kappa_m\right) \cdot e^{-2\pi i\vec{k}\bullet\vec{r}_{0m}} - \sum_n \frac{\chi_n 4\pi(a_{1,n}a_{2,n}a_{3,n})}{(2\pi\kappa_n)^2} \left(\frac{\sin 2\pi\kappa_n}{2\pi\kappa_n} - \cos 2\pi\kappa_n\right) \cdot e^{-2\pi i\vec{k}\bullet\vec{r}_{0n}} \quad (26)$$

where m and n are the indices for the primary and complementary ellipsoids, respectively, with:

$$\kappa_p^2 = \left(\vec{a}_{1,p}\bullet\vec{k}\right)^2 + \left(\vec{a}_{2,p}\bullet\vec{k}\right)^2 + \left(\vec{a}_{3,p}\bullet\vec{k}\right)^2 \quad (27)$$

and replacing the ellipsoid spin density (ρ) with the magnetic susceptibility (χ).

The magnetic field perturbation for each voxel $\mathbf{B_e}$ was calculated in $\mathbf{x}$, transformed into the slice frame ($\mathbf{x}'$) using the affine transformation matrix $\mathbf{A_p}$, and then aligned with each corresponding slice derived from o_{2D}. The spatial shift $\mathbf{B_e/G_{read}}$ was calculated separately for each voxel and requires N^2 convolutions for a matrix dimension of NxN. The spatial shift from a Cartesian k-space trajectory with a constant read-out gradient was given by:

$$I_{post}\left(\vec{p},n\right) = \sum_{j=0}^{N_r-1} I_{pre}(j,n)\cdot DFT\left[\exp\left[2\pi i\left(\frac{B_e(j,n)}{G_r}+j\right)\frac{\vec{l}}{N_r}\right]\right] \quad (28)$$

where N_r is the dimension along the readout direction (here chosen as corresponding to the index m), l is an N_r-dimensional vector that runs from 0 to N_r-1, and DFT is the discrete Fourier transform. We ignored the effects of the phase-encoding gradient.

4 Results

Figure 1 shows proton density axial images through the phantom. Figure 2 shows example images from the quadric phantom in various orientations and with various T_1 and T_2 weightings. Note that the ventricles are sloped to emulate true anatomy. All of the images appear in DICOM radiological coordinates.

5 Discussion

We successfully demonstrated a 2D multi-slice phantom using a 3D quadric head phantom with MRI properties. We extended the work presented by Koay et al. by providing: 1) a 3D phantom with MRI properties and enhanced anatomical details; and 2) providing an algorithm for acquiring the 2D slices through the phantom at any position and orientation, and in either image or k-space domains [2]. Although more sophisticated models are available, they require significantly more processing time and are typically represented in discrete space. Our simulations use closed-form analytic k-space representations that are suitable for evaluating and comparing k-space trajectories and simulating artifacts.

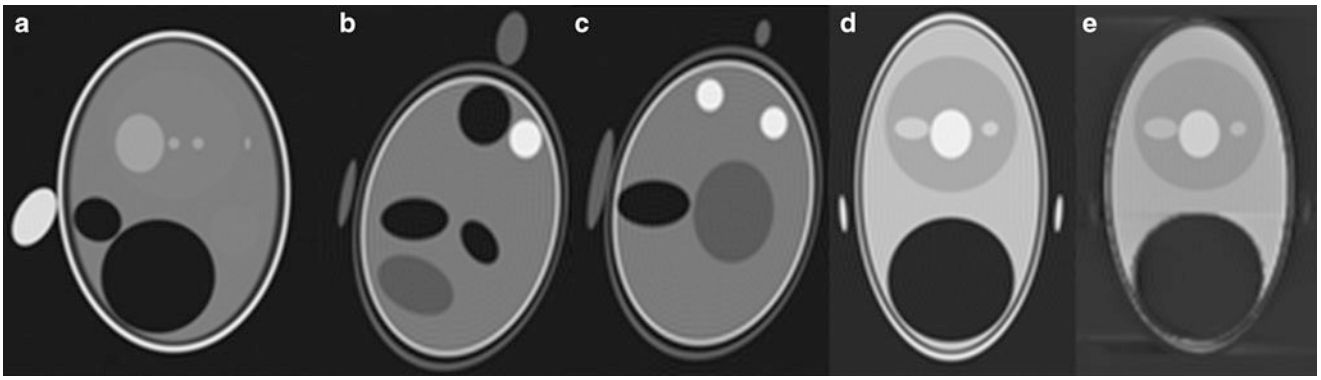


Fig. 2 Sagittal T_1-weighted (FLASH) 2D slice through the phantom without magnetic susceptibility (Matrix: 128x128, TR: 0.1 s, TE: 0.005 s, Flip angle: 30^0, B_0: 3 T) (a). Oblique axial T_2-weighted (spin-echo) 2D slices through the phantom (Matrix: 128x128, TR: 4 s, TE: 0.1 s, Flip angle: 90^0, B_0: 3 T). The imaging volume was centered at $x = y = z = 0$. The slice spacing was 0.15 units (b, c). The slice rotation angles are $\varphi = \psi = 20^0$. Coronal T_2*-weighted (FLASH) 2D slices through the phantom without (d) and with (e) magnetic susceptibility artifacts (Matrix: 128x128, TR: 0.5 s, TE: 0.03 s, Flip angle: 30^0, Read Gradient: 4.7 mT/m, B_0: 3 T). The imaging volume was centered at $x = z = 0$. The slices were centered at $x = -0.075$ (d, e). The readout gradient was placed along the x axis.

Acknowledgements. This research was conducted with the support of the Nevada Cancer Institute, the University of Pittsburgh, and the National Institutes for Health National Cancer Institute grant R01 CA159471-01. We are grateful to Fernando Boada and Costin Tanase for their early models that inspired this work.

References

1. Shepp, L.A., Logan, B.F.: The fourier reconstruction of a head section. IEEE Transactions on Nuclear Science NS-21, 21-43 (1974)
2. Koay, C.G., Sarlls, J.E., Ozarslan, E.: Three-dimensional analytical magnetic resonance imaging phantom in the Fourier domain. Magnetic Resonance in Medicine 58, 430-436 (2007)
3. Van de Walle, R., Barrett, H.H., Myers, K.J., Altbach, M.I., Desplanques, B., Gmitro, A.F., Cornelis, J., Lemahieu, I.: Reconstruction of MR images from data acquired on a general nonregular grid by pseudoinverse calculation. IEEE Transactions on Medical Imaging 19, 1160-1167 (2000)
4. Pan, S., Kak, A.: A computational study of reconstruction algorithms for diffraction tomography: interpolation versus filtered-backpropagation. IEEE Transactions on Acoustics Speech and Signal Processing 31, 1262-1275 (1983)
5. Collins, D.L., Zijdenbos, A.P., LKollokian, V., Sled, J.G., Kabani, N.J., Holmes, C.J., Evans, A.C.: Design and construction of a realistic digital brain phantom. IEEE Transactions on Medical Imaging 17, 463-468 (1998)
6. Ferguson, C.C.: Intersections of ellipsoids and planes of arbitrary orientation and position. Mathematical Geology 11, 329-335 (1979)
7. Gendzwill, D.J., Stauffer, M.R.: Analysis of triaxial ellipsoids: Their shapes, plane sections, and plane projections. Mathematical Geology 13, 135-152 (1981)
8. Shene, C.-K., Johnstone, J.K.: Computing the intersection of a plane and a revolute quadric. Computers and Graphics 18, 47-59 (1994)
9. Kellman, P., Derbyshire, J.A., McVeigh, E.R.: Automatic in-plane rotation for doubly-oblique cardiac imaging. Journal of Magnetic Resonance Imaging 18, 612-615 (2003)
10. Klein, P.P.: On the Ellipsoid and Plane Intersection Equation. Applied Mathematics 3, 1634-1640 (2012)
11. Bottomley, P.A., Foster, T.H., Argersinger, R.E., Pfeifer, L.M.: A review of normal tissue hydrogen NMR relaxation times and relaxation mechanisms from 1-100 MHz: dependence on tissue type, NMR frequency, temperature, species, excision, and age. Medical Physics 11, 425-448 (1984)
12. Alfano, S., Greer, M.L.: Determining if two solid ellipsoids intersect. Journal of Guidance, Control, and Dynamics 26, 106-110 (2003)
13. Callaghan, P.T.: Principles of Nuclear Magnetic Resonance Spectroscopy. Clarendon Press, Oxford (1993)
14. Yoder, D.A., Zhao, Y., Paschal, C.B., Fitzpatrick, J.M.: MRI simulator with object-specific field map calculations. Magnetic Resonance Imaging 22, 315-328 (2004)

Classification of Lungs Nodule using Hybrid Features from CT Scan Images

M. Arfan Jaffar and Eisa Al Eisa

1 Introduction

Lung cancer is the most important and leading cause of deaths in Western Countries. In 2004, there were 341.800 lung cancer deaths in Europe [1]. It is considered that nearly 85 percent cases occur due to Tobacco. Heavy smokers have a much higher risk of dying of lung cancer i.e. about 10 times that of non-smokers [2]. Thus it is always assumed that early detection and treatment at stage 1 have high survival rate. But unfortunately, lung cancer is usually detected late due to the lack of symptoms in its early stages. This is the reason why lung screening programs have been investigated to detect pulmonary nodules: they are small lesions which can be calcified or not, almost spherical in shape or with irregular borders. Lung segmentation is a precursor to all of these quantitative analysis applications. In the early and untimely detection of lung abnormalities, Computer Aided Diagnosis (CAD) of lung CT image has proved to be an important and innovative development. It has been instrumental in supporting the radiologists in their final decisions. The accurateness and higher decision confidence value of any lung abnormality identification system relies strongly on an efficient lung segmentation technique. Therefore it is very important for effective performance of these systems to provide them with entire and complete lung part of the image. Samuel et al. [6] has used Ball-Algorithm for the segmentation of lungs. It has been observed that selected ball size for morphological closing did not work for the whole database of a single patient. Binsheng et al. [7] have used histogram for threshold selection. In order to obtain complete hollow free lung mask, morphological closing is applied. Morphological operators help to choose spherical shape of the structural element so that filter size can be approximately determined. With the help of 3D mask, the lungs can be readily extracted from the original chest CT images [8]. An automatic CAD system for lung cancer screening was developed by Ayman El-Baz et al. [9]. Optimal gray-level thresholding is applied by Ayman El-Baz et al. [10] for the extraction of thorax area. Shiying et al. [11] have developed an automatic method for identifying lungs in 3D pulmonary X-Ray CT images. It can be observed that that this technique has the following very serious short comings: (i) Fixed ball size (ii) addition of unnecessary areas as lung regions (iii) Processing time overhead. This proposed method ensures that no nodule information is lost after segmentation. This will help us for better classification of nodules [11].

The paper is organized as follows. Section II contains a detailed description of the proposed system follows. Implementation and relevant results are presented in (Section IIII). Finally, Section IV ends the paper with several conclusions drawn from the design and the work with the proposed system.

2 Proposed Method

The proposed system consists of two major phases. In the first phase, Lungs segmentation has been performed by using fuzzy entropy. In the second phase, different type of features has been extracted for nodule classification by using artificial neural network classifier. In the following, the various steps of the method are described with more details [11]. It has been noted that the threshold computed for one slice cannot be used for all the slices since the grey level variations in each image for a specific case are quite large and must be catered to by the segmentation algorithm. This variation in grey levels is highlighted by the following histograms: Figure 2(a),(b) and (c). Secondly, the number of voxels which fall in a particular grey-level bin differs in

M.A. Jaffar (✉) • E.A. Eisa
College of Computer and Information Sciences, Al-Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia
e-mail: arfan.jaffar@ccis.imamu.edu.sa; aleisa@ccis.imamu.edu.sa

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_91,

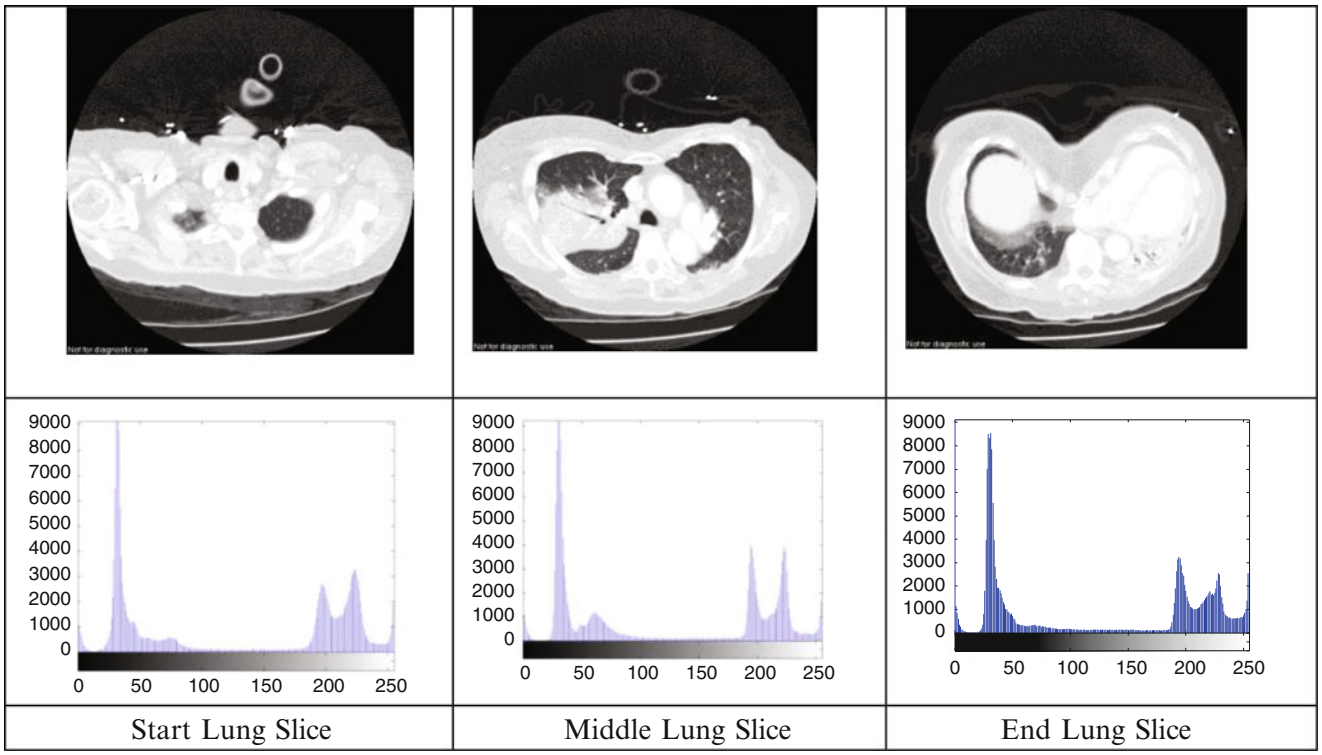


Figure 1 Test images of LIDC dataset

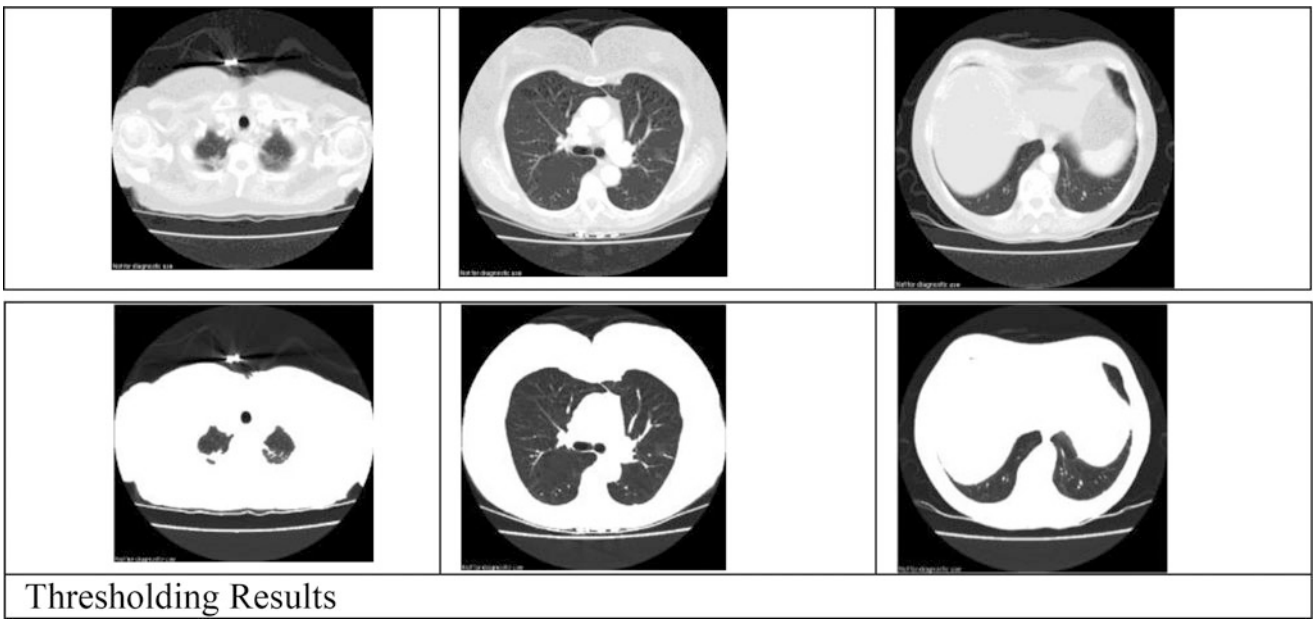


Figure 2 Threshold results on some test images of LIDC dataset

each slide. This means that the number of object and background voxels differs considerably in each slide. Hence, the iterative threshold for each image is different which yields optimum results. Therefore, we have calculated an optimal threshold which works in dynamic environment and calculate according to the slice [11]. This spatial relationship is important in clustering, but it is not utilized in a standard FCM algorithm.Weighted Spatial Fuzzy c-means (WSFCM) is an improved method of clustering which incorporates spatial information in standard FCM. One of the significant

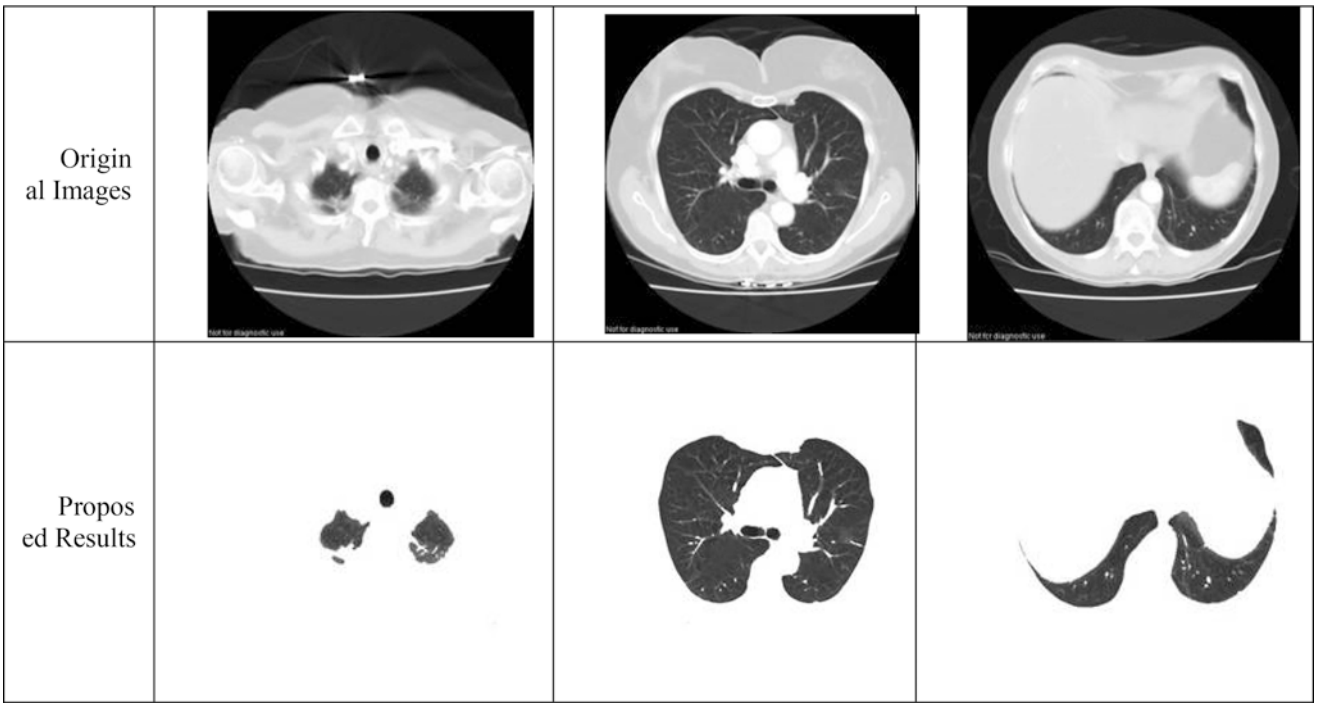


Figure 3 Lungs Segmentation results

uniqueness of an image is that neighboring pixels are extremely correlated. In other terms these neighboring pixels hold similar feature values, and the probability that they belong to the same cluster is large. This spatial relationship is important in clustering, but it is not utilized in a standard FCM algorithm [11]. To develop the spatial information, a spatial function is defined as

$$h_{ij} = \sum_{k \in NB(x_j)} W_{ij}\, U_{ik}$$

where$NB(x_j)$ stands for a square window centered on pixel x_j in the spatial domain and i,j are data points. W_{ij} is the weights that have been incorporated in spatial neighborhood in such a way that:

$$w_{i,j} = \begin{cases} 2, & (i,j) \in \Omega^3 \\ 1, & otherwise \end{cases}$$
$$\Omega^3 = \{(i,j) : -1 \le i,j \le 1\}$$

A 5x5 window has been used throughout in this research work. Just like the membership function, the spatial function h_{ij} stands for the probability that pixel x_j belongs to i^{th} cluster. We have used exponential entropy fuzzy entropy based error functions.The thresholds values are obtained from the error function, as the grey levels with the maximal levels of fuzziness respectively [17].

3 Lung Region Extraction

After thresholding, the lung region has to extract. For this purpose, we are interested in extracting lung regions from non-body voxels. To extract lung region, we have used 3-dimensional component labeling which use 18-connected neighborhood. In the first step, the labeled volumes LV are obtained and then remove the air in the vicinity of the body. The largest and the second largest volumes in LV are then selected [11]. We use the labeled volumes at median slice to reduce the computational complexity. After selecting the lung volume, air outside the body and gas in the intestine are removed. And also, remained holes in lung region are filled by using morphological operators. The extracted lung region shows the initial lung segmentation contours superimposed on the original section [11]. It occurs some times that the thresholding value selected for lung segmentation has erroneously resulted in the exclusion of a prominent juxta-pleural nodule such as two circle of figure 4 (a). To compensate for this type of segmentation error, morphology operator has to use. There are two types of morphology operators namely opening and closing. For this purpose we need an optimal opening method at the extracted lung region. For optimal opening method, we need shape and size of structuring element. We have find out that our nodules are approximately oval and circular type. Therefore, we have used circular structuring element and we also know

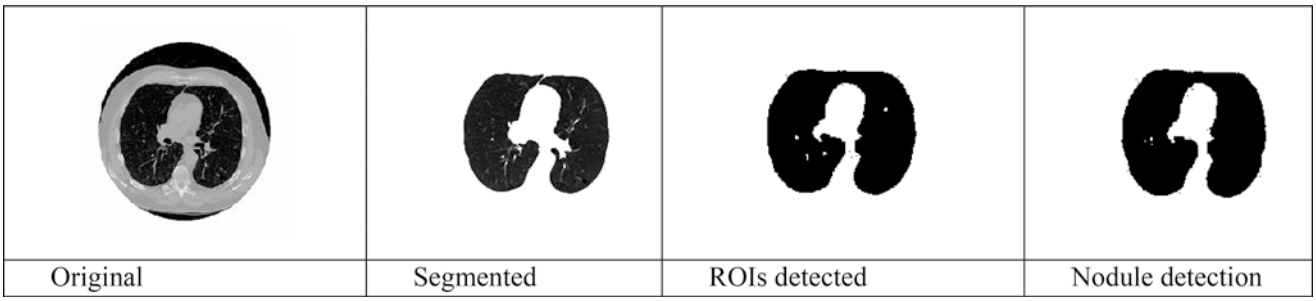


Figure 4 Nodule Detection results

the size of nodules. We have selected size of structuring element that is 13. This type of structuring element works well. For nodule detection, the region has been highly reduced. We have eliminated all the external structures except lungs lobes. Simple thresholding allows us to isolate the internal part of the lung, i.e., blood vessels, bronchi and nodules [11].

After extracting lungs part, we calculate histogram of lungs part only and calculate threshold. We segment the lungs part by using threshold. Nodules are of limited in size (3.5 to 7.3 mm) and most of the areas smaller than 3 pixels are false areas. Thus there is need a cleaning process that eliminate objects with area of 1 or 2 pixels. Therefore we have used an erosion filter of 3x3. In a typical image, the number of objects after the erosion decreases from 40/130 to 10/65, and most of the deleted objects are 1 or 2 pixels in area. The filter was set 3 by 3 because heavier erosion by a larger filter would have deleted too many pixels. The diamond shape proved to delete less objects than the square one, so it was chosen for precaution [11].

Nodules Detection. To reduce the complexity, we have to extract ROIs. Instead of scanning the whole lung image with the template pixel by pixel, only ROIs were considered in the scan. As a result the computation time and the detection time were reduced. To extract ROIs, eight directional searches have been used as shown in Fig. 5and it was considered that pixels which form a ROI must be members of a set of adjacent neighbor pixels with suitable intensities [11]. It has been observed that diameters of nodules are between the upper and lower boundaries. Therefore, to understand whether a pixel was in the center region of the ROI, diameter of the ROI was considered initially. In this stage, we introduce two more thresholds which form the boundaries, one is the minimum distance threshold representing the lower boundary and the other is the maximum distance threshold representing the upper boundary. These threshold values dealt with the resolution of the CT scan image and used to avoid very big or very small objects that are not part of lung nodule. We have to count adjacent neighbors of a pixel [11]. On serial images, vessels maintain a similar cross-sectional size and their in-plane circular appearance changes its location. But the true lung nodules remain at the same location from slice to slice. To check whether a ROI is on same location on consecutive slices, we calculate Euclidiandistance of current ROI to the all ROIs in the upper slice and lower slice. We calculated it by using equation (10) and procedure is shown in the figure. From these distances, we find out the minimum distance from upper slice and minimum distance from the lower slice. If these minimum distances are less than a threshold then that ROI is considered as a nodule otherwise deleted from that image. As a result, we got an image with less number of ROIs in the image [11].

After pruning ROIs in the slices, we construct a 3D image by using all slices of ROIs images. A 3D model of the nodule has been modeled by experienced radiologists. We have used a 3D model, which extracted the imaging features of a nodule, and then we performed a search through the 3D ROIs for objects that were similar to our 3D prismatic nodule template by using 3D model. There were 12 layers with 4x4 pixels in the 3D template. The values of the elements that constitute the 3D template were chosen as 1 and 2. The classification task was performed by convolving the 3D ROI image with our 3D nodule model that we call the 3D template [11].

Feature Extraction and Classification. After detecting nodules, we have extracted features of those nodules. These features have been used for classification. Artificial Neural Network has been used for classification.

Feature selection helps to reduce the feature space which improves the prediction accuracy and minimizes the computation time. This is achieved by removing irrelevant, redundant and noisy features .i.e., it selects the subset of features

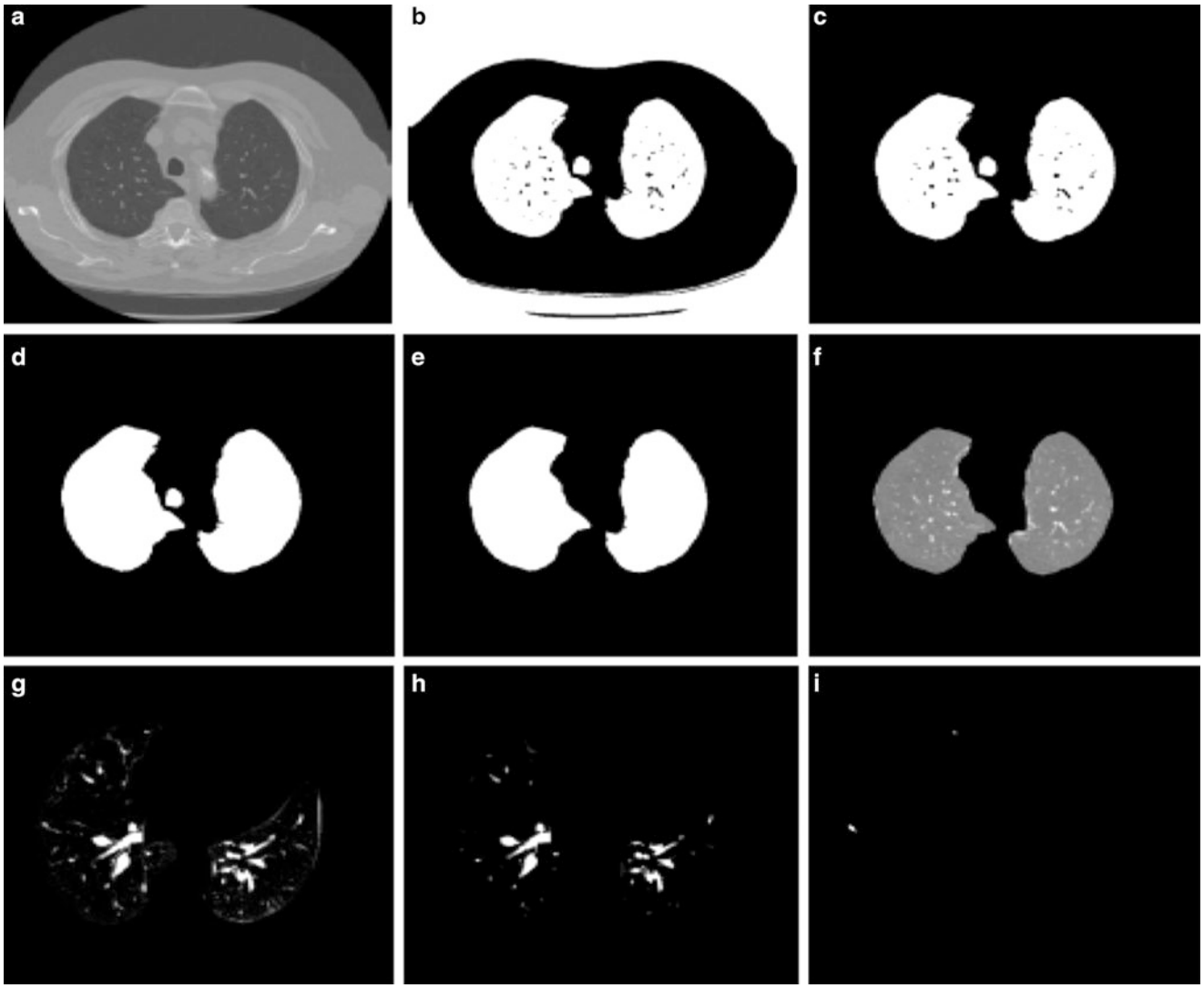


Figure 5 Nodule Detection Steps

that can achieve the best performance in terms of accuracy and computation time. It performs the dimensionality reduction.

Among a variety of features, we have selected different type of features. The 2D features are extracted from median slice of nodule candidates because the area of median slice is the largest in the nodule candidate volume. Proposed features consists of 2D geometric features consist of area (f1), diameter (f2), perimeter (f3) and circularity (f4), 2D intensity features comprise of minimum value inside (f5), mean inside (f6), mean outside (f7), varianceinside (f8), skewness inside (f9) and kurtosisinside (f10). Moreover, 8 largest eigenvalues (f11 ~ f18) of pixel intensities at median slice are extracted.

The feature set is divided into three subsets from all selected features **F**. The three selected feature vectors {**F1, F2, F3**} are provided as input for classifier.

F1 = {f1,f2,f3,f4}

F2 = {f5, f6, f7, f8, f9, f10, f11 ~ f18}

F3 = {f1,f2,f3,f4,f5,f6,f7, f8, f9, f10, f11 ~ f18} = F1UF2

2D geometric features	2D intensity feature
F1 = Area	F5 = Minimum value inside
F2 = Diameter	F6 = Mean inside
F3 = Perimeter	F7 = Mean outside
F4 = Circularity	F8 = Variance inside
	F9 = Skewness inside
	f10 = Kurtosis inside
	f11 ~ f18 = Eigenvalues

4 Experimental Results and Discussion

We have implemented the proposed system by using the MATLAB environment. We obtained datasets from Lung Image Database Consortium (LIDC) dataset and Aga Khan Medical University, Pakistan.

4.1 Segmentation and Nodule Detection Results

Fig. 7 gives the result of our proposed technique on the test image. It is evident through observation that the proposed system produces much smoother results. Results of our proposed method that are shown in Fig. 7demonstrate significant improvement. The figure shows that using our method, we are able to segment and extract nodules from that segment part is very good and promising. There is also no loss of lung nodules in our proposed method. The essence of our segmentation method lies in its ability to fully automatically segment the lungs part from whole CT scan image and detect nodules and classify.

Classification Results. We have performed three different types of experiments for classification and used ANN classifiers to test nodule classification. We generate results by repeating experiments 100 times and taking the average value of results. We have repeated the same process multiple times to cancel out the random effect. Table 1 shows theresults of these experiments.

Experiment I: 20–80/fixed. In this experiment, 20 % of the images of the each of the noduleand non-nodule classes were used to form the training set, andthe other 80 % of the images were used to form the test set. Table 1 shows theresults of this experiment.

Experiment II: 50–50/fixed. In this experiment, 50 % of the images of the each of the nodule and non-nodule classes were used to form the training set, and the other 50 % of the images were used to form the test set. Table 2 shows the results of this experiment.

Experiment III: 80–20/fixed. In this experiment, 80 % of the images of the each of the nodule and non-nodule classes were used to form the training set, and the other 20 % of the images were used to form the test set. Table 3 shows the results of this experiment.

I: I represent Intensity features (F2)
G: G represents geometric features (F1)
C: C represents combined features (F3)

Table 1 The classification result using features (rate 80 %-20 %)

20 %-80 %		Accuracy (%)	Sensitivity(%)	Specificity (%)
SVM	I	81.3846	83.6923	62.7215
	C	**90.0769**	**89.9385**	**91.5154**
	G	82.8462	83.3846	74.7344
NN	I	80.5385	84.0000	62.3781
	C	89.8462	89.0769	90.6154
	G	71.3846	73.8462	63.7646

Table 2 The classification result using features (rate 80 %-20 %)

50 %-50 %		Accuracy (%)	Sensitivity(%)	Specificity (%)
SVM	I	83.1262	85.2953	64.2272
	C	**92.9915**	**93.7365**	**92.9324**
	G	84.8572	84.9246	75.4834
NN	I	82.2743	86.7030	67.2631
	C	90.4723	90.4789	91.2164
	G	72.2635	74.7412	66.4626

Table 3 The classification result using features (rate 80 %-20 %)

80 %-20 %		Accuracy (%)	Sensitivity(%)	Specificity (%)
SVM	I	86.3846	87.6923	89.7215
	C	**95.9369**	**94.9385**	**97.5154**
	G	85.8462	87.3846	83.7344
NN	I	85.5385	88.0000	66.3781
	C	92.8462	92.0769	92.6154
	G	77.3846	78.8462	69.7646

Table 4 The comparison of classification result

Method	TPR	FPR	SPC	ACC
Decision tree	92.69	9.65	90.35	91.52
EM SVM	90.53	4.83	95.17	92.85
EM Random	95.68	5.49	94.51	95.10
Proposed	96.85	4.13	95.78	97.37

From these experiments, we have found that SVM performs well as compared to other classifiers and experiment III is better as compared to experiment I and II. So we have used SVM with experiment III for further experiments to compare results with other existing methods. We have compared our results with some previous techniques. Results in Table (4) shows that our proposed method performs better as compared to all other existing techniques.

5 Conclusions and Future Work

We have described an adaptive and fully automatic method for segmentation of pulmonary parenchyma. The "heart" of the FEMS system, FCM that performs the adaptive thresholding, used to determine the thresholds and the error function based upon fuzzy entropy measures, was designed. This is just the first step of a CAD system which is still under development. The results we obtained are comparable with those of other known methods. In addition, the proposed system has the advantage that it does not require any human expert intervention or any a priori information to the number of clusters (segments) that appear in the image. Not only does it performs efficiently on normal images, but also works equally well on images containing abnormality patches and nodules at any part of the lung, as verified by the results. Methodologies used in [5],[6],[9] and [11] show a serious shortcoming when they encounter with abnormal thoracic CT images, containing abnormality at the margins, which is a common case in images of most of the fatal lung diseases, like Cancer, TB etc.

References

1. Atam P. Dhawan, "Medical Image Analysis", IEEE press series in Biomedical Engineering, John Wiley & Sons. Inc. Publications, 2003
2. Hoffman, E. A, and McLennan, G., "Assessment of the pulmonary structure-function relationship and clinical outcomes measures: Quantitative volumetric CT of the lung", Academic Radiology, vol. 4, no. 11, pp. 758–776, 1997
3. M. Arfan Jaffar, Ayyaz Hussain and Anwar Majid Mirza, Fuzzy entropy based optimization of clusters for the segmentation of lungs in CT scanned images, Knowledge and Information Systems, 2010, Volume 24, Number 1, Pages 91–111
4. Aristofanes C. Cilva, Paulo Cezar, Marcello Gattas, "Diagnosis of Lung Nodule using Gini Coefficient and skeletonization in computerized Tomography images", ACM symposium on Applied Computing, March 2004.
5. MichelaAntonelli, Beatrice Lazzerini, Francesco Marcelloni, "Segmentation and reconstruction of the lung volume in CT images", ACM Symposium on Applied Computing, 2005
6. Samuel G. Armato III, Maryellen L. Giger and Catherine J. Moran, "Computerized Detection of Pulmonary Nodules on CT Scans", RadioGraphics, vol. 19, pp. 1303–1311, 1999
7. Binsheng Zhao, Gordon Gamsu, Michelle S. Ginsberg, "Automatic detection of small lung nodules on CT utilizing a local density maximum algorithm", Journal of Applied Clinical Medical Physics, vol. 4, No. 3, summer 2003
8. N. A. Memon, A. M. Mirza, S.A.M. Gilani, "Deficiencies of Lung Segmentation Techniques using CT Scan Images for CAD", Proceedings of World Academy of Science, Engineering and Technology, vol 14, aug 2006.
9. Ayman El-Baz, Aly A. Farag, Robert Falk, Renato La Rocca, "Detection, Visualization and identification of Lung Abnormalities in Chest Spiral CT Scan: Phase-I", International Conference on Biomedical Engineering, Cairo, Egypt, 12-01-2002
10. Ayman El-Baz, Aly A. Farag, Robert Falk, Renato La Rocca, "A Unified Approach for Detection, Visualization and Identification of Lung Abnormalities in Chest Spiral CT Scan", Proceedings of Computer Assisted Radiology and Surgery, London 2003
11. M. Arfan Jaffar, Ayyaz Hussain, Anwar M. Mirza, Asmat ullah Ch, "Fuzzy Entropy and Morphology based fully automated Segmentation of Lungs from CT Scan Images", International Journal of Innovative Computing, Information and Control (IJICIC) Vol.5, No.12, December 2009

A Smart Carpet Design for Monitoring People with Dementia

Osamu Tanaka, Toshin Ryu, Akira Hayashida, Vasily G. Moshnyaga, and Koji Hashimoto

1 Introduction

1.1 Motivation

Dementia is a syndrome which deteriorates memory, thinking, behavior and the ability to perform everyday activities. There many different types of dementia although some are far more common than others (e.g. Alzheimer's disease). According to World Health Organization over 35.6 million people worldwide suffer from dementia and there are 7.7 M new cases every year [1]. The growth in the number of people having dementia is so fast that it is predicted to almost double every 20 years, to 65.7 M in 2030 and 115.4 M in 2050 [1].

Dementia care is difficult because it depends heavily on patient's behavior, surrounding and assistive equipment. Because a person with cognitive impairment can make a judgmental error if left unsupervised, caregivers must alter their behavior, working hours and sleep pattern in an effort to provide supervision. The combination of unsupervised nighttime awakenings and cognitive impairment is particularly difficult in the home setting. In many cases family members of a person suffering from dementia become physically and mentally exhausted taking care that they consider placing him or her into mental hospitals. The hospitalization tends to be long-term, averaging several years [2] and usually is very expensive [3]. Neither dementia people nor their spouses and families want to opt for the nursing home. However, almost 70 % of families are forced to do that because of the ongoing sleep disruption and stress they suffer as a result of erratic night- and day-time activity of a person with dementia [4]. Hiring a personal caregiver is already difficult and is going to be harder with the fast aging of population. Hence, the only solution to this pending crisis is the development and deployment of smart technologies that compensate for the specific physical and cognitive deficits of older people with dementia, and thereby reduce burden of family caregivers.

Numerous technologies have been developed to assist old people with cognitive impairment. Such technologies can enable remote monitoring of individuals and early detection of potential problems; so that early interventions can help older adults remain as healthy and independent as possible. The technologies can be active, enforcing a person wear a sensor, pull a cord or push an alarm; or be passive embedded into environment to detect potential problems [5]. A survey of assistive technologies and systems can be found in [6].

Active technologies usually are wearable (e.g. fall detection pendants, watches, etc.), inexpensive, and easy to identify and track. However users of active devices must always wear them. Majority of users are not willing to wear the devices. Some question the need for device. Others consider it an unwelcome admission of vulnerability [7].

Furthermore, many active devices require their users not only to remember how to use the device but also be conscious to push a button to call for help. It is assumed, the user has quick access to transmitter and is able to activate it and initiate the call, even if he or she has fallen and is immobile. However, not all old people especially those with cognitive impairment can do that. People with dementia have problems recognizing things and their purposes; they frequently forget how to use objects, tools or appliances. Even if an older adult knows how to operate the device, there is a high risk that he or she may become unconscious during or after the fall [8]. Consequently, it is necessary that tools satisfy the following requirements:

- Non-intrusiveness. The devices have to be neither worn nor operated by a person with cognitive impairment;
- Independence of device operation form the person's condition (conscious or unconscious);

O. Tanaka (✉) • T. Ryu • A. Hayashida • V.G. Moshnyaga
K. Hashimoto
Dept. Electronics Engineering and Computer Science, Fukuoka University, 8-19-1, Nanakuma, Jonan-ku, Fukuoka 814-0180, Japan
e-mail: td132011@cis.fukuoka-u.ac.jp; td132020@cis.fukuoka-u.ac.jp; td142012@cis.fukuoka-u.ac.jp; vasily@fukuoka-u.ac.jp; khashi@fukuoka-u.ac.jp

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_92,

- Automatic fall detection;
- Automatic alerting the caregiver if fall occurs.

A general approach to satisfy the aforementioned requirements is to put intelligence into the home environment, making it able to monitor human. Several solutions have been proposed. One of them is to install wireless motion sensors on inside the home, and use their readings to detect the fall [9]. Another solution is to equip a home with multiple infra-red video cameras [10] of Microsoft Kinect devices [11] and use computer-vision techniques for fall detection. Although these solutions are promising they still lack fall detection accuracy.

In this paper we focus on embedding intelligence into a carpet or a home floor and present a novel smart carpet design.

1.2 Related Research

Many proposals for the smart carpet have been reported in literature over the years. Paradisso, et al, [12] use a grid of piezoelectric wires hidden under carpet and a pair of Doppler radars to monitor motion, velocity, dynamic foot position and pressure of people. Adlesse, et al [13] utilized small rectangular plates, each of which having four load cells at the corners. Each load cell measures Ground Reaction Force (GRF) caused by the weight and inertia of the body while walking. A nearest-neighbor classifier computes the GRF profile and feds it to a Hidden Markov Model to identify an individual. An algorithm optimization for profile identification is reported in [14]. The main problem of the design is that the tiles are very rigid and expensive.

Aud, et al, [15] develop a Smart Carpet from an array of signal scavenging sensors that use energy available throughout the environment. Although this design does not have a power supply, the electronics put into the carpet make it thick and heavy. For instance, a small mat of the carpet has four blocks of foil reading electronics parts, four AD convertors, one microcontroller and one wireless transmitter. A similar drawback is inherent in Z-tiles [16], which have 24 sensors per a hexagonal tile.

To make the smart flooring unperceivable for the users, researchers from Univ. of Manchester place a 2D mesh of optical fibers beneath the carpet [17]. These fibers can detect and plot movements as pressure blend them, changing the light at the edges of the carpet. Sensors around the carpet's edges then relay signals to a computer which analyzes the 2D footstep pattern. When a change is detected - such as a sudden stumble and fall - an alarm is produced [17]. Similarly, Savio and Ludwig [18] proposed to embed electronics in textiles and interweaving them in Smart Carpet.

An extension of smart carpet to determine the weight, age, and sex of the individuals stepping across has been reported in [19]. The carpet's intelligence is derived from a layer of silicon rubber with built-in electrodes to measure changes in electrical resistance and current flow caused by someone walking across it.

A more fundamental implementation of Smart Carpet is the ELSI Smart Floor [20][21] from Aalto University, Finland. It is a copper based sensor installed under the floor surface, such as vinyl flooring or laminate and when connected to the system creates a very low power capacitive field. A person positioned over this flooring, conducts the sensor, changing the capacitive field as it done in touch panel. This solution looks well but seems to be more suitable for hospitals or nursing homes due to expensive installation. Moreover, it is difficult to maintain and repair, since if anything happen someone has to lift the entire floor.

1.3 Contribution

In this paper we present a novel design of intelligent carpet for in-home monitoring of a person with cognitive impairment. The proposed smart carpet is capable of tracking a person non-intrusively, detecting falls and alarming the caregiver wirelessly. Unlike existing systems, the system is inexpensive, provides high quality of fall detection, is easy to install, maintain and extend.

In the next section we describe architecture and prototype implementation of the smart carpet. Section 3 shows results of experimental testing. Section 4 presents conclusion and outline work for the future.

2 The Proposed Smart Carpet

2.1 Architecture

The proposed smart carpet is dedicated to unobtrusively detect location of a person in apartment or room, his/her status such as walking, staying on the floor, sitting on the floor and lying on the floor. The carpet consists of an array of mats or carpet tiles, each of which having a pressure sensor FSR406 [22]. The sensor is placed under an expanse of carpeting and connected to a corresponding row- and column-lines, as shown in Fig.1. When a pressure is put over the mat, the sensor produces an active signal on both the row and the column lines. These signals are coded and sent wirelessly to server, which decodes the signals received, evaluates the pressure pattern, and if it corresponds to a fall or another pre-defined risk event (e.g. no motion), an alert is sent to the caregiver through the internet. Because each mat has a fixed position within the carpet, the location of a person

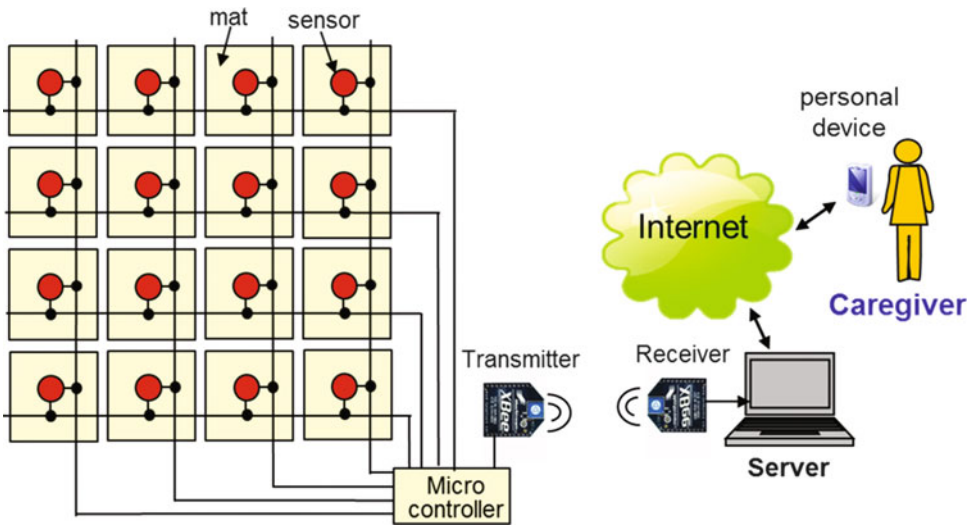


Fig. 1 An illustration of the proposed smart carpet (4x4mats)

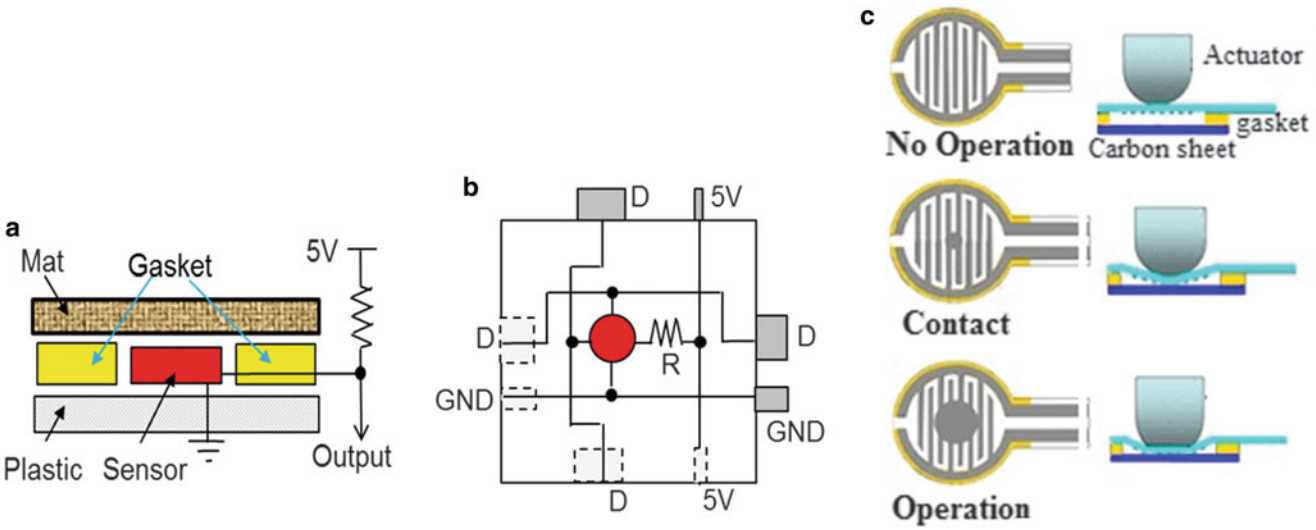


Fig. 2 (a) Cross-sectional view of a mat; (b) electronic circuit embedded into the mat; (c) sensor operation

over the mat can be easily defined through the reading of mat sensors. Walking, staying, sitting or lying down over the mats is reflected by a proper signal pattern.

2.2 Mat Design

Fig.2 (a) and (b) show cross-sectional view of a mat and the electronic circuit, embedded into the mat, respectively. As one can see the design is simple and has minimum electronics. The circuit embedded into a mat contains one small sensor FSR406 [22], one (10kΩ) resistor, and four wires crossing the mat top-down and left-right. The D ports of the mat are for the row- and column-line connections, respectively; 5 V and GND are the voltage supply and the ground ports, respectively. The left-right or top-down ports that have same label provide male-female connectors.

Fig.2(c) illustrates the sensing operation. When no pressure is applied to the mat, the sensor idles. Otherwise, the sensor's actuator makes a contact with the carbon sheet (see the center-right picture of Fig.2, c) reducing the sensor' resistance and causing voltage at the output to change. With increase in force applied to the mat and to the surface of the sensor, the contact area grows in size, increasing the current flow through the sensor (see Fig.2(c), the bottom-right picture). The output signals from the mat sensors are delivered to microcontroller and then sent wirelessly (via XBee transmitter) to the server.

The main feature of our mat design is simplicity and low cost. All mats have same shape and same number of

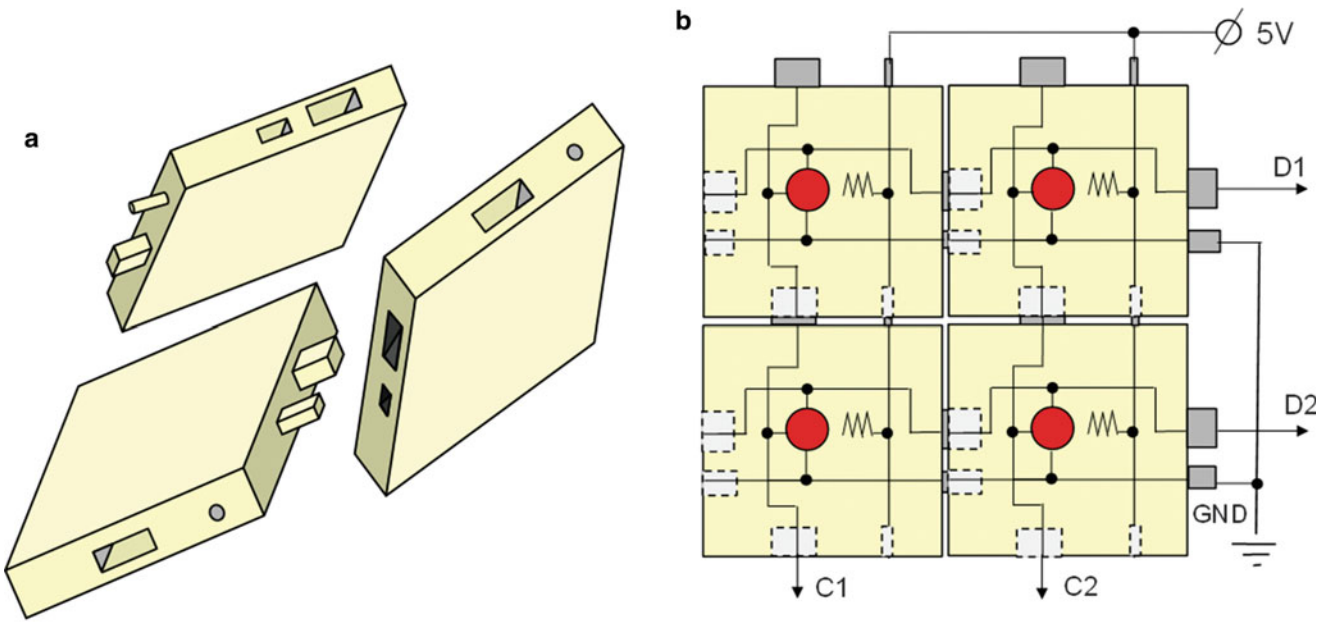


Fig. 3 Views of mats (a); the circuitry of 2x2 smart carpet (b)

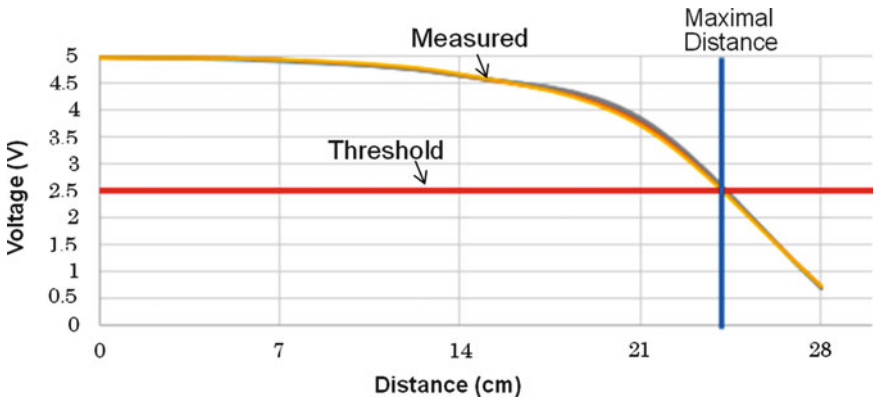


Fig. 4 The voltage variation observed at the mat ports (D and C) with the distance between the person's location and the mat center.

connectors. In order to prevent error connections, the connectors have different shapes: the voltage supply connectors are cylindrical while the others are rectangular of different size, as shown in Fig.3 (left). The mats can be grouped very easily just by attaching one to another as Lego building blocks. Fig.3 (right) exemplifies a circuit of a 2x2 smart carpet core built just by joining the mats. Although the number of mats in the carpet is limited by the number of pins on microcontroller, a 18-pin microcontroller, such as Arduino Uno or PIC, allows using up to 8x8 mats in the carpet.

In order to find the optimal mat size, we empirically measured the voltage produced at the Output port of the mat when a person (60 kg in weight) steps over at the different distance of the mat center. Fig.4 shows the results. The yellow line in this figure depicts the measured voltage; the red line shows the threshold voltage (used in microcontroller); the blue line shows the maximum distance from the center which assumed acceptable. Based on the results, we determined that a mat of 40x40cm in size can effectively sense a person weighting 60 kg or more.

2.3 Implementation

For experimental evaluation, we built a prototype smart carpet consisting of 4x4 mats, each of which having 40 cm $\times$ 40 cm in size. The prototype was implemented based on Arduino Uno R3 microcontroller and XBee wireless communication protocol. As server we used Toshiba PC (2GHz

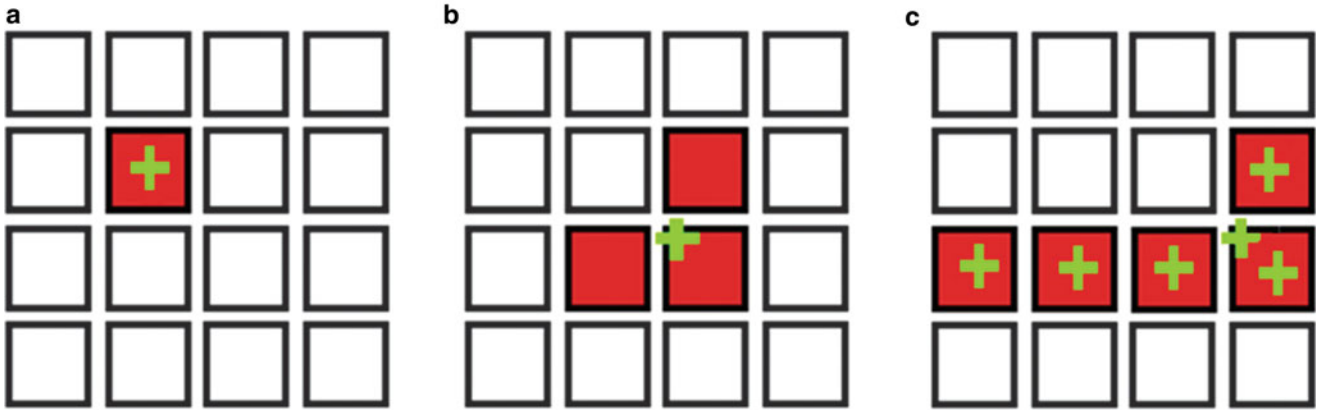


Fig. 5 Examples of patterns sensed by 4x4mats: when a person stands inside a mat (a); when a person stands on the border of 3 mats (b), and when a person lies over the area of 5 mats

Intel Core Duo CPU, 2GB memory), Windows 8 OS. The client device was 2012 Asus Nexus 7 that runs Android™ 4.3 OS and is equipped with NVIDIA Tegra3 T30L, 1.3GHz CPU, 1GB internal storage, 7-inch display. The client-server communication was implemented through Internet Socket API (ws2_32.lib), WiFi Local Area Network (LAN) and TCP/IP transport protocol. To support the OS-based control of the communication, a dedicated programming interface was also created.

We also developed application software that monitored a person by the smart carpet, assessing his/her motion, detecting falls and displaying the results on PC (server) and the mobile device (client). The software decodes data received from sensors, determines the number of mats, which produced active signals, locations of the mats, and the "motion pattern" which had been sensed. Fig.5. exemplifies three motion patterns reflected by the smart carpet when a person stays within area of a single mat, on the border of three mats and lying over the area of five mats. The cross signs in the figures show the pressured places. The mats that sensed the pressure are depicted in red.

Based on the difference between the current pattern and the pattern previously detected, and the duration of the pattern, the system determines whether the monitored person stays or walks or lies on the carpet. For instance, if both the number of active mats exceeds a given threshold ($k = 4$) and the pattern lasts longer than a pre-defined time limit (e.g. $T = 1$ sec), the system detects person's fall, generates alarm and sends it for display at the server as well as the caregiver's device (client).

The software was created in Java using Java-script (Node.js) interpreter, the Android software development kit, Notepad++, the Arduino IDE design environment, and the Dropbox file management and exchange system.

3 Evaluation and Results

To test functionality of the proposed smart carpet design, we conducted a set of experiments with 6 volunteers (faculty and students ages 20 to 60 without health concerns). The subjects were asked to stand, walk, sit, and fall over the carpet in different directions and places.

All volunteers stated that walking on the smart carpet had no perceptible difference with walking on standard carpet. The sensors embedded into the mats were not perceptible to the people as they walked across the smart carpet and successfully detected gait characteristics.

Fig.6 shows the snapshots of the notebook and Nexus 7 screens displaying the monitoring results. The red patterns on the screens show the activated sensing mats as a person walks (upper image) or lies on the carpet (bottom image). The inactive mats are shown by dark squares. The green window in the screen is activated when fall is detected. As one can see, the results, displayed both on the notebook and the Nexus 7 device, are correct. Note the results were obtained in real time as a person walked over the carpet. The delay of displaying the results on Nexus 7 device was very small (0.6 sec).

To evaluate the fall detection accuracy, we asked the volunteers to perform a series of postures, namely walking, standing, sitting, lying down in a "stretched" position, and lying down in a "tucked" position. These five scenarios were repeated four times by each subject in a random order and various directions. These test positions totaled 40 fall-

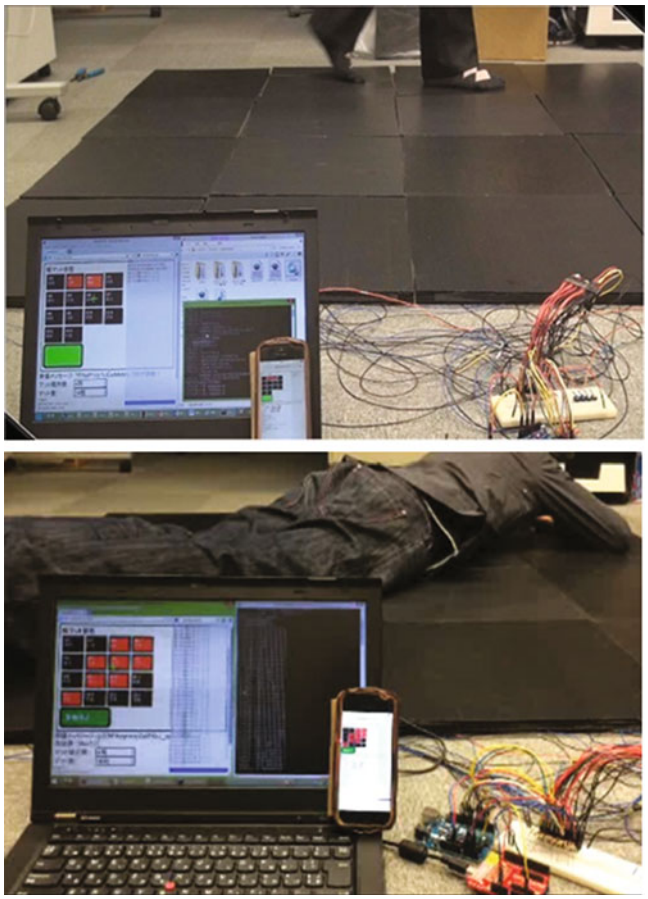


Fig. 6 Snapshots of the smart-carpet results as a person walks (top), and falls down (bottom)

simulated tasks and 60 non-fall–simulated tasks. The true positive rate for fall detection was 98 % with a false positive rate of 0.03 %.

Finally, we tested the smart carpet for different shapes (1x16 and 2x8) and different number of mats in a row/column. The experiment revealed that the carpet shape does not affect the functionality. For the tested shapes, the carpet always produced good results. The total cost of the off-the-shell components to implement the smart carpet (4x4 mats) was 228US$.

4 Conclusion

In this paper we presented a simple yet efficient design of a smart carpet for monitoring persons with cognitive impairment. The tests show that the proposed design can accurately determine the person's position, motion and falls reporting the results online (on server PC and mobile device) in real time. Although much work remains to perfect the design, we plan to install the prototype in a room of dementia person to conduct a rigorous clinical trial.

References

1. New 'flow of care' to steer dementia patients away from hospitals , The Japan Times Sep. 6, 2012
2. Hope T, Keene J, Gedling K, Fairburn C, Jacoby R. Predictors of institutionalization for people with dementia living at home with a carer. Int J of Geriatr Psychiatry, vol.13, pp. 682–90, 1998.
3. CareGiver Technology: independent living through technology (for family CareGiver) http://www.caregivertechnology.com/for-family-caregivers/
4. C.P.Pollak, D.Perlick, G.Alexopoulos, A.Gonzales, "Disruptive nigh-ttime behaviors in elder caregiver pairs", Sleep Res, vol.23, 305, 1994
5. M.J.Raintz, F.M.Skubic, G.Alexander, M.A.Aud, et al, "Improving Nurse Care Coordination with technology", Computers, Informatics. Nursing, vol.28, no.6, pp.325-332, 2010.
6. A.J.Bharucha, V.Anand, J.Forizzi, M.A.Dew, C.F.Reynolds, S. Stevens, H.Wactlar, "Intelligent assistive technology applicationsto dementia care; Current capabilities, limitations and future challenges", Am. J. Geriatric Psychiatry, vol.17, no.2, pp. 88-104, Feb. 2009
7. E.J. Porter, "Wearing and using personal emergency emergency response system buttons: Older frail windows' intentions". J.of Gerontological Nursing, vol.31, no.10, pp.26-33, 2005
8. J.A.Langlois, S.R.Kegler, J.A.Gotsch, et al., "Traumatic brain injury-related hospital discharges: Results from a 14-state surveillance system", 2003, MMWR Surveilance Summaries 52(SS04), Centers for Desease Control and prevention, available from website: www.cdc.gov/MMWR/preview/mmwrhtml/ss5204a1.html
9. Just Checking, Helping people to stay at home, available online at http://www.justchecking.com.au/
10. K.Williams, A.Arhur, M.Niedens, et al, "In-Home Monitoring Support for Dementia Caregivers: A Feasibility Study", Clin Nurs Res., vol. 22, no.2, pp.139-150, May 2013.
11. M.Rantz, T.S.banerjee, E.Cattoor, S.D.Scott, M.Skubic, and M. Popescu, Automated fall detection with quality improvement "rewind" to reduce falls in hospital rooms. J.Gerontological Nursing, v.40, No.1, 2014.
12. J.Paradiso, K. H. C. Abler, M. Reynolds, "The magic carpet: Physical sensing for immersive environments", CHI97, pp. 277-278, 1997
13. M. Addlesee, A. Jones, F. Livesey, and F. Samaria, "Orl active floor," IEEE Personal Communications, vol. 4.5, pp. 35 – 41, 1997
14. R. Orr and G. Abowd, "The smart floor: A mechanism for natural user identification and tracking," Proc. 2000 Conf. Human Factors in Computing Systems, 2000.
15. M.A. Aud, C.C. Abbott, H.W. Tyrer, R.V. Neelgund, U.G. Shriniwar, A.Mohammed, K.K.Devarakonda, "Smart Carpet: Developing a Sensor System to Detect Falls and Summon Assistance", J. Gerontological Nursing, vol.36, Issue 7, pp.8-12, July 2010
16. R. Richardson, J. Paradiso, K. Leydon, and M. Fernstrom, "Z-tiles: Building blocks for modular, pressure-sensing," in CHI2004, 2004.

17. S.Tutoru, “Smart Carpet: detecting falls and predicting mobility problems”, Prescouter Journal, Sep.6, 2012.
18. D.Savio, T.Ludwig, “Smart Carpet: A Footstep Tracking Interface”, 21st Int. Conf. Advanced Information Networking and Application Workshops (AINAW’07), 2007.
19. D.Murph, Intelligent carpet can autodiscriminate, July 26, 2006, available from http://smart-home-blog.com/2006/07/27/intelligent-carpet/
20. http://telecareaware.com/smart-flooring-that-can-simplify-alerting/
21. A. Ropponen, H. Rimminen, & R. Sepponen. “Robust system for indoor localization and identification for the healthcare environment”, Wireless Personal Communications, vol. 59, no. 1, pp.57-71, 2010.
22. FSR 406 Data Sheet, FSR 400 Square Force Sensing Resistor, Interlink Electronics, available at www.interlinkelectronics.com.

Rationalisation of the Maintenance Process of Transport Telematics System Comprising two Types of Periodic Inspections

Adam Rosinski

1 Introduction

The term "Telematics" first appeared in literature in the beginning of 70's last century. It was coined using two French words: telecommunications (Fr. télécommunications) and information technology (Fr. informatique). It did not find its way to the broader public immediately after the term was coined. Only when EU programmes aimed at developing telematics and deploying it in different areas, the term became more recognisable and popular. This dates back to the 90's. Today the term telematics is used to describe sciences integrating telecommunications and IT solutions.

From among different areas of application, one of the biggest and most developing (in Poland, Europe and worldwide) is transport. Transport telematics marked its presence in Polish publications as late as in mid-nineties.

Transport telematics is defined as a field of knowledge and technical activity integrating IT with telecommunications, intended to address transport systems' needs. Many bodies of work concerning that subject were published [7], [13], [17] including those on intelligent transport [6], [19].

The issue of maintaining electronic equipment, particularly those used in transport telematics systems is an important problem [2], [4]. This stems from the fact correct reliability and operating parameters have to be assured. Many renowned papers have already been written on the matter [3], [9], [16]. By carrying out an adequate reliability analysis [5], [8], [18] of systems, their reliability structures are determined which provide correct reliability parameters. This applies both to the entire system, as well as its constituting elements e.g. power supply [11], [15] and transmission media [1], [10], [14]). This approach guarantees the designed system becomes more reliable. It does not, however, assure high enough availability of the system [12]. Hence, maintenance analysis has to be carried out taking account of selected operating properties of the systems. It will enable to propose optimum parameters of maintenance strategy, including rationalisation of routine inspections and their length relative to requirements to those systems in respect of their availability in the transport process.

2 The process of maintaining transport telematics systems

An analysis of transport telematics systems maintenance proves they could be reflecting one of the following states (fig. 1):

- usage state (S_{PZ}),
- repair state (S_B).

Denotations in figures 1:

λ - failure rate,

μ - repair rate.

The following availability rates were determined for maintenance process shown in figure 1.

$$K_g = \frac{T_m}{T_m + T_n} \tag{1}$$

where: T_m- mean correct operation time between failures,
T_n- mean time to repair.

Fig. 1 presents graph showing switching between states which does not include all possible and actual state. Hence the following states were added (fig. 2):

- state S_{ZB1} (basic servicing required by specification of type I inspection),

A. Rosinski (✉)
Faculty of Transport, Warsaw University of Technology,
Warsaw, Poland
e-mail: adro@wt.pw.edu.pl

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_93,

– state S_{ZB2} (basic servicing required by specification of type II inspection).

Denotations in figures 2:

$R_O(t)$ – the function of probability of system staying in usage state S_{PZ},

$Q_{ZB1}(t)$ – the function of probability of system staying in state of I type inspection S_{ZB1},

$Q_{ZB2}(t)$ – the function of probability of system staying in state of II type inspection S_{ZB2},

$Q_B(t)$ – the function of probability of system staying in state of repair S_B,

λ_{ZB1} – transition rate from usage state S_{PZ} to state of type I inspection S_{ZB1},

λ_{ZB2} – transition rate from usage state S_{PZ} to state of type II inspection S_{ZB2},

μ_{PZ1} – transition rate from state of type I inspection S_{ZB1} to usage state S_{PZ},

μ_{PZ2} – transition rate from state of type II inspection S_{ZB2} to usage state S_{PZ},

λ_{B1} – transition rate from state of type I inspection S_{ZB1} to state of repair S_B,

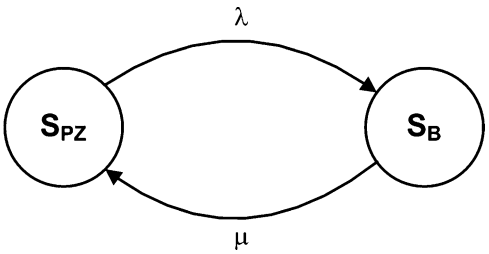


Fig. 1 Graph showing switching between usage and repair states

λ_{B2} – transition rate from state of type II inspection S_{ZB2} to state of repair S_B,

λ_B – transition rate from usage state S_{PZ} to state of repair S_B,

μ_{B1} – transition rate from state of repair S_B to state of type I inspection S_{ZB1}.

The usage state S_{PZ} is a state whereby transport telematics system completes all tasks for which it was designed. The state of type I inspection S_{ZB1} is a state whereby a periodic inspection is performed consisting of basic checks. The state of type II inspection S_{ZB2} is a state whereby a periodic inspection is performed consisting of wider range of checks. State of repair Q_B is a state whereby the system's state of full ability is restored.

When transport telematics system is in usage state S_{PZ}, then it is possible for the system to transition to the state of type I inspection S_{ZB1} at a rate λ_{ZB1}. When the system is in state of type I inspection S_{ZB1} it is possible (after basic checks are completed) it transitions to the usage state S_{PZ}.

When state of type I inspection S_{ZB1} is active and failure demanding repair is detected then system transitions to the state of repair Q_B at a rate λ_{B1}. Reverse transition back to state of type I inspection S_{ZB1} from state of repair Q_B is possible provided measures restoring system's state of full ability are undertaken.

When transport telematics system is in usage state S_{PZ}, then it is possible for the system to transition to the state of type II inspection S_{ZB2} at a rate λ_{ZB2}. When the system is in state of type II inspection S_{ZB2} it is possible (after wider checks are completed) it transitions to the usage state S_{PZ}.

When state of type II inspection S_{ZB2} is active and failure demanding repair is detected then system transitions to the state of repair Q_B at a rate λ_{B2}. After restoring system's state of full ability it transitions to the state of type I inspection S_{ZB1}.

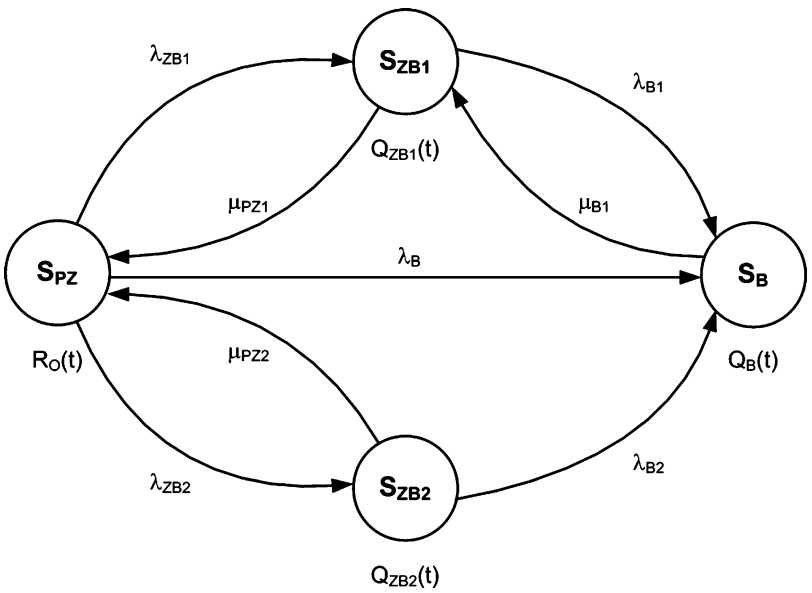


Fig. 2 Relationships in the system

When transport telematics system is in usage state S_{PZ}, then it is possible for the system to transition, in case of failure, to the state of repair S_B at a rate λ_B.

The system illustrated in fig. 2 may be described by the following Chapman–Kolmogorov equations:

$$\begin{aligned}
R_0^{'}(t) &= -\lambda_{ZB1}\cdot R_0(t) + \mu_{PZ1}\cdot Q_{ZB1}(t) - \lambda_{ZB2}\cdot R_0(t)\\
&\quad +\mu_{PZ2}\cdot Q_{ZB2}(t) - \lambda_B\cdot R_0(t)\\
Q_{ZB1}^{'}(t) &= \lambda_{ZB1}\cdot R_0(t) + \mu_{PZ1}\cdot Q_{ZB1}(t) - \lambda_{B1}\cdot Q_{ZB1}(t)\\
&\quad +\mu_{B1}\cdot Q_B(t)\\
Q_{ZB2}^{'}(t) &= \lambda_{ZB2}\cdot R_0(t) + \mu_{PZ2}\cdot Q_{ZB2}(t) - \lambda_{B2}\cdot Q_{ZB2}(t)\\
Q_B^{'}(t) &= \lambda_{B1}\cdot Q_{ZB1}(t) + \lambda_{B2}\cdot Q_{ZB2}(t) - \mu_{B1}\cdot Q_B(t)\\
&\quad +\lambda_B\cdot R_0(t)
\end{aligned} \tag{2}$$

Given the initial conditions:

$$\begin{aligned}
R_0(0) &= 1\\
Q_{ZB1}(0) &= Q_{ZB2}(0) = Q_B(0) = 0
\end{aligned} \tag{3}$$

Laplace transform yields the following system of linear equations:

$$\begin{aligned}
s\cdot R_0^*(s) - 1 &= -\lambda_{ZB1}\cdot R_0^*(s) + \mu_{PZ1}\cdot Q_{ZB1}^*(s)\\
&\quad -\lambda_{ZB2}\cdot R_0^*(s) + \mu_{PZ2}\cdot Q_{ZB2}^*(s) - \lambda_B\cdot R_0^*(s)\\
s\cdot Q_{ZB1}^*(s) &= \lambda_{ZB1}\cdot R_0^*(s) - \mu_{PZ1}\cdot Q_{ZB1}^*(s) - \lambda_{B1}\cdot Q_{ZB1}^*(s)\\
&\quad +\mu_{B1}\cdot Q_B^*(s)\\
s\cdot Q_{ZB2}^*(s) &= \lambda_{ZB2}\cdot R_0^*(s) - \mu_{PZ2}\cdot Q_{ZB2}^*(s) - \lambda_{B2}\cdot Q_{ZB2}^*(s)\\
s\cdot Q_B^*(s) &= \lambda_{B1}\cdot Q_{ZB1}^*(s) + \lambda_{B2}\cdot Q_{ZB2}^*(s) - \mu_{B1}\cdot Q_B^*(s)\\
&\quad +\lambda_B\cdot R_0^*(s)
\end{aligned} \tag{4}$$

Probability of system staying in specific state denoted symbolically (Laplace) reads as follows:

$$\begin{aligned}
R_0^*(s) &= -\frac{b_1\cdot b_2\cdot s + b_1\cdot b_2\cdot\mu_{B1} - b_2\cdot\lambda_{B1}\cdot\mu_{B1}}{a\cdot b_2\cdot\lambda_{B1}\cdot\mu_{B1} - a\cdot b_1\cdot b_2\cdot\mu_{B1} - a\cdot b_1\cdot b_2\cdot s + b_2\cdot s\cdot\lambda_{ZB1}\cdot\mu_{PZ1} + b_1\cdot s\cdot\lambda_{ZB2}\cdot\mu_{PZ2} +}\\
&\quad +b_2\cdot\lambda_B\cdot\mu_{B1}\cdot\mu_{PZ1} + b_2\cdot\mu_{B1}\cdot\lambda_{ZB1}\cdot\mu_{PZ1} + b_1\cdot\mu_{B1}\cdot\lambda_{ZB2}\cdot\mu_{PZ2} - \lambda_{B1}\cdot\mu_{B1}\cdot\lambda_{ZB2}\cdot\mu_{PZ2} +\\
&\quad +\lambda_{B2}\cdot\mu_{B1}\cdot\mu_{PZ1}\cdot\lambda_{ZB2}\\
Q_{ZB1}^*(s) &= -\frac{b_2\cdot s\cdot\lambda_{ZB1} + b_2\cdot\lambda_B\cdot\mu_{B1} + b_2\cdot\mu_{B1}\cdot\lambda_{ZB1} + \lambda_{B2}\cdot\mu_{B1}\cdot_{ZB2}}{a\cdot b_2\cdot\lambda_{B1}\cdot\mu_{B1} - a\cdot b_1\cdot b_2\cdot\mu_{B1} - a\cdot b_1\cdot b_2\cdot s + b_2\cdot s\cdot\lambda_{ZB1}\cdot\mu_{PZ1} + b_1\cdot s\cdot\lambda_{ZB2}\cdot\mu_{PZ2} +}\\
&\quad +b_2\cdot\lambda_B\cdot\mu_{B1}\cdot\mu_{PZ1} + b_2\cdot\mu_{B1}\cdot\lambda_{ZB1}\cdot\mu_{PZ1} + b_1\cdot\mu_{B1}\cdot\lambda_{ZB2}\cdot\mu_{PZ2} - \lambda_{B1}\cdot\mu_{B1}\cdot\lambda_{ZB2}\cdot\mu_{PZ2} +\\
&\quad +\lambda_{B2}\cdot\mu_{B1}\cdot\mu_{PZ1}\cdot\lambda_{ZB2}\\
Q_{ZB2}^*(s) &= -\frac{b_1\cdot s\cdot\lambda_{ZB2} + b_1\cdot\mu_{B1}\cdot\lambda_{ZB2} - \lambda_{B1}\cdot\mu_{B1}\cdot\lambda_{ZB2}}{a\cdot b_2\cdot\lambda_{B1}\cdot\mu_{B1} - a\cdot b_1\cdot b_2\cdot\mu_{B1} - a\cdot b_1\cdot b_2\cdot s + b_2\cdot s\cdot\lambda_{ZB1}\cdot\mu_{PZ1} + b_1\cdot s\cdot\lambda_{ZB2}\cdot\mu_{PZ2} +}\\
&\quad +b_2\cdot\lambda_B\cdot\mu_{B1}\cdot\mu_{PZ1} + b_2\cdot\mu_{B1}\cdot\lambda_{ZB1}\cdot\mu_{PZ1} + b_1\cdot\mu_{B1}\cdot\lambda_{ZB2}\cdot\mu_{PZ2} - \lambda_{B1}\cdot\mu_{B1}\cdot\lambda_{ZB2}\cdot\mu_{PZ2} +\\
&\quad +\lambda_{B2}\cdot\mu_{B1}\cdot\mu_{PZ1}\cdot\lambda_{ZB2}\\
Q_B^*(s) &= -\frac{b_1\cdot b_2\cdot\lambda_B + b_2\cdot\lambda_{B1}\cdot\lambda_{ZB1} + b_1\cdot\lambda_{B2}\cdot\lambda_{ZB2}}{a\cdot b_2\cdot\lambda_{B1}\cdot\mu_{B1} - a\cdot b_1\cdot b_2\cdot\mu_{B1} - a\cdot b_1\cdot b_2\cdot s + b_2\cdot s\cdot\lambda_{ZB1}\cdot\mu_{PZ1} + b_1\cdot s\cdot\lambda_{ZB2}\cdot\mu_{PZ2} +}\\
&\quad +b_2\cdot\lambda_B\cdot\mu_{B1}\cdot\mu_{PZ1} + b_2\cdot\mu_{B1}\cdot\lambda_{ZB1}\cdot\mu_{PZ1} + b_1\cdot\mu_{B1}\cdot\lambda_{ZB2}\cdot\mu_{PZ2} - \lambda_{B1}\cdot\mu_{B1}\cdot\lambda_{ZB2}\cdot\mu_{PZ2} +\\
&\quad +\lambda_{B2}\cdot\mu_{B1}\cdot\mu_{PZ1}\cdot\lambda_{ZB2}
\end{aligned} \tag{5}$$

where:

$$\begin{aligned}
a &= s + \lambda_{ZB1} + \lambda_{ZB2} + \lambda_B\\
b_1 &= s + \mu_{PZ1} + \lambda_{B1}\\
b_2 &= s + \mu_{PZ2} + \lambda_{B2}
\end{aligned}$$

Solving above equations in time domain is a next step of the analysis which has not been included in this paper.

3 Modelling of process of power supply maintenance

Computer simulation and computer-aided analysis facilitate to relatively quickly determine the influence of change in reliability-maintenance parameters of individual components on reliability of the entire system. Of course, the reliability structure of both the entire system and its components has to be known beforehand.

Using computer aided allows to perform the calculation of the value of probability of system staying in usage state R_O. That procedure is illustrated with below example.

Example

The following quantities were defined for the system:

– test duration - 1 year (values of this parameter is given in [h]):

$$t = 8760[h]$$

– transition rate from usage state SPZ to state of type I inspection SZB1:

$$\lambda_{ZB1} = 0,000006$$

– transition rate from usage state S_{PZ} to state of type II inspection S_{ZB2}:

$$\lambda_{ZB2} = 0,000001$$

– transition rate from usage state S_{PZ} to state of repair S_B:

$$\lambda_B = 0,0000001$$

– transition rate from state of type I inspection S_{ZB1} to state of repair S_B:

$$\lambda_{B1} = 0,000001$$

– transition rate from state of type II inspection S_{ZB1} to state of repair S_B:

$$\lambda_{B2} = 0,000006$$

– transition rate from state of repair S_B to state of type I inspection S_{ZB1}:

$$\mu_{B1} = 0,1$$

For above input values using equation (5) we get:

$$R_0^*(s) = \frac{\begin{array}{l}3,00003 \cdot 10^{12} \cdot s + 3 \cdot 10^{12} \cdot \mu_{PZ1} + 5 \cdot 10^{18} \cdot s^2 \cdot \mu_{PZ1} + 5 \cdot 10^{18} \cdot s^2 \cdot \mu_{PZ2} + \\ +5,00035 \cdot 10^{17} \cdot s^2 + 5 \cdot 10^{18} \cdot s^3 + 5,0003 \cdot 10^{17} \cdot s \cdot \mu_{PZ1} + 5,00005 \cdot 10^{17} \cdot s \cdot \mu_{PZ2} + \\ +5 \cdot 10^{17} \cdot \mu_{PZ1} \cdot \mu_{PZ2} + 5 \cdot 10^{18} \cdot s \cdot \mu_{PZ1} \cdot \mu_{PZ2}\end{array}}{\begin{array}{l}2,1300213 \cdot 10^7 \cdot s + 5,000355 \cdot 10^{17} \cdot s^2 \cdot \mu_{PZ1} + 5,000355 \cdot 10^{17} \cdot s^2 \cdot \mu_{PZ2} + \\ +5 \cdot 10^{18} \cdot s^3 \cdot \mu_{PZ1} + 5 \cdot 10^{18} \cdot s^3 \cdot \mu_{PZ2} + 6,5502785 \cdot 10^{12} \cdot s^2 + 5,000705 \cdot 10^{17} \cdot s^3 + \\ +5 \cdot 10^{18} \cdot s^4 + 3,500033 \cdot 10^{12} \cdot s \cdot \mu_{PZ1} + 3,0500305 \cdot 10^{12} \cdot s \cdot \mu_{PZ2} + \\ +5,000005 \cdot 10^{17} \cdot s \cdot \mu_{PZ1} \cdot \mu_{PZ2} + 5 \cdot 10^{18} \cdot s^2 \cdot \mu_{PZ1} \cdot \mu_{PZ2}\end{array}}$$

Assuming $\mu_{PZ1} = 0,1$, $\mu_{PZ2} = 0,2$ and using Laplace's equation we get:

$$\begin{array}{l}R_0(t) = 0,000197588682639 \cdot e^{-0,099703548965 \cdot t} + \\ +0,0000050004249 \cdot e^{-0,200007 \cdot t} \\ -0,00013559253646 \cdot e^{-0,10030355097 \cdot t} + \\ +0,999933003428851\end{array}$$

Finally obtained is:

$$R_O = K_g = 0,999933003$$

The relationships (5) were used to determine impact of transition rates from state of type I and II inspection to the usage state (μ_{PZ1} and μ_{PZ2}) on probability of system staying in usage state K_g. Rates μ_{PZ1} and μ_{PZ2} should be read as inverse times t_{PZ1} and t_{PZ2} and which determine the time taken to restore the usage state. That way, presented analysis could be deployed in practice.

4 Summary

The rationalisation of two type periodic inspections part of maintenance strategy aimed at maximising the availability rate concerning transport telematics systems could optimise K_g by obtaining information regarding inspection rates. Given rapid development of electronic systems employed in transport and the diagnostics subsystem they use, it is fair to say the trend in designing tends towards developing and implementing systems with diagnosing and therapeutic capabilities. They will have overseen the system and taken ever complex (factoring in the reliability theory, the maintenance theory) therapeutic measures, preventing the system from collapsing into the state of reached operational ability.

The presented method of optimising the maintenance process requires knowledge on theoretical notions behind the reliability and maintenance theory. Hence, there is a need to develop a computer application which would

determine optimum period inspection rates. A solution of this calibre would have enabled the users and maintenance officers to quickly and correctly deploy the developed method. It would be the next step in research work facilitating faster implementation of presented solutions into practice.

References

1. Bajda, A., Wrażeń, M., Laskowski, D.: Diagnostics the quality of data transfer in the management of crisis situation. Electrical Review 87(9A), 72-78 (2011)
2. Będkowski, L., Dąbrowski, T.: The basis of exploitation, part II: The basis of exploational reliability. Military Academy of Technology, Warsaw (2006)
3. Duer, S., Zajkowski, K., Duer, R., Paś, J.: Designing of an effective structure of system for the maintenance of a technical object with the using information from an artificial neural network. Neural Computing & Applications (2012) DOI: 10.1007/s00521-012-1016-0.
4. Dyduch, J., Paś, J., Rosiński, A.: Basics of maintaining electronic transport systems. Publishing House of Radom University of Technology, Radom (2011)
5. Epstein, B., Weissman, I.: Mathematical models for systems reliability. CRC Press / Taylor & Francis Group (2008)
6. Ghosh, S., Lee, T.: Intelligent Transportation Systems: Smart and Green Infrastructure Design. CRC Press / Taylor & Francis Group (2010)
7. Kasprzyk, Z.: Delivering payment services through manual toll collection system. In: Mikulski J. (eds.) Telematics in the transport environment, given as the monographic publishing series – „Communications in Computer and Information Science", Vol. 329, pp. 60–68. Springer-Verlag, Berlin Heidelberg 2012.
8. Kołowrocki, K., Soszyńska-Budny, J.: Reliability and safety of complex technical systems and processes. Springer, London (2011)
9. Nakagawa, T.: Advanced Reliability Models and Maintenance Policies. Springer-Verlag, London (2008)
10. Paś, J., Duer, S.: Determination of the impact indicators of electromagnetic interferences on computer information systems. Neural Computing & Applications (2012) DOI:10.1007/s00521-012-1165-1.
11. Rosinski, A., Dabrowski, T.: Modelling reliability of uninterruptible power supply units. Eksploatacja i Niezawodnosc – Maintenance and Reliability, Vol.15, No. 4, 409-413 (2013)
12. Rosiński, A.: Maintenance strategy maximising availability rate. Archives of Transport, Vol. XXIV, No 4, 553–569 (2012)
13. Siergiejczyk, M., Paś, J., Rosiński, A.: Application of closed circuit television for highway telematics. In: Mikulski J. (eds.) Telematics in the transport environment, given as the monographic publishing series – „Communications in Computer and Information Science", Vol. 329, pp. 159–165. Springer-Verlag, Berlin Heidelberg (2012)
14. Siergiejczyk, M., Rosinski, A.: Reliability analysis of electronic protection systems using optical links. In: Zamojski W., Kacprzyk J., Mazurkiewicz J., Sugier J., Walkowiak T. (ed.) Dependable Computer Systems, given as the monographic publishing series – „Advances in intelligent and soft computing", Vol. 97, pp. 193-203. Springer-Verlag, Berlin Heidelberg (2011).
15. Siergiejczyk, M., Rosinski, A.: Reliability analysis of power supply systems for devices used in transport telematic systems. In: Mikulski J. (eds.) Modern Transport Telematics, given as the monographic publishing series – „Communications in Computer and Information Science", Vol. 239, pp. 314-319. Springer-Verlag, Berlin Heidelberg (2011)
16. Stapelberg, R. F.: Handbook of Reliability, Availability, Maintainability and Safety in Engineering Design. Springer-Verlag, London (2009)
17. Sumila, M.: Selected aspects of message transmission management in its systems. In: Mikulski J. (eds.) Telematics in the transport environment, given as the monographic publishing series – „Communications in Computer and Information Science", Vol. 329, pp. 141–147. Springer-Verlag, Berlin Heidelberg (2012)
18. Verma, A.K., Ajit, S., Karanki D.R.: Reliability and safety engineering. Springer, London (2010)
19. Williams, B.: Intelligent transport systems standards. Artech House (2008)

An Adaptive Controller of Traffic Lights using Genetic Algorithms

Kalum Udagepola, Belal Ali Alshami, Naveed Afzal, and Xiang Li

1 Introduction

Road traffic system is one of the primary modes of private and public transport. Due to the massive growth of modern cities in terms of population and number of vehicles on the roads, has put an enormous pressure on the road traffic system. Therefore, there is a constant need for more effective and efficient traffic management systems [4]. Many modern cities such as Beijing, London and Paris are suffering from the traffic congestion problem. This traffic congestion problem has also resulted into a lot amount of time wastes and energy. Apart from economic impact, there is also an environmental impact like pollution of CO_2 and black carbon emissions that cause a high temperature on ozone and many serious illnesses [6]. In 2007, the Urban Mobility Report estimated that the total annual cost of congestion for the 75 US urban areas at 89.6 billion dollars, 4.5 billion hours of delay and 6.9 billion gallons of excess fuel consumed [7].

One way to alleviate this problem of roads and highway's congestion is to build new high capacity roads and highways but this solution is very expensive, time-consuming and in majority of scenarios, it is not feasible due to space limitations. On the other hand, the most realistic solution is to optimize the usage of existing roads and highways by better management and control strategy of traffic light systems. Therefore, in this paper, we are proposing a model based on Genetic Algorithms that can optimize the usage of traffic light systems and reduce time wastes, money and energy.

2 Related Work

In order to control and manage this problem, most present traffic light control systems use static models of flow patterns at intersections, where the shape and cycle of flow at regular or rush hours are assumed to be the same at all times. Some researches fixed the cycle time and over flow of each intersection and changed only on green split [1, 2, 8].

Our objective in this paper is to reduce the delay and optimize the flow ratio at intersections. In [1], the probability to enter main road is 80 %, and side road is 20 % (with the exception of intersection of two main roads where the probabilities of the choice of the two target roads are equal). A fixed cycle time, however, causes crowding in junctions with lost time because in each intersection, the length of overflow queue will grow from cycle to cycle [5]. The assumption in the older delay models is that the overflow queue is static, constant from cycle to cycle [5, 2, 8]. The offset has proven that we can attain better results by using the coordinating control method when the distance between neighboring intersections is not more than 800 meters [8].

3 Our Approach

In developing countries, most traffic light control systems are pre-timed. These systems correspond to the predicated traffic changes via preset changes on a time clock. Moreover, there are different places that have congestion at junctions where a study is required to reduce the delay time. Our work presents a dynamic Genetic Algorithm

K. Udagepola (✉) • N. Afzal
Faculty of Computing and Information Technology, North Branch, King Abdulaziz University, Jeddah, Saudi Arabia

B.A. Alshami
College of Computing and Information Technology, Arab Academy for Science and Technology and Maritime Transport, Egypt, Alexandria

X. Li
College of Computer and Science Technology, Harbin Engineering University, Harbin, PR China

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_94,

Traffic Signal Timing Management System (GATSTMS) that adapts and coordinates most traffic light models. In order to execute the GATSTMS there are two files as inputs, the structure file which contains structure of traffic lights where the user can define the parameters such as number of intersections N, number of phases (group lights) P at each intersection, number of roads R connected to the intersections, and number of lanes movement L at each road. There is also the data file that contains the traffic lights data such as arrival, departure rate for each lane and vehicles distribution for each intermediate lane. In addition, the user can change phase shape and the sequence order. The GATSTMS can get the optimal solution for cycle times, green splits and offset times.

3.1 Genetic Algorithm in Traffic Signal Optimization

The Genetic Algorithm (GA) is a search technique that solves optimization problems. The structure of GA is same as the evolution algorithm. In order to solve any problem as traffic lights using GA, four questions should be answered [3]:

- What fitness function is used in traffic lights?
- How an individual is represented?
- How individuals are selected?
- How individuals do reproduce?

In GA, we represent a solution of our problem as a chromosome. The main idea is to maintain a population of chromosomes that symbolize acceptable solutions to the particular problem that evolves over continuous iterations (generations) through a process of competition and controlled variation. Each chromosome in the population has an associated fitness to control which chromosomes. We should use to form new ones in the competition process. The new chromosomes are created using genetic operators such as crossover and mutation. In GA model [9], we represent the chromosome as binary digits with length *Ch_L* for each intersection and calculate it as follows:

$$Ch\_L = (N + \sum_{i=1}^{N}(G(i) + R^{i}_{intermediate})) \times 8 \qquad (1)$$

Chromosome length depends on a total number of intersections N, a number of lighting groups $G(i)$ inside i^{th} intersection and a number of intermediate roads $R^{i}_{intermediate}$ connected to i^{th} intersection. Each chromosome represents parameters Cycle Times (*CT*), Offset Times (*FT*) and Green split Times (*GT*). Eight multiplies each parameter because the highest value for *CT* is 256 seconds and some of intersections may have one *GT*.

Table 1 Directions lane movement for one intersection

lane move / direction	Right=1	Left=2	Direct=3
N=1	NE=11	NW=12	NS=13
E=2	ES=21	EN=22	EW=23
S=3	SW=31	SE=32	SN=33
W=4	WN=41	WS=42	WE=43

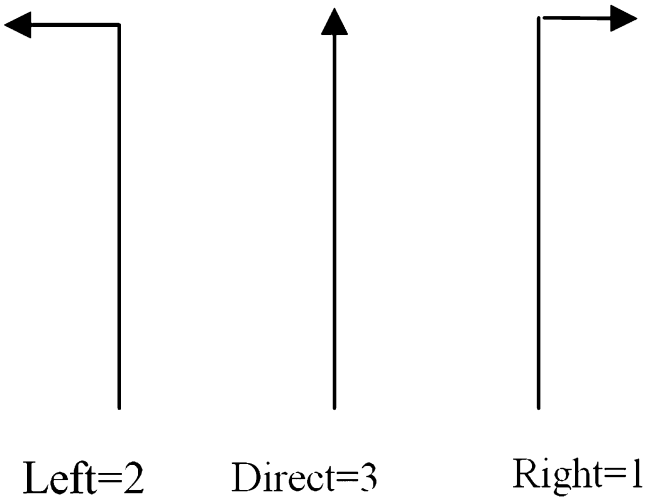


Fig. 1 Type of lanes for one road

4 Proposed Model

The proposed model is a general dynamic model based on GA, where the number of intersections is in the range (*2-N*), number of road R that connected to the desired intersection is in the range (1-4) ($N = 1$, $E = 2$, $S = 3$, $W = 4$) see Table 1, number of group lights P for each intersection are in the range (1-4). If group light is 1 this means there is one stage for the cycle in the desired intersection. In addition, the lanes movement type *L* for each road is in the range (1-3) as shown in Figure 1. We assume that each lane of the input road has detector to detect the arrival vehicles, flow ratio dependence on the priority of the road, and average speed for each road will optimize the offset time. In this model, we assume that every vehicle is on the desirable lane of the road.

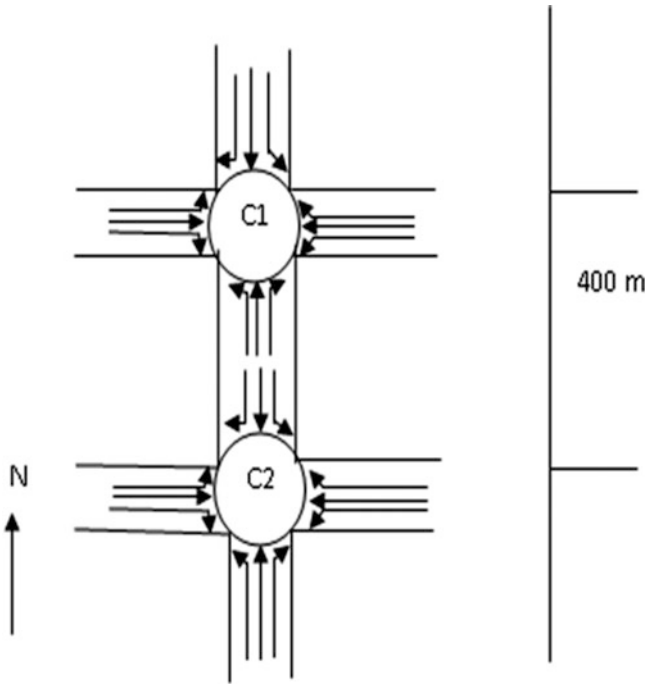


Fig. 2 Suggested network

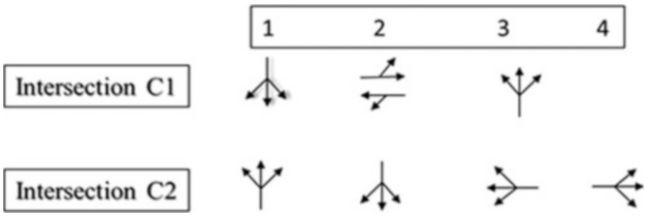


Fig. 3 Sequence order for network

5 Simulation and Results

The proposed model of the GATSTMS is simulated using input from two files that contain structure and data of traffic lights. The structure file contains all parameters needed to define the dynamic model, while the data file contains the data needed for this model. We assume the detector detects flow per hour, and then the GATSTMS update the data file with arrival and departure for each lane. In Figure 2, the modeled network consists of two intersections (C1 and C2) with three phases for C1, four phases for C2, also, each intersection consists of four roads (N, E, S and W).

Figure 3 shows the phases for intersection C1 and intersection C2, where the sequence order in the GATSTMS select phase 1 from intersection C1 and C2, then select phase 2 from intersection C1 and intersection C2, etc.

The results of the simulation show that the GATSTMS can deliver an optimal solution (see Figures 4 and 5). We test

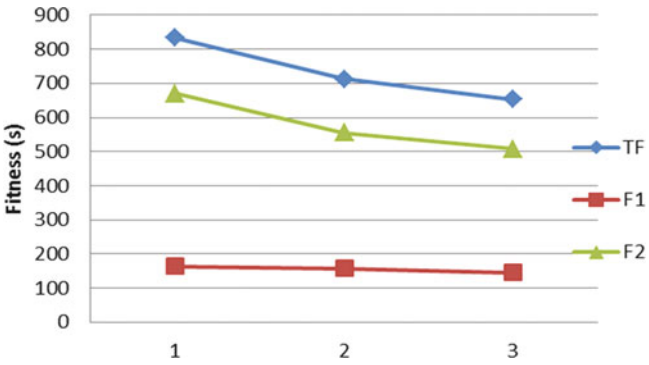


Fig. 4 Fitness's VS generation

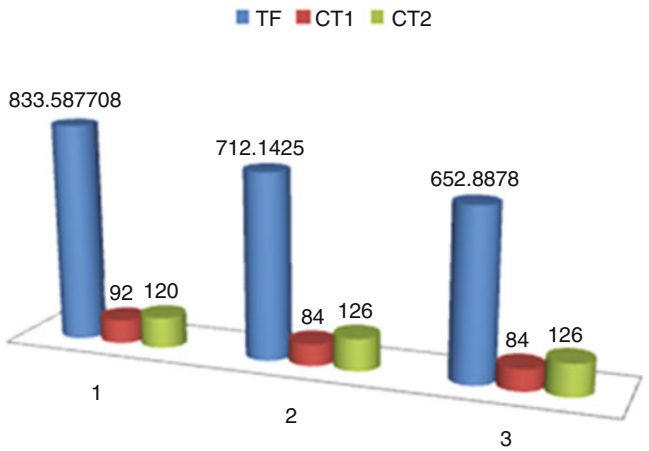


Fig. 5 Fitness for optimal CT

our network under values 48 < =CT < =120, 16 < =GT < =40 for intersections C1 and C2. Figure 4 shows the results for the minimum of Total fitness (TF), minimum Fitness for intersection C1 (F1) and minimum fitness for intersection C2 (F2) for each generation based on data table and interaction of vehicles for each generation, where in first generation TF recorded was 833.5877's, F1 was 163.4693's and F2 was 670.1184's. In the second and third generation, the GATSTMS get the optimal solution respectively (712.1425's, 652.8878's) for TF, (157.5795's, 145.7803's) for F1 and (554.563's, 507.1075's) for F2.

Figure 5, shows the relation between TF and optimal CT for intersection C1 and C2 for each generation, where the TF, CT1 and CT2 respectively (833.5877's, 92's and 120's), (712.1425's, 84's and 126's) and (652.8878's, 84's and 126's).

6 Conclusions and Future Work

In this paper, we have addressed the vehicle traffic congestion problem that is one of the most prominent problems. We have presented a dynamic model for most of

the current traffic light control systems using the Genetic Algorithm. In our work, the user defines the structure of the traffic lights, while the GA program optimizes the traffic lights fluency, which can be achieved by finding the minimum delay in the entire system. The cycle time is defined at each intersection to enhance the fluency and avoid long queues in each intersection. The work presents a more efficient GATSTMS that is suitable for wide range of different traffic models while considering a number of dynamic constraints. The paper presents the green splits such as dynamic as to control lanes movement of any road.

In the future, we will use our proposed GATSTMS on more complex road networks, especially when there are more intersections and evaluate the performance of our model in more details.

References

1. Kwasnicka, H. and Stanek M.: Genetic approach to optimize traffic flow by timing plan manipulation. Intelligent Systems Design and Applications, ISDA'06. Sixth International Conference on, IEEE (2006).
2. Purohit, G., Sherry, M. and Saraswat, M.: Time optimization for real time traffic signal control system using Genetic Algorithm. Global Journal of Enterprise Information System, Volume 3, Issue 4, (2011).
3. Russell, S. J. and Norvig, P.: Artificial intelligence: a modern approach. Prentice hall Englewood Cliffs (1995).
4. T.-T.: Deployment of intelligent transport systems on the trans european road network, expert group on ITS for road traffic management (2002).
5. Van Zuylen, H. J. and Viti, F.: Delay at controlled intersections: the old theory revised. Intelligent Transportation Systems Conference, ITSC'06. IEEE, (2006).
6. Wargo, J., Wargo, L. and Alderman, N.: The harmful effects of vehicle exhaust: A case for policy change, environment & human health, (2006).
7. Yousef, K. M., Al-Karaki, J. N. and Shatnawi, A.: Intelligent traffic light flow control system using wireless sensors networks. Journal of Information Science and Engineering Volume 26, No. 3, pp. 753-768, (2010).
8. Zang, L., Hu, P. and Zhu, W.: Study on dynamic coordinated control of traffic signals for oversaturated arterials. Journal of Information and Computational Science, Volume 9, Issue 12, pp. 3625-3632, (2012).
9. Donis-Díaza,C.A., Muroa,A.G., Bello-Péreza, R., Morales, E.V.: A hybrid model of genetic algorithm with local search to discover linguistic data summaries from creep data. Expert Systems with Applications, Volume 41, Issue 4, pp. 2035–2042(2014).

Parameters Analysis of Satellite Support System in Air Navigation

Miroslaw Siergiejczyk, Karolina Krzykowska, and Adam Rosinski

1 Introduction

The constant development of satellite navigation has a definite influence on safety and improve the quality of positioning and navigating vehicles. GPS (Global Positioning System) and GLONASS (Globalnaja Nawigacjonnaja Sputnikowa Sistiema) systems are the first ones, the other start-ups base on their performance. In Europe, the concept of using satellite navigation was launched in 1990.

2 Architecture of satellite systems including supporting systems

EGNOS (European Geostationary Navigation Overlay Service) is part of the GNSS (Global Navigation Satellite System) [1, 3, 7]. It is one of the SBAS systems (Satellite Based Augmentation System), which transmit differential corrections via geostationary satellites. It was created for military ground and air system users in order to improve the quality of positioning GPS and GLONASS, which itself bases action. Efforts on the use of EGNOS for aviation are primarily dictated by the low cost of its use. The EGNOS system differs from GPS / GLONASS with several important elements:

- ensuring the reliability of positioning quality SOL (Safety -of- Life),
- increasing accuracy of determining the position (1 to 2 m),
- the ability to determine the geographical location with the guarantee correctness.

Space Segment of the system consists of three geostationary satellites:

- Inmarsat -3 AOR -E located at 15.5 ° W,
- Inmarsat -3 IOR -W located at 25.0 ° E,
- ESA Artemis located at 21.5 ° E.

EGNOS satellites, unlike the GPS ones do not generate signals and are equipped only with a transponder, by means of which, all users transmit signal generated in the ground segment. The EGNOS system provides users with free access to civilian signal reception. Additional ease of use is that the user must have a receiver compatible with that used for the reception of the GPS signal. It is possible, because EGNOS and GPS signals have the same frequency and modulation [2, 4].

GPS space segment consists of 31 satellites. The satellites are not evenly distributed on each orbit, the position of theirs is chosen so that at any point of the Earth there is visible as many satellites as possible. According to the system specifications it can be expected with a probability of 0.9996 on average to see five satellites. Ground segment of each satellite systems fulfills the functions of control, supervision and correction. Distribution of ground stations is chosen in the way that the areas to monitor overlap which translates into greater reliability supervision. This also allows a comparison of received signal for several different stations. This form of land-based system gives the ability to accurately determine the orbits of all the satellites, as well as to synchronize their system time scale adopted by the improvement of on-board display patterns.

Pseudorange measurement of the GPS satellites is a measurement of the distance between the satellite and the receiver by with the time interval between the time of signal transmission from the satellite and the time when signal is received by the receiver. This interval is called the time shift and it occurs between the coded signals generated from the satellite as a PRN (Pseudo -Random Noise), and the signals of the same form, however, generated by the receiver. It is the PRN code form respectively, replicated, shifted until the maximum correlation with the PRN code obtained from the satellite. The user position is calculated on the basis of knowledge of the position of all currently seen by the

M. Siergiejczyk (✉) • K. Krzykowska • A. Rosinski
Faculty of Transport, Warsaw University of Technology, Warsaw, Poland
e-mail: Msi@wt.pw.edu.pl; kkrzykowska@wt.pw.edu.pl; adro@wt.pw.edu.pl

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_95,

receiver satellites and their pseudoranges. All the information needed to determine the coordination of the satellites receiver are contained in the navigation message.

The main errors now exchanged for the GPS system are:

- Ionospheric Delay - distance error associated with the propagation of radio waves, the inaccuracy of 2.3 - 5 m,
- Delay troposphere - the result of changes in the speed signal when passing through the atmosphere, the inaccuracy of 1.5 m,
- Ephemeris Error - this is the difference between the actual and the set of the orbital position of the satellite data, inaccuracy of the order of 4.2 m,
- Satellite clock error - the difference between the exact GPS time and indicated by the satellite clock, inaccuracy order 3 m,
- Multipath Error - arises when the user reaches the signal reflected from obstacles, uncertainty of the order of 1 - 2 m,
- Receiver Error - Receiver noise may generate an error of 0.2 - 1 m.

3 Requirements in relation to the parameters of satellite systems

Satellite navigation systems have been developed for users so that they could get the desired useful data. Whether these data are used in learning, everyday life or transport, they must meet a number of standards and requirements. In all these areas the system performance is determined by four parameters: credibility, accuracy, availability and continuity. For use in aviation – these requirements are higher, there are more important parameters and their values are different for phases of flight or ground operations [5], [9], [11].

Credibility is a key parameter of the system due to the quality, safety and warning users when crossing the limits of tolerance. A separate definition of credibility was written by the ICAO SARP (Standards and Recommended Practices), according to which, reliability is a measure of the confidence placed in the transmitted information. Credibility includes the ability of the system to ensure important warnings to the user on time. Please note that the converse is also true definition -as the risk of credibility, and therefore the probability of sending a signal out of tolerance. The figure below shows a Stanford diagram for credibility in EGNOS.

Credibility is defined in three categories as an alarm limit, the time to alert and the probability of detection credibility. The probability of not detecting the credibility is the probability that the error exceeds the alarm limit and the user is not notified in time to alert. Stanford diagram (Figure 1) indicates when the satellite service is available ($PE < AL < PL$) or unavailable ($AL < PE < PL$, $AL < PL < PE$) within the framework of credibility. There are also distortion of information which is also included in the diagram, MI (Misleading Information) occurs when $PL < PE < AL$, $AL < PL < PE$ and HMI (hazardously Misleading Information) for $PL < AL. < PE$, where PE is the level of safety (Protection Level), PE is the position error (position error) and AL is the time limit for alarm (Alert limit).

Accuracy is the difference between the actual and measured by the system position. Availability is the ratio of the percentage of time when the broadcast beacons are available for use. The continuity of the system is defined as the ability to operate properly for a given period of use. These requirements are the most important (Table 1) and must be

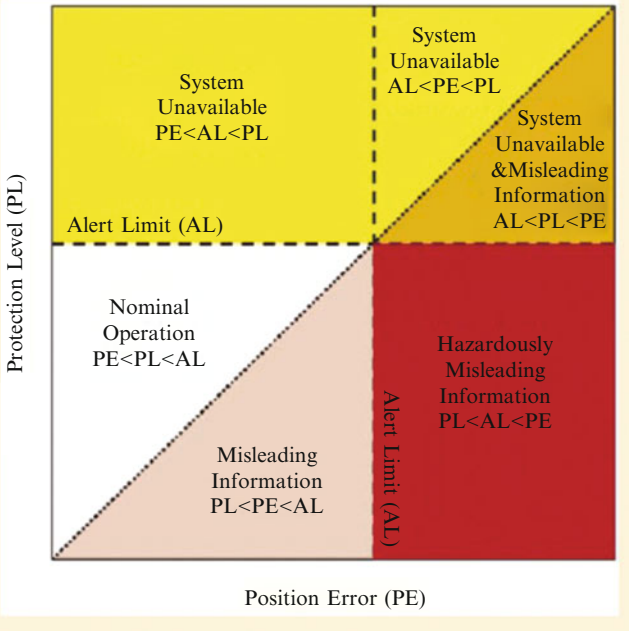


Fig.1 Stanford diagram for credibility in EGNOS

Table 1 Requirements for parameters in EGNOS

Operation	Accuracy horizontal 95 %	Accuracy vertical 95 %	Credibility	Continuity	Availability
APV - I	16.0 m (52 ft)	20 m (66 ft)	$1 - 2 * 10^{-7}$	$1 - 8 * 10^{-6}$ in any 15 s	0.99 to 0.9999
APV - II	16.0 m (52 ft)	8.0 m (26 ft)	$1 - 2 * 10^{-7}$	$1 - 8 * 10^{-6}$ in any 15 s	0.99 to 0.99999
CAT - I	16.0 m (52 ft)	6.0 m to 4.0 m (20 ft to 13 ft)	$1 - 2 * 10^{-7}$	$1 - 8 * 10^{-6}$ in any 15 s	0.99 to 0.99999

to consider development of satellite systems in aviation. It is also clear that even if all these conditions would not cause that navigation would be carried out only with the help of satellite systems. In aviation, from the point of view of safety, all systems are duplicated or others reserved for emergencies failure. In the Table 1 below there is a juxtaposition of requirements for parameters in EGNOS.

EGNOS satellites send signals through the receiver, using the appropriate decoding to determine the user"s position. PEGASUS program, created for research purposes, operates in a similar way, even after the signal received determines the position. It has, however, a fundamental difference, namely when its use is determined by the difference between the position determined by the EGNOS system, and the known position of the receiver signal by specifying its exact location. This makes it possible to obtain the positioning error and a number of other significant size. The program was created in order to investigate the operation of the SBAS (Satellite Based Augmentation System), including EGNOS and GBAS (Ground Based Augmentation System) before deploying them as operative. PEGASUS generates an overview of the most important characteristics of the system EGNOS for two categories of approaches with vertical and horizontal guidance (APV), and one category of precision approach (CAT-I), such as:

- Determining the position error (accuracy) for both levels (HNSE - Horizontal Navigation Service Error) and vertical (VNSE - Vertical Navigation Service Error), these values are calculated by two different assumptions:
  - Measured (Meas.) - is taken straight from the measurement of the positioning error,
  - Scaled (Scal.) - determined by the ratio of AL to the UK, and the data scaled to the worst geometry in order to avoid the volatility of accuracy caused by the geometry of orbiting satellites.
- Availability, checked on an assumed 99 % level,
- Continuity, described in the table as the number of events discontinuity,
- Credibility, described in the table as the amount of distortions of information (MI HMI).

Requirements relating to the results presented above:

- The accuracy of determining the position of:
  - APV- I: The horizontal plane - 16 m, The vertical plane - 20 m
  - APV- II of The horizontal plane - 16 m, The vertical plane - 8 m,
  - CAT - I: The horizontal plane - 16 m, the vertical plane - 4 m.
- Availability - 99 % of the time covered by EGNOS system,
- Continuity - no instances of discontinuity events,
- Authenticity - no instances of falsification of information MI, the HMI.

4 Results of the analysis of the parameters for EGNOS (from PEGASUS) for Warsaw

It should be noted that the satellite of PRN126 on 10.09.11r. Did not send a beacon signal, which may mean damage, or temporary exclusion. This means no data in the respective tables. The following chart shows the errors of determining the position in the horizontal plane and the vertical (HNSE, VNSE) of EGNOS in the dates for the three geostationary satellites (PRN120, PRN124, PRN126). The division also occurs due to an approach (APV -I, - II APV, CAT - I), and further for two estimates, the measured and scaled of that error. Chart 1. contains errors of determining position, the values of which are very satisfactory, because there has been no exceedance of limit. There can be seen the difference between the measured and scaled values, and this is due to the amount and type of factors that affect the measurement result. Along with the restrictions for category landing approach decreases the error determined position.

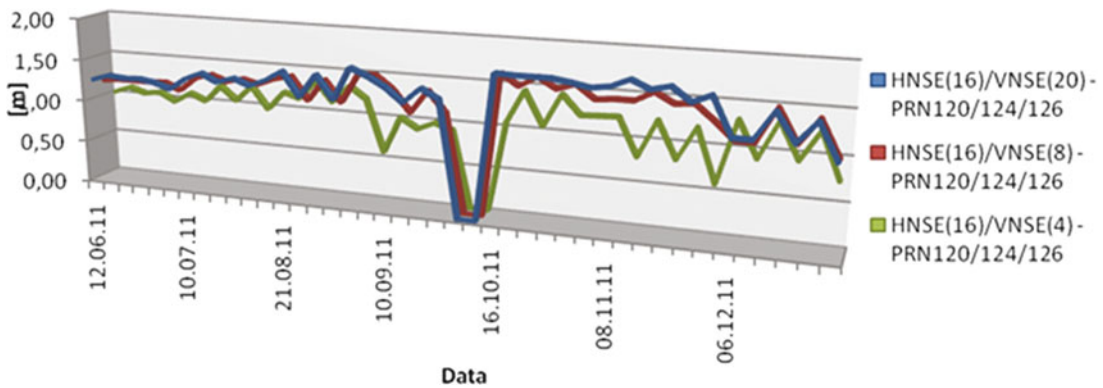


Chart 1 Error determining the position (measured)

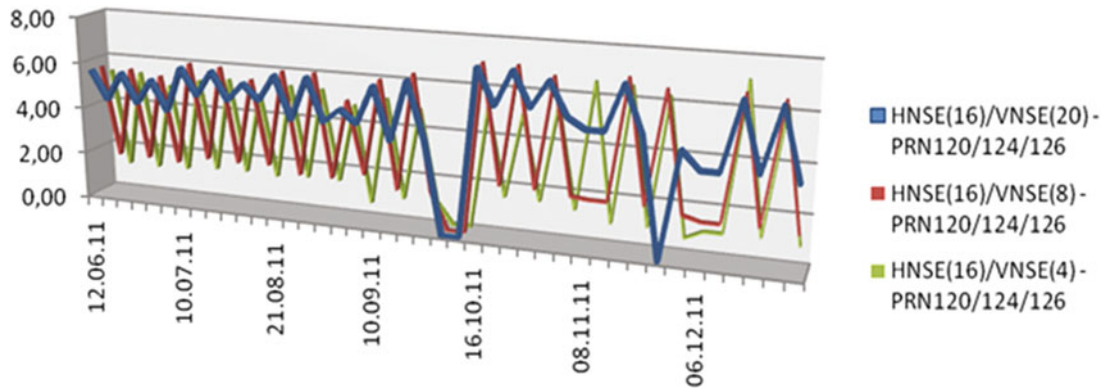


Chart 2 Error determining the position (scaled)

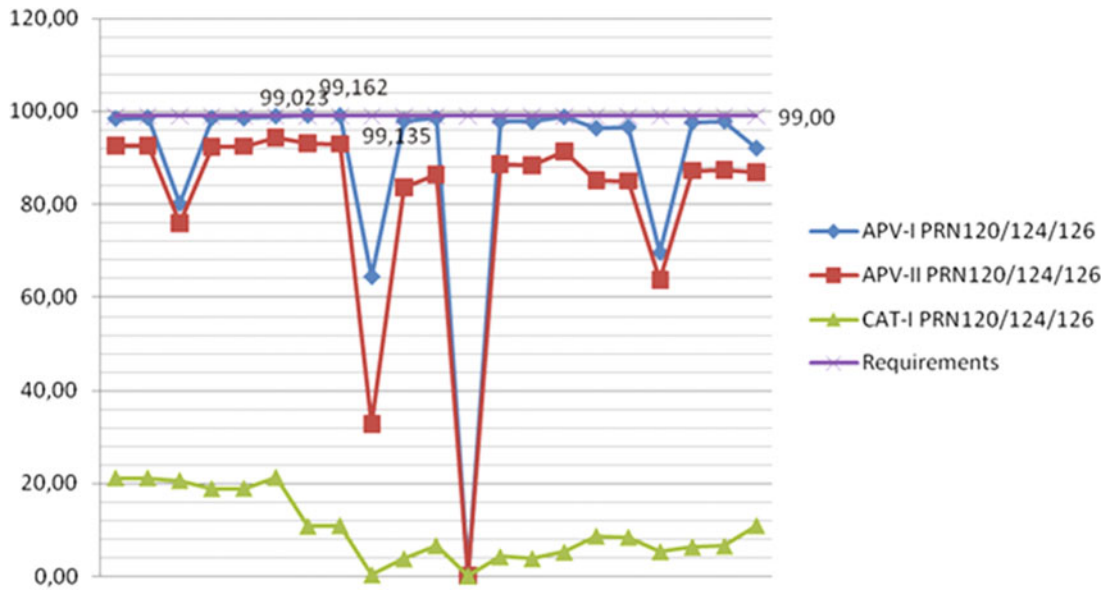


Chart 3 Data on system availability

Chart 3. presents data on system availability for chosen dates for the three categories of approaches (APV-I, APV-II, CAT-I), and three geostationary satellites (PRN120, PRN124, PRN126) EGNOS. The results displayed in the chart provide a picture of insufficient quality of this parameter because the majority did not exceed the assumed 99 %. Indeed it is required that at least two of the three satellites comply with the criteria.

Table 2. contains the results for the continuity of the EGNOS system in the selected days. Breakdown occurs due to the geostationary satellites (PRN120, PRN124, PRN126) and categories of approaches (APV-I, APV-II, CAT-I).

The continuity of the system is defined as the amount of discontinuity events and they are the result comprised in Table 2. The occurrence of these events for any category approaches, regardless of the number of satellites, cause failing the test. It is seen that for any of the analyzed days parameter does not met expectations. The amount of discontinuity is very large, which makes the signal quality as continuity very low.

Credibility is analyzed based on the occurrence of distortions information (MI HMI). No table indicates the absence of these errors leads to the conclusion that the navigation service EGNOS in the context of the accuracy of this parameter is ensured.

5 Analysis of satellite support system in air navigation

It is possible to create an advanced integrated system which consists of different sources of surveillance, for example GPS and EGNOS.

Damage of one of the systems moves it from the state of full ability $R_0(t)$ to the state of the impendency over safety unreliability $Q_{ZB}(t)$. Figure 2 shows the relationships in the integrated surveillance system in terms of reliability [6, 8, 10].

Denotations in figures:

$R_O(t)$ – the function of probability of system staying in state of full operational capability,

$Q_{ZB}(t)$ – the function of probability of system staying in state of the impendency over safety,

$Q_B(t)$ – the function of probability of system staying in state of unreliability of safety,

λ_{ZB1} – transition rate from the state of full operational capability into the state of the impendency over safety,

μ_{PZ1} – transition rate from the state of the impendency over safety into the state of full operational capability,

λ_{ZB2} – transition rate from the state of the impendency over safety into the state of the unreliability of safety,

μ_{PZ} – transition rate from the state of unreliability of safety into the state of full ability.

The system illustrated in fig. 2 may be described by the following Chapman–Kolmogorov equations:

$$\begin{aligned} R_0'(t) &= -\lambda_{ZB1} \cdot R_0(t) + \mu_{PZ1} \cdot Q_{ZB}(t) + \mu_{PZ} \cdot Q_B(t) \\ Q_{ZB}'(t) &= \lambda_{ZB1} \cdot R_0(t) - \mu_{PZ1} \cdot Q_{ZB}(t) - \lambda_{ZB2} \cdot Q_{ZB}(t) \\ Q_B'(t) &= \lambda_{ZB2} \cdot Q_{ZB}(t) - \mu_{PZ} \cdot Q_B(t) \end{aligned} \quad (1)$$

Given the initial conditions:

$$\begin{aligned} R_0(0) &= 1 \\ Q_{ZB}(0) &= Q_B(0) = 0 \end{aligned} \quad (2)$$

The following system of linear equations we get after Laplace transform:

$$\begin{aligned} s \cdot R_0^*(s) - 1 &= -\lambda_{ZB1} \cdot R_0^*(s) + \mu_{PZ1} \cdot Q_{ZB}^*(s) + \mu_{PZ} \cdot Q_B^*(s) \\ s \cdot Q_{ZB}^*(s) &= \lambda_{ZB1} \cdot R_0^*(s) - \mu_{PZ1} \cdot Q_{ZB}^*(s) - \lambda_{ZB2} \cdot Q_{ZB}^*(s) \\ s \cdot Q_B^*(s) &= \lambda_{ZB2} \cdot Q_{ZB}^*(s) - \mu_{PZ} \cdot Q_B^*(s) \end{aligned} \quad (3)$$

Probabilities of system staying in a distinguished functional states in symbolic (Laplace) terms have the following form:

Table 2 Results for the continuity of the EGNOS system in the selected days

		Continuity		
Warsaw		APV-I	APV-II	CAT-I
Requirements		No Discontinuities Event		
12/06/11	PRN120	**4**	**261**	**584**
	PRN124	**4**	**191**	**616**
	PRN126	**5**	**73**	**858**
10/07/11	PRN120	**5**	**251**	**876**
	PRN124	**8**	**161**	**763**
	PRN126	**20**	**192**	**898**
21/08/11	PRN120	**3**	**291**	**923**
	PRN124	**3**	**244**	**944**
	PRN126	**46**	**540**	**131**
10/09/11	PRN120	**38**	**284**	**712**
	PRN124	**17**	**224**	**558**
	PRN126	**-**	**-**	**-**
16/10/11	PRN120	**5**	**185**	**416**
	PRN124	**5**	**157**	**351**
	PRN126	**15**	**131**	**603**
08/11/11	PRN120	**30**	**341**	**509**
	PRN124	**33**	**299**	**443**
	PRN126	**6**	**256**	**429**
06/12/11	PRN120	**44**	**251**	**804**
	PRN124	**18**	**248**	**873**
	PRN126	**48**	**181**	**998**

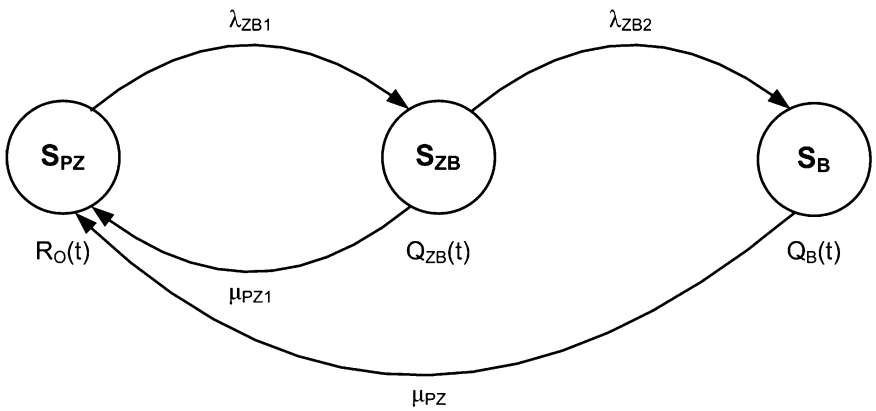


Fig. 2 Relations in the system

$$
\begin{aligned}
R_0^*(s) &= \frac{s^2 + s \cdot \mu_{PZ} + s \cdot \mu_{PZ1} + s \cdot \lambda_{ZB2} + \mu_{PZ} \cdot \mu_{PZ1} + \mu_{PZ} \cdot \lambda_{ZB2}}{s^2 \cdot \mu_{PZ} + s^2 \cdot \lambda_{ZB1} + s^2 \cdot \mu_{PZ1} + s^2 \cdot \lambda_{ZB2} + s^3 + s \cdot \mu_{PZ} \cdot \lambda_{ZB1} + s \cdot \mu_{PZ} \cdot \mu_{PZ1} + \\
&\quad + s \cdot \mu_{PZ} \cdot \lambda_{ZB2} + s \cdot \lambda_{ZB1} \cdot \lambda_{ZB2}} \\
Q_{ZB1}^*(s) &= \frac{s \cdot \lambda_{ZB1} + \mu_{PZ} \cdot \lambda_{ZB1}}{s^2 \cdot \mu_{PZ} + s^2 \cdot \lambda_{ZB1} + s^2 \cdot \mu_{PZ1} + s^2 \cdot \lambda_{ZB2} + s^3 + s \cdot \mu_{PZ} \cdot \lambda_{ZB1} + s \cdot \mu_{PZ} \cdot \mu_{PZ1} + \\
&\quad + s \cdot \mu_{PZ} \cdot \lambda_{ZB2} + s \cdot \lambda_{ZB1} \cdot \lambda_{ZB2}} \\
Q_B^*(s) &= \frac{\lambda_{ZB1} \cdot \lambda_{ZB2}}{s^2 \cdot \mu_{PZ} + s^2 \cdot \lambda_{ZB1} + s^2 \cdot \mu_{PZ1} + s^2 \cdot \lambda_{ZB2} + s^3 + s \cdot \mu_{PZ} \cdot \lambda_{ZB1} + s \cdot \mu_{PZ} \cdot \mu_{PZ1} + \\
&\quad + s \cdot \mu_{PZ} \cdot \lambda_{ZB2} + s \cdot \lambda_{ZB1} \cdot \lambda_{ZB2}}
\end{aligned} \tag{4}
$$

Solution to the above set of equations in the time domain is the next step in the analysis and is not discussed here.

6 Summary

European Geostationary Navigation Overlay Service (EGNOS) is a satellite navigation system that was developed to assist in the positioning GPS and GLONASS. With precision fed information about the location it can be applied to maritime transport, land and air. In the article analyzed and interpretated were all parameters that describe the operation of EGNOS for two categories of approaches with vertical and horizontal guidance, and for one category of precision approach. The study consisted of several days reception in Warsaw. The results of the analysis showed a lack of fulfillment by the beacon signal of the EGNOS system requirements of two important parameters. The size of the biggest mistakes is burdened with the continuity of the system. It describes the ability of the system to work properly during use. The negative outcome is the result of occurrence of discontinuities. For category approach APV -I disorders are the smallest, and their size increases with the more accurate category approaches. For the category of APV -II and CAT -I results illustrate a very bad work of EGNOS. The second parameter is the availability meeting the requirements specifying the percentage of the time signal suitable for navigation. For the category of approaches APV -II and CAT -I results are very poor. Promising results are, however, for other parameters, including two extremely important. The accuracy of determining the position, regardless of the plane reaches a level of less than 2 m. Moreover, for the first category of precision approach accuracy error oscillates around 1 m. The second parameter is called credibility. It is the size that specifies the error indication system. The EGNOS system does not generate distortions of information. Disclosed the status of the EGNOS system does not allow the use of services within the airport traffic area in Poland. This may be due to the limited scope of the system in the eastern part of Europe.

References

1. EGNOS Fact Sheet 3: Integrity explained. ESA (2005)
2. ESA (2012) European Geostationary Navigation Overlay Service, www.cbk.waw.pl Accessed 27 December 2012
3. European Organisation for the Safety of Air Navigation EUROCONTROL (2007) NAV-GNSS Global Navigation Satellite System Training Provided by the IANS ATM Unit. Luxembourg, Kirchberg
4. Flament, D.; Ventura-Traveset, J.: EGNOS The European Navigation Overlay Service – A cornerstone of Galileo. Noordwijk (2006)
5. Konaktchiev, I., Butzmuehlen, C.: Agenda Item 4: First Glance Algorithm Description. AIRSPACE AND NAVIGATION TEAM (EUROCONTROL) (2005)
6. Laskowski, D., Łubkowki, P.: The end-to-end rate adaptation application for real-time video monitoring. In Advances in Intelligent Systems and Computing, pp. 295-305. Springer International Publishing AG (2013)
7. Narkiewicz, J.: Podstawy układów nawigacyjnych. Warszawa (1999)
8. Rosinski, A., Dabrowski, T.: Modelling reliability of uninterruptible power supply units. Eksploatacja i Niezawodnosc – Maintenance and Reliability, Vol.15, No. 4, 409-413 (2013)
9. Siergiejczyk, M., Krzykowska, K.: The analysis of implementation needs for automatic dependent surveillance in air traffic in Poland. TransNav - The International Journal on Marine Navigation and Safety of Sea Transportation, 241 – 245, London UK (2013)
10. Siergiejczyk M., Rosiński A., Krzykowska K.: Reliability assessment of supporting satellite system EGNOS. New results in dependability and computer systems. Springer (2013)
11. Skorupski, J., Malarski, M., Stelmach, A.: Air traffic safety investigation problem. Proceedings and Monograph in Engineering, Water and Earth Sciences, ESREL, Portugal (2006)

Selected Issues of the Reliability Analysis of GSM-R in Poland

Mirosław Siergiejczyk

1 Introduction

Development plan for European Rail Traffic Management System in Poland adopted by the Council of Ministers envisages to implement the ERTMS standard on 15000 km of railways. The need to implement the solution which combines two systems namely ETCS (*European Trail Control System*) and GSM-R (*Global System for Mobile Communications - Railway*) is motivated by both technical (technological progress in telecommunications and rail traffic management) and legislative (necessity to enhance interoperability on the basis of EU and Polish legislation). Interoperability also refers to telecommunications - the GSM-R system has been acknowledged as a rail radiocommunication system on a European level.

The GSM-R system was developed to facilitate communication between the regulation centre, train driver and railway operational staff across the entire line. [6] Thanks to built-in priority control mechanism, each call has a predefined priority so that the most important ones are served first. Thanks to the aforementioned, traffic density on given line can be planned at an optimum level, thus makes better use of that line and improves safety. GSM-R standard was developed by workgroup 7B9 as part of the EIRENE project implemented under the auspices of International Union of Railways UIC [11]. That system displaced various railway communication systems across the European Union. In Poland the voice radio system operates the frequency VHF 150 MHz. Once implemented, the quality of radio telecommunications improves and barriers for data transmission and voice radio systems caused by the sheer multitude of different solutions used across European Union are lifted.

M. Siergiejczyk (✉)
Railway Institute, Warsaw, Poland

Faculty of Transport, Warsaw University of Technology, Warsaw, Poland
e-mail: msiergiejczyk@ikolej.pl; msi@wt.pw.edu.pl

The GSM-R system is also a transmission medium for ETCS system at levels 2 and 3 designed to, among other things, authorise a clean line via the Radio Block Centre for trains located within an area of given RBC [12].

2 Overview of the GSM-R System

GSM-R is a Global System for Mobile Communications used for purposes of rail transport. It enables digital voice communication and digital data transmission. Base stations dedicated for GSM-R are usually laid out in different ways depending on required level of safety. The layout and connection between base stations should be determined by the class and intended use of railway line, its throughput and required level of safety. There are the following three typical types of cells. They were presented in figure 1. The first are cells which cover only the area of the railway line. They are long and narrow. The second are cells covering areas of railway stations and partly railway lines. They are usually circular or elliptical. The third type is big cells (3) which cover other areas like rail yard, complexes of railway buildings etc. Each type of cell is supported by all types of mobile phones. The size and shape of cells may be adjusted by controlling power levels and using omnidirectional antenna, both wide and linear angle. The GSM-R system for all intents and purposes is for non-public use only, hence it was not envisaged to provide coverage over areas other than railway areas.

GSM-R is based on GSM phase 2 offering all basic and additional services complimented by GSM phase 2+ (voice broadcast service, voice group call service, GPRS, Multi-Level Precedence and Pre-emption Service - call priority control). Hence the following services were introduced [6], [12]:

- *Voice Broadcast Service* – only the initiator of the call can speak. The others who join the call can only be listeners. This kind of call is mainly used to broadcast recorded messages or to make announcements.

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_96,

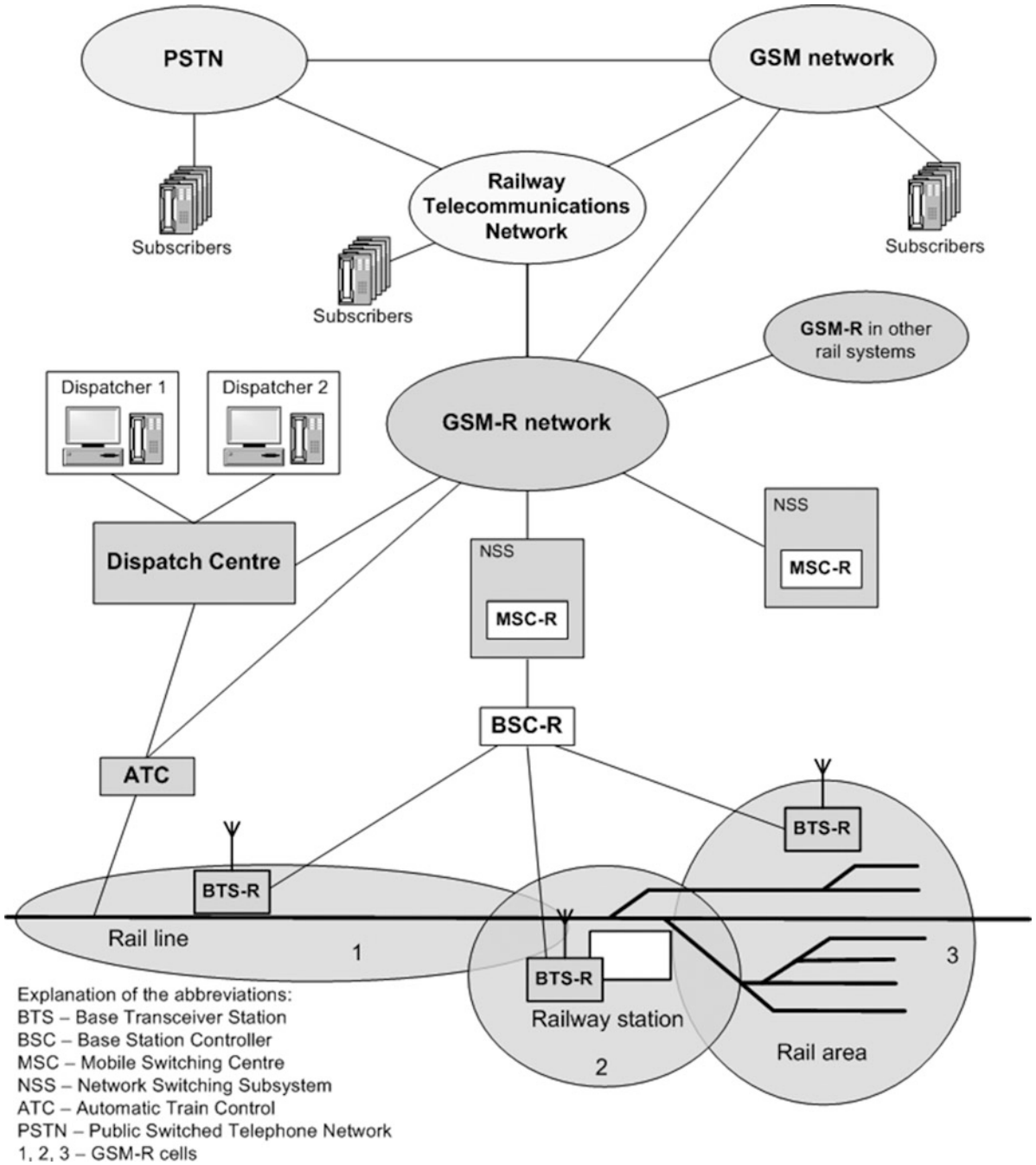


Fig. 1 Organisation of GSM-R network in rail transport (Source: Logistyka 6/2012)

- *Voice Group Call Service* – enables simultaneous and mutual communication between predefined user groups (SIM card) and each participant may both activate and terminate its connection. The dispatcher or meeting organiser moderates participants list, selects talkers and oversees the entire group call.
- *enhanced Multi-Level Precedence and Pre-emption*) – service which defines the user's priority and helps in emergencies. A caller with higher priority would terminate calls of lower priority users. Inbound calls to lower priority users are automatically answered. The call establishing time for priority call should be short (below 2 s inclusive of dropping current call) [5].
- *Functional Addressing* - alias system to call someone registered on the GSM-R network, only by knowing the temporary function user. This way a connection is made by dialling desired function by its identification number as opposed to having to dial a particular physical

terminal. Thanks to this feature, the train driver may be reached not only on his personal number but also by typing train number or locomotive number.

- *Location Dependent Addressing* – provides the automatic routing of Mobile Originated Calls to predefined destinations relative to the geographic area where the subscriber is roaming. This type of addressing enables the train driver to dial the dispatcher by pressing a single button. The location is determined based on identifier of the cell where the train currently is located. Due to various sizes of the cells, accuracy of information about location may be improved by using other sources as well i.e. on-board train positioning system, balises or data transmitted from systems using fixed infrastructure.
- *General Packet Radio Service* – is intended mainly for rail-vehicle transmission (ETCS).
- *enhanced Railway Emergency Call* - emergency call informs the train driver, traffic controllers and other personnel about hazards and dangers necessitating possible stopping of the train within given area or other actions to be taken. The type of call being initiated is established automatically on the basis of operating mode of initiating terminal. Railway emergency call needs to reach all train drivers and traffic controllers within emergency area.
- *Shunting mode* – shunting team can communicate with each other.
- *Direct mode* – supported by mobile terminals and is set up when GSM-R network is not used for communication. This feature is intended for situations like network outage or no GSM-R reception.
- *Data transmission*. Data transmission in GSM-R supports four main groups of services: SMS, core data reliant applications, automated facsimile and applications assisting train driving. Text messages are sent in two ways: point-to-point between two users or point-to-multipoint to many users at the same time. Data transmission service is required to remotely control on-board units and track-side equipment, automatically control train traffic, control railway traffic safety and enable passenger applications. Some of the applications dedicated for train passengers are railway timetables, weather information or internet access. GPRS and EDGE packet data services, known from public cellular communication systems, are being introduced into GSM-R.

In Poland the GSM-R system is currently being implemented on the E30 railway line. Due to substantial role GSM-R plays in railway traffic management and assuring railway traffic safety it is paramount that the system is highly reliable. The key parameter is availability i.e. the probability an object is ready to complete its functions at given time *t*. [3], [5] Because ERTMS solutions are to a large extent automated, rely on continuous data feeds over, among others GSM-R, the system needs to be highly available. Outages and unplanned downtimes could lead to miscommunication between equipment and personnel ultimately leading to a rail disaster.

3 GSM-R network availability

ERTMS system was designed around the overarching principle of assuring highest possible rail traffic safety. ETCS and GSM-R are autonomous only when ETCS implemented on given railway is level other than two and three - then despite interoperability of both systems, one does not require the other to operate correctly - failure of GSM-R would not render ETCS inefficient as well. The relationship between ETCS and GSM-R systems in terms of reliability occurs when ETCS enabled (level two and three) rail line is concerned. In case where the GSM-R system constitutes a transmission medium used by ETCS failure of the former renders inefficient the latter. It is for that reason why EIRENE System Requirements Specification [11] determines availability of GSM-R relative to the level of ETCS with which it inter-operates. Requirements towards reliability of GSM-R are the following:

- Failure-free system operating time should be greater than 13500 h $>$ ($>$1.5 year);
- Service life greater than 25 years;
- In order to assure uninterrupted operation in cause of failure, the remote control system components should be designed to provide spinning reserve;
- Availability as high as 0.99995 for purposes of ETCS levels 2/3 and 0.9991 for purposes of voice and information transmission.

It is fair to say that availability required for purposes of ECTS 2/3 is relatively high. As far as availability metrics for IT systems are concerned, it would score between "fault tolerant" (availability $=$ 0.9999) and "continuous availability" (availability $=$ 0.99999). Availability of 0.9995 % means that the network is not able to provide its services for as little as 26 minutes per annum. Reliability of GSM-R network may be considered in many ways.

Layers of GSM-R architecture often consist of several dozen complicated devices. In case of a system that extensive, it is absolutely vital from the viewpoint of reliability to pinpoint “weakest links” i.e. system components whose reliability parameters are worse and undertake correcting measures to improve them. According to the vulnerability theory which states that in order to maximise a system’s reliability its most vulnerable elements need to be targeted to improve i.e. element’s exhibiting worse reliability parameters. [5], [8]. The most simple yet most popular solution is to introduce devices providing the spinning

reserve for elements displaying worse reliability parameters relative to other components. Usually one main element is accompanied by one redundant element both of which are operating (redundancy technique 1 + 1). In case of substantial series structures showing high reliability, manufacturers decide sometimes to include a single device providing spinning reserve (redundancy technique n + 1). Much rarer are series structures with multiple redundant devices (redundancy technique n + p) [7], [8].

When considering reliability of equipment constituting architecture of the GSM-R system, one shall remember about supporting infrastructure e.g. optical fibre networks assuring communication between elements (often set apart hundreds of kilometres) or power supply infrastructure [4], [9], [10]. Distances between individual elements and exposure to weather conditions such as rain, freeze and high temperatures mean that designers need to use only top of the range equipment, often best in class.

As far as BSS (*Base Station Subsystem*), components are concerned it is difficult to establish geographical location of balises. In order to assure radio coverage required by EIRENE SRS, designers often place them at a distance from roads and transportation routes. This obviously hinders attending to failed component thus increases time period over which device is not operational. In pursuance of optimum balises' location more optical fibre cables are required. Their reliability in turn is closely linked to the length of the cable. Effectively, the best solution is to implement BTS in series and parallel layout.

In terms of NSS (*Network Switching Subsystem*) reliability is usually less of an issue. Its elements are located inside a building therefore technical staff can react very fast. Failures are diagnosed and repaired quickly, often thanks to replacement parts supplied by the manufacturer, thus the downtimes are kept at minimum. Therefore NSS devices are often single pieces of equipment or implemented in line with 1 + 1 redundancy technique.

Remaining layers, although important from viewpoint of reliable system operation, are not key for the system to operate, hence OMC (*Operations and Maintenance Centre*) layer equipment - control and monitoring applications - have no redundancy. Should they fail, it would not render the entire GSM-R system inefficient, hence their contribution to system's reliability was ignored in computations.

4 Estimating reliability for pilot section of E30 railway line

Due to unavailability of manufacturer's data on devices used in the system as well as fact that currently operating solutions are at pilot testing stage, the most suitable from viewpoint of determining availability of Polish GSM-R system is the analytical method. One of such is MTBF (*Mean Time Between Failures*). Analytical methods - they involve calculating reliability parameters based on a mathematical model. The set of produced parameters is therefore derived from the adopted model and input dataset. The key issue when it comes to using those methods are simplifying assumptions whose impact is often unrecognised.

$$MTBF = \frac{1}{\lambda} \tag{1}$$

Note, however, forecasting failure rate and MTBF may be highly erroneous should device in question house a structure which has not been disclosed publicly by the manufacturer. In that case, internal components of device may be determined solely based on knowledge about construction of similar devices. Where the solution is unique, certain assumptions with regards to the internal structure may be drawn from function and features revealed by the manufacturer.

When MTBF of individual devices is known, their availability is determined. Availability of complex systems (telecommunication connections) may be analysed once its constituting components' availability is determined. For purposes of analysis an assumption needs to be taken that each complex system is a more or less complicated combination of individual units. When considering complex telecommunication connections, one needs to start with computing the equivalent availability [1], [2]. Bottom up, beginning with the most simple units, availability of complex systems is determined. This method resembles calculating equivalent resistance of electronic system. An assumption needs to be taken that according to probability mathematics all failures are independent faults. Block diagrams of reliability, which in a transparent manner visualise analysed problem, are helpful to determine availability of complex telecommunications systems by breaking down reliability structure of given network. Should single failure cause downtime of the entire system then that system has a series reliability structure and its equivalent availability is given by [1]:

$$A_{ser} = \prod_{i=1}^{n} A_i \tag{2}$$

where:

$A_i, \ldots, A_i, \ldots, A_n$ – availability of individual n elements in series layout;

As per formula above, availability of each element is always greater or equal to availability of the entire structure. When a system has a parallel structure (failure of individual element would not render the entire system ineffective) it equivalent availability is:

$$A_{par} = 1 - \prod_{i=1}^{n} (1 - A_i) \quad (3)$$

where:

A_i - availability of element number i

Note that formula would apply only to parallel structures whereby only one operating element is sufficient to support the system. The availability of any constituting element is lower or equal to equivalent availability. Where a system has reliable structure "*k-out-of-n*", one may use the reliability function for "*k-out-of-n*" structure" [5]:

$$A_{kon} = \sum_{i=k}^{n} \frac{n!}{i!(n-i)!} (1 - A)^{n-i} \quad (4)$$

where:

i – minimum no of elements that need to be serviceable in order for the system to be operational,

n – total number of constituting elements.

Note that formula (4) is valid only for a system which has a uniform structure.

On the analyzed segment of the E30 railway line it is possible to calculate the availability of the GSM-R system accepting, according to made earlier assumptions, that main blocks of this system (BSS layer, NSS layer and cable connections) are creating the series reliability structure (Fig.2).

About the reliability of devices of BSS layer an availability of the net of BTS station including their topology, BSC(*Base Station Controller*) availability and TCU (*Transcoding Unit*) availability. It is possible for example to consider the single base station. Her reliability structure they presented in the Fig. 3.

Fig. 2 Reliability structure of the analysed segment of the E30 railway line

The method of calculating availability of complex telecommunications systems presented hereinbefore was used to estimate availability of GSM-R system on E30 rail line. The result was 0.99999. Estimated availability complies with EIRENE SRS specification concerning availability of a GSM-R system on ETCS-enabled line (level 2 and 3) stipulating required availability of ≥ 0.99995 [7], [8].

Achieved availability is very high for such a large scale system - both in terms of its architecture and area it covers given the largest distance between elements most apart exceeds 500 kilometres. Only best, and also very expensive, hardware and systemic solutions as well as operational reserve for not only individual modules or devices, but also entire GSM-R architecture (including e.g. switching centres) meant high availability of a system that complex could be assured.

Analysing individual components, a few conclusions were drawn. First and foremost, implementation of redundant elements or devices to provide spinning reserve considerably increases reliability even of elements with low "starting" availability. The issue here, however, is the cost of manufacturing the redundant element, its physical implementation inside a device or in remote location as well as costs of powering and supervising it. An interesting alternative in case of modules or other elements easy to transport and fit is for manufacturers to provide back-up elements and modular structure of devices. Upon failure, a structure as that and back-up elements supplied by the manufacturer would cut Mean Time to Repair (MTTR) down to several minutes thus significantly improving the availability of the entire device without having to implement its modules in parallel structure. Another advantage is the fact that back-up

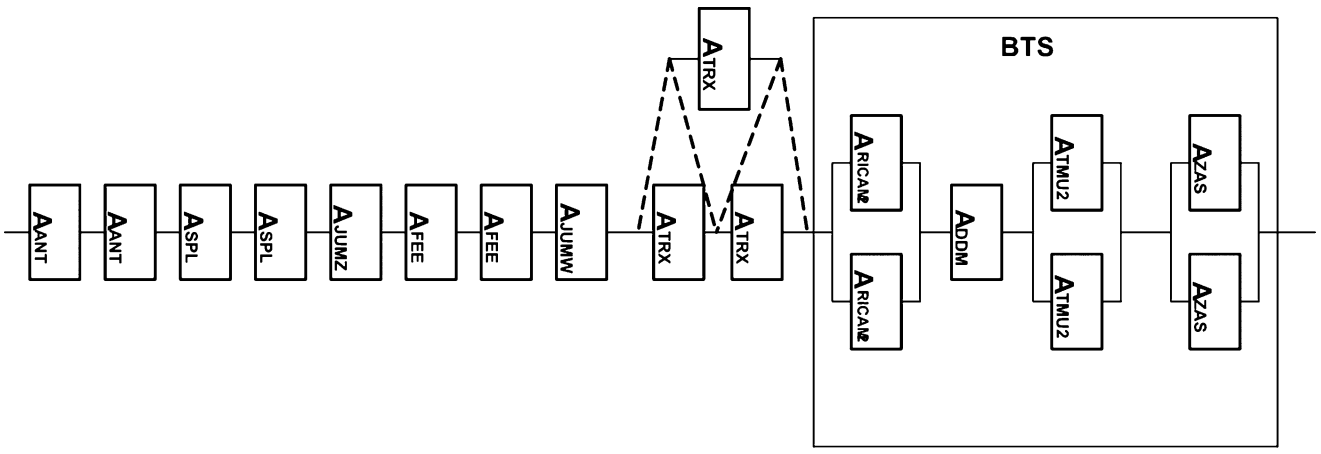


Fig. 3 Reliability structure of the single BTS

element does not need power. A disadvantage on the other hand is that failure of one element would cause a downtime which could potentially lead to disruption in device operation, especially when it comes to failure of Mobile Switching Centre (MSC) or other element that would transpire into outage of the entire system. Hence the solution that guarantees highest possible availability (sometimes of 0.99999999999) is to lay out devices and modules in parallel whilst supplying back-up elements which could be quickly put in place of a failed element. Unfortunately that is the most expensive solution used in case of most prone to failure or vital components of the infrastructure. In case of devices whose construction does not allow easy fitting and disassembly of components, the only solution is to implement those devices in parallel or reliable structure.

In design process, the most important task in terms of system reliability is identifying the "weakest link" in the system and assuring an adequate redundancy in order to increase reliability of the entire system. As far as GSM-R system is concerned, its weakest link is the Base Transceiver Station (BTS). Base stations located in rural areas where access via road infrastructure is challenging would be pending repair longer than normally, thus their availability is not high enough for them to be implemented as independent components. Instead they are laid out in series-parallel assuring best availability. Estimated - using method presented hereinbefore - results are comparable to system availability declared by the manufacturer (0.99999) on test section of E30 line.

5 Summary

Availability analysis is an important issue in a time when GSM-R system is being implemented in Poland. In near future, the GSM-R system will have been implemented on E20 and E65 railway lines and in longer term - on all railway lines in Poland. The most challenging for Poland is the migration from analogue system utilising RadioStop to digital GSM-R. The transition period, when both systems will be used simultaneously will also pose a challenge in terms of assuring required availability of both systems on railway lines using them. The analytical method used in this paper is seemingly a very effective tool to project and estimate availability of radiocommunication system on a railway line. It could also be used to estimate availability of analogue networks and hybrid system used during transition period. It shows in a simple way which system elements are its weakest links in terms of reliability and thus facilitates better design of network using redundancy techniques.

References

1. Chołda P., Jajszczyk A. Assessment of availability in telecommunication networks'. Telecommunication Review and Telecommunication News. No. 2-3/2003. Publication by Sigma NOT Warsaw (2003) (in Polish)
2. Florek J. and others 'Analysis of the telecommunication services quality monitoring systems'. Works of the Institute of Communication, Warsaw, (2000) (in Polish)
3. Kołowrocki K., Soszyńska-Budny J. Reliability and safety of complex technical systems and processes. Springer-Verlag, London (2011)
4. Kwestarz W., Krygier J. Reliability requirements for telecommunication lines. Proceedings of the National Telecommunication Symposium, Warsaw, Poland (1999) (in Polish)
5. Migdalski J. and others: Guide of the reliability. Mathematical bases. WEMA, Warsaw (1982) (in Polish)
6. Siergiejczyk M., Gago S.: Safety issues of GSMR system in the context of supporting of rail transport. Logistyka 6/2012. Poznan (2012) (in Polish)
7. Siergiejczyk M., Gago S.: Problems of reliability and safety of data transmission in the GSM-R system. Proceedings of International Conference RELIABILITY IN RAIL TRANSPORT AND THE POSSIBILITY OF INCREASING. Railway Institute Warsaw (2013) (in Polish)
8. Szmigiel P., Siergiejczyk M.: Selected issues relating to the availability of GSM-R network in Poland. Proceedings of International Conference RELIABILITY IN RAIL TRANSPORT AND THE POSSIBILITY OF INCREASING. Railway Institute Warsaw (2013) (in Polish)
9. Siergiejczyk M., Rosiński A.: Reliability analysis of electronic protection systems using optical links. The Monograph "Dependable Computer Systems". Vol. 97. Springer-Verlag, Berlin Heidelberg (2011)
10. Siergiejczyk M., Rosiński A.: Reliability analysis of power supply systems for devices used in transport telematic systems. The monograph "Modern Transport Telematics", Vol. 239. Springer-Verlag, Berlin Heidelberg (2011)
11. UIC Project EIRENE, System Requirements Specification, GSM-R Operators Group, SRS v 15, 2006
12. Winter P.: International Union of Railways, compendium on ERTMS, Eurail Press, Hamburg, (2009)

Speed-Volume Relationship Model for Speed Estimation on Urban Roads in Intelligent Transportation Systems

Zilu Liang and Yasushi Wakahara

1 Introduction

Traditionally, transportation research has been extensively focusing on highway studies with the relationship among three fundamental variables, i.e., traffic flow, density, and speed, as the theory foundation. Although the exact pairwise relationship between any two of the three variables still remains unclear til today, several models were proposed to capture the relationship between speed and density under equilibrium traffic conditions on highways, among which Greenshield's Model (GM) [5] is the earliest and a most famous one. These models help estimate average speed on highways and have wide applications in highway Intelligent Transportation Systems (ITS).

The recent development of ITS in urban areas have brought out the necessity for effective speed estimation models tailored for urban roads. Traffic density is usually not available in urban traffic network; instead, traffic volume (the number of vehicles on an urban road) is usually available and is often taken as important variale in many ITS applications, such as dynamic traffic assignment [1], traffic simulation [2], and real-time route guidance [3]. Consequently, it is a more feasible way to estimate speed on urban roads from traffic volume instead of density. Therefore, there is need to reclarify fundamental variables and develope effective speed-volume relationship models in urban traffic networks.

The current research trend on this topic is converting highway speed-density models to speed-volume models, and then apply them to urban roads either directly [16] or under trivial modification [2]. However, the effectiveness of this methodology is highly doutble, as urban roads have different charactersitics from highways. Urban roads are generallly short, and thus the traffic conditions on one road could have strong impact on other roads. Under congested traffic conditions, the remaining traffic queue on one road could extend to its upstream roads and thus spillover of congestion [4] could easily happen. Obviously highway models cannot accommodate the above characteristic of urban traffic networks. In this paper, we seek to tackle the problem and propose an effective speed-volume relationship model for the purpose of speed estimation in urban traffic networks. The modeling approach that we adopt is to consider the impedance effect of exit intersections on travel speed on urban roads. The impedance of an exit intersection is caused by multiple factors, and in this paper we mainly consider the impact of the traffic conditions on downstream roads of the concerned one.

The rest of the paper is organized as follows. In the next section we discuss related works. In Section 3, we formulate the problem that has been addressed. In Section 4, we describe the proposed speed estimation model of urban roads. The evaluation of the proposed model using simulated traffic data is presented in Section 5. In the last section we draw conclusions.

2 Related Works

The pioneering work on highway speed-density relationship modeling was done by Greenshield [5] who developed a linear function to summarize such relationship. Due to its simplicity and elegance, Greenshield's Model (GM) has been widely used in transportation research community. However, empirical observations reveal that this model lacks accuracy. In order to tackle the problem, several other models have been proposed to approximate empirical observations closer than GM [17, 18]. A comparison among typical highway speed-density relationship models can be found in [19].

When it comes to urban roads, the study on speed-volume relationship for an individual road is limited. In [16] the

Z. Liang (✉) • Y. Wakahara
Graduate School of Engineering, The University of Tokyo,
Yayoi 2-11-16, Bunkyo-Ku, Tokyo, Japan 113-8658
e-mail: z.liang@cnl.t.u-tokyo.ac.jp; wakahara@nc.u-tokyo.ac.jp

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_97,

authors make a straightforward translation of GM to speed-volume relationship which is shown in (1), and use it to estimate average speed on urban roads.

$$\hat{v} = V\left(1 - \frac{X}{C}\right) \quad (1)$$

where $\hat{v}$ and V are the estimated average speed and the freeflow speed respectively. X is the traffic volume on the concerned road, and C is the capacity of the concerned road. It is worth noting that the capacity used in (1) is in fact the queuing capacity [20] of a road defined as the number of vehicles that can be stored on the road in a queue. When this storage capacity is exceeded the queue will spill back onto the upstream of this road and often block intersections.

This model describes the speed-volume relationship for an individual urban road, and assums that speed on one road is only decided by the vehicles on this road itself. However, urban roads are usually short and thus their exit intersections play a more important role in deciding how fast the vehicles can run on the roads. Under congested traffic conditions, the exit intersection of a road could demonstrate high impedance to the vehicles that attempt to exit from this road. Such impedance effect is caused by various factors, including traffic conditions on other conflicting links, drivers' route choice, etc. The speed models proposed in [21] attempt to capacture the impact of traffic congestion propagating from downstream on the speed of the concerned road. However, these models require the knowledge of the kinematic wave speed (i.e. the propagation speed of congestion frontline), which is not easy to obtain in practice. In this paper, we devote to constructing effective models to explicitly accommodate the impedance effect of intersections on urban road speed.

3 Problem Formulation

Suppose there is a centralized traffic control center in an urban network, and the control center considers the traffic network as a discrete-time system; in other words, the time horizon is divided into discrete time intervals whose length is τ seconds. The traffic conditions are supposed to stay constant within each time interval. Real-time traffic data are aggregated over each time interval and then transmitted to the traffic control center. Speed estimation is conducted every time a new set of traffic data, e.g. trafic volume, becomes available at the control center. This framework is similar to the rolling horizon approach [7] that is widely adopted in various traffic control systems. Suppose there are traffic sensors (e.g. loop detectors and probe cars) installed on all roads. The sensors provide traffic speed and volume

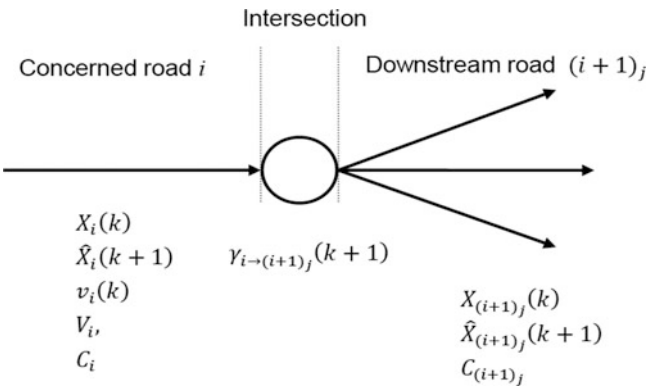


Fig. 1 Summary of model input.

data aggregated during time interval k, and traffic volume in time interval $k + 1$ can be predicted using existing methods [22]. We formulate the urban road speed estimation problem as follows.

*Definition*1: Given $X_i(k)$, $X_{(i+1)_j}(k)$, $\hat{X}_i(k+1)$, $\hat{X}_{(i+1)_j}(k+1)$, $v_i(k)$, V_i, C_i, $C_{(i+1)_j}$, $\gamma_{i\to(i+1)_j}(k+1)$, find $\hat{v}_i(k+1)$,

where $\hat{v}_i(k+1)$ is the estimated average speed on the concerned road i in time interval $k + 1$, $X_i(k)$ and $X_{(i+1)_j}(k)$ are the observed traffic volume on the concerned road i and its j-th downstream in time interval k, $\hat{X}_i(k+1)$ and $\hat{X}_{(i+1)_j}(k+1)$ are the predicted traffic volume on road i and its j-th downstream in time interval $k + 1$, $v_i(k)$ is the observed average speed on road i in time interval k, V_i is the free flow speed on road i, C_i and $C_{(i+1)_j}$ are the queuing capacity of road i and its j-th downstream, and $\gamma_{i\to(i+1)_j}(k+1)$ is the ratio of vehicles that intend to enter the j-th downstream out of all the vehicles on road i in time interval $k + 1$. The input variables are illustrated and summarized in Fig. 1.

4 The Proposed Model

4.1 Assumptions

Before describing the proposal in detail, we present several assumptions here. We presume that the configurations of traffic lights are fixed and known. The split rate of traffic flows at the intersections are obtained beforehand, e.g. via vehicle tracking [8] or by collecting drivers feedback on their route choice [9]. Moreover, the split rates should satisfy (2) and (3).

$$\gamma_{i\to(i+1)_j}(k+1) = \frac{1}{J} \quad \text{if} \quad \hat{X}_i(k+1) = 0 \quad \text{for} \quad \forall i,j \quad (2)$$

$$\sum_j \gamma_{i\to(i+1)_j}(k+1) = 1 \tag{3}$$

where J is the total number of downstream roads that road i has.

4.2 Model Construction

We propose a speed-volume relationship model represented by (4). The outcome of this model is a weighted average over the speed of vehicles that turn in each direction at the exit intersection. For vehicles that turn in a certain direction and enter the corresponding downstream, the speed is an adjustment of GM using split rate of this direction and traffic conditions on corresponding downstream road. It is worth noticing that the capacity used in (4) is also the queuing capacity [20] of a road, which is defined as the number of vehicles that can be stored on the road in a queue.

$$v_i(k) = v_i^{GM}(k)\sum_j \gamma_{i\to(i+1)_j}(k)\cdot exp[-\frac{\gamma_{i\to(i+1)_j}(k)\cdot X_i(k)}{C_i + C_{(i+1)_j}}] \tag{4}$$

$$v_i^{GM}(k) = V_i[1 - \frac{X_i(k)}{C_i}] \tag{5}$$

Since speed estimation is a major application of the above model and the goal of such estimation research is to achieve high accuracy, we further utilize real-time traffic data available to potentially improve estimation accuracy. We distinct two special cases: $\hat{X}_i(k+1) = 0$ and $|\Delta\hat{X}_i(k+1)| = |\hat{X}_i(k+1) - X_i(k)| \le \varepsilon$, where ε is a small number. In the former case, there is no vehicle on road i, consequently the estimated speed should be the maximum speed on this road; in the latter case, the change in traffic volume is not enough to cause significant change in speed, therefore, it is reasonable to assume that the estimated speed in time interval $k + 1$ is equivelent to the measured value in previous time interval k.

Based on the speed-volume relationship model as well as the usage of real-time data for two special cases, the speed on an urban road is estiamted using (6).

$$\hat{v}_i(k+1) = \begin{cases} V_i & \hat{X}_i(k+1) = 0 \\ v_i(k) & \hat{X}_i(k+1) \neq 0 \quad \text{and} \quad |\Delta\hat{X}_i(k+1)| < \varepsilon \\ \Omega(k+1) & \hat{X}_i(k+1) \neq 0 \quad \text{and} \quad |\Delta\hat{X}_i(k+1)| \ge \varepsilon \end{cases} \tag{6}$$

and $\Omega(k + 1)$ can be calculated by (4) with $X_i(k + 1)$ and $X_{(i+1)_j}(k+1)$ replaced with $\hat{X}_i(k+1)$ and $\hat{X}_{(i+1)_j}(k+1)$ respectively.

5 Accuracy Evaluation on Speed Estimation

In the previous section a speed estimation model for urban roads has been proposed. We use traffic data of Cologne, Germany to evaluate the effectiveness of the proposed model. We employ the microscopic traffic simulator SUMO [10] to simulate and collect the real speed on individual roads, mainly because of the unavailability of the real speed data in urban scenarios. We input the real data of traffic demand between 6:00 and 8:00 a.m. of a day available in the "TAPAS Cologne" Scenario [11] to the simulator and collect the average travel speed on the roads as the real values for our evaluation.

There are in total 97828 roads in "TAPAS Cologne"; part of the roads have more than one lanes, and some intersections are equipped with traffic lights. The topology of the traffic network is shown in Fig. 2. The default setting in SUMO 0.15.0 is used to configure vehicles. The vehicle length is 5m, which is a reasonable value according to [12]. The minimal gap between vehicles is 2.5m, and the Krauss model [13] is used as a car following model. We use the proposed model to estimate average speed on 12 randomly selected roads from the traffic network.

We compare the estimation accuracy of the proposed model with that of the GM which is represented by (1) in terms of Mean Absolute Error (*MAE*) and Symmetric Mean Absolute Percent Error (*SMAPE*) [14]. The definitions of these measures are shown in (7) and (8) respectively.

$$MAE = \frac{1}{I}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K}|\tilde{v}_i(k) - v_i(k)| \tag{7}$$

$$SMAPE = \frac{1}{I}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\frac{|\tilde{v}_i(k) - v_i(k)|}{\tilde{v}_i(k) + v_i(k)} \tag{8}$$

where I and K are the number of roads studied and the total number of time intervals respectively, $v(k)$ are the values collected from the simulations in SUMO, which are considered as real traffic data, while $\tilde{v}(k)$ are the estimated values. In this paper I is 12 as we studied 12 randomly selected roads, and the value of K varies on each road depending on the duration of traffic demand. It is worth mentioning that we only include the estimation results when the traffic volume is none zero on the concerned road to ensure meanful comparisons. The effectivness of the proposed model in

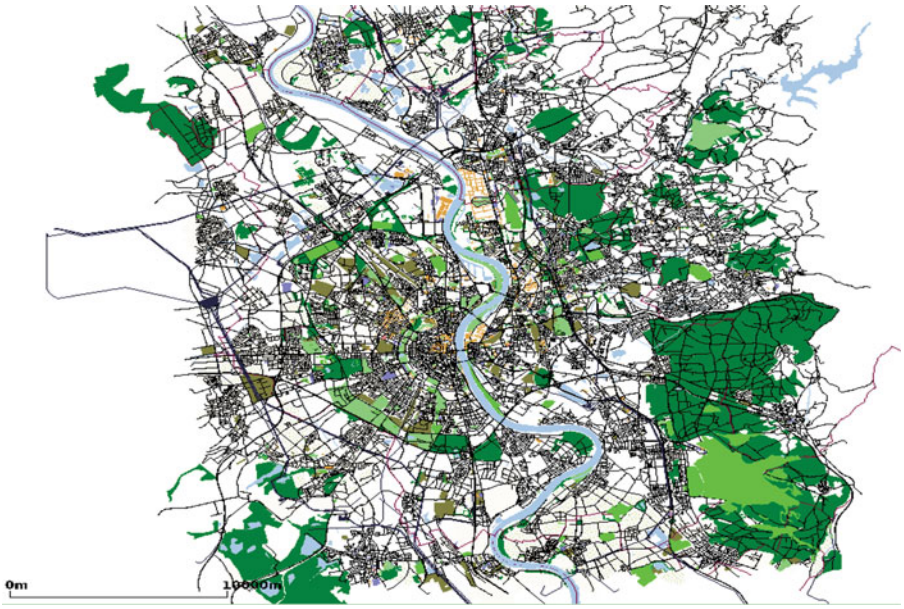


Fig. 2 The topology of Cologne City, Germany.

term of the reduction of *MAE* and *SMAPE* based on that of GM are denoted by ΔMAE and $\Delta SMAPE$. The definition of these two values are given in (9) and (10), where $MAE^{proposal}$ and MAE^{GM} represent the *MAE* of the proposed model and the GM respectively, and the same rule goes with *SMAPE* as well.

$$\Delta MAE = \frac{MAE^{GM} - MAE^{proposal}}{MAE^{GM}} \times 100\% \quad (9)$$

$$\Delta SMAPE = \frac{SMAPE^{GM} - SMAPE^{proposal}}{SMAPE^{GM}} \times 100\% \quad (10)$$

There is a tuning parameter ε in the proposed model, which characterizes the shifting threshold of the model. In the evaluation, we set ε to typical values from 0 to 3 increasing at a step of 0.5. The corresponding result for the proposed model under a fixed ε value is denoted as "proposal-ε". The time interval τ is set to 300 seconds. The average estimation errors over 12 roads under a certain ε value is demonstrated in TABLE 1. The proposed model considerably improves estimation accuracy regardless of the value of ε, which confirms the effectiveness of the modeling approach. The estiamtion accuracy of the model reaches its maximum when $\varepsilon = 2$, which translates into on average more than 50% reduction in *MAE* and *SMAPE* compared to GM. The speed-volume relationship alone ($\varepsilon = 0$) also yields 24% reduction of *MAE* and 27% reduction of *SMAPE*. This indicates that even when $v_i(k)$ is not available, the relationship model itself still works better than GM to estimate $\hat{v}_i(k+1)$. From practical point of view, the proposed model is a good candidate for speed estimation, as its *SMAPE* is constantly lower than 20%.

Table 1 Error Analysis

Model	*MAE*	ΔMAE	*SMAPE*	$\Delta SMAPE$
GM	2.89	/	27.2%	/
Proposal-0	2.19	24.3%	19.8%	27.1%
Proposal-0.5	1.60	44.9%	14.6%	46.1%
Proposal-1	1.40	51.7%	12.9%	52.2%
Proposal-1.5	1.34	53.8%	12.9%	52.2%
Proposal-2	1.33	53.9%	12.3%	54.7%
Proposal-2.5	1.38	52.2%	12.9%	52.6%
Proposal-3	1.62	43.9%	14.9%	45.0%

With respect to the characteristics of different models, we discovered the following phenomenon. First, the GM produces similar accurate estimation when the traffic volume is low and the speed is high. However, as traffic volume increases, the accuracy of GM sharply decreases, whereas our proposed model still works well. Second, the GM may have the tendancy of overestimation. The fundamental reason for such overestimation lies in the fact that GM does not consider the capacity of the exit intersection of urban roads, which could be low especially under congested traffic conditions. Consequently, the low capacity of exit intersection would become a strong impedance to the traffic flow on the road, and the average speed would thereby be considerably reduced. As GM was origianlly proposed for highways where intersection is rare, it cannot capture such impedance effect on travel speed by an exit intersection and thus tends to yields overestimated speed. Third, the GM can capture the

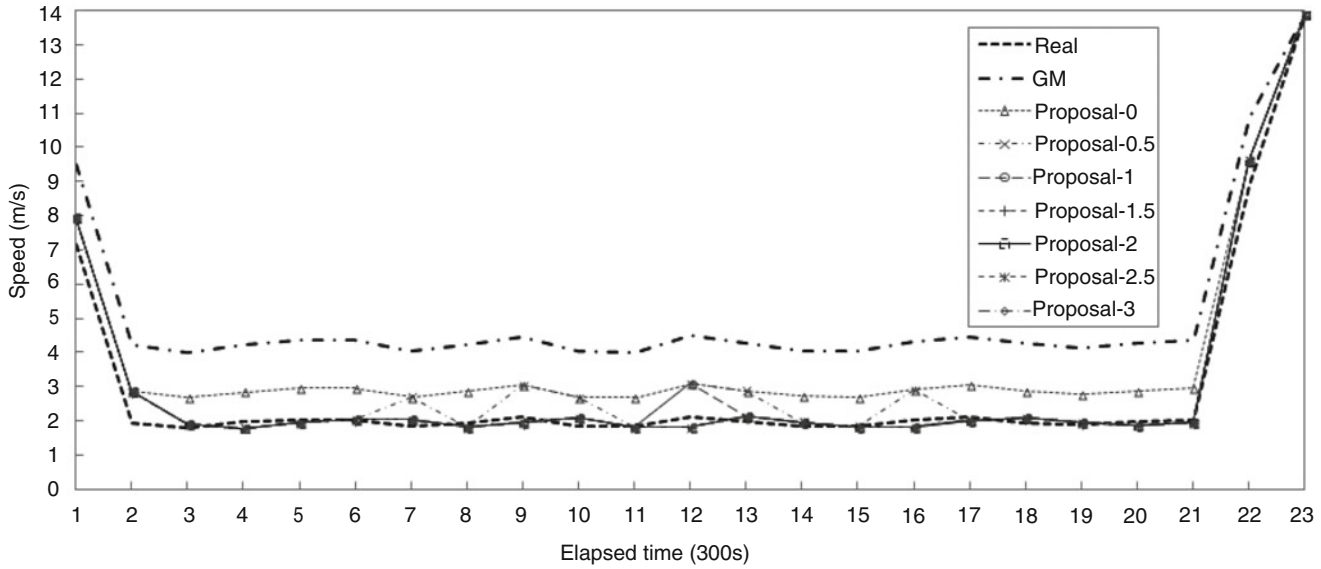


Fig. 3 Comparison between estimated speed and real speed versus elapsed time on Road-1 (ID=-10272255♯ 1, length=227m, max.speed=14m/s, 1 lane, two downstreams with 1 lane, traffic lights equipped at the exit intersection).

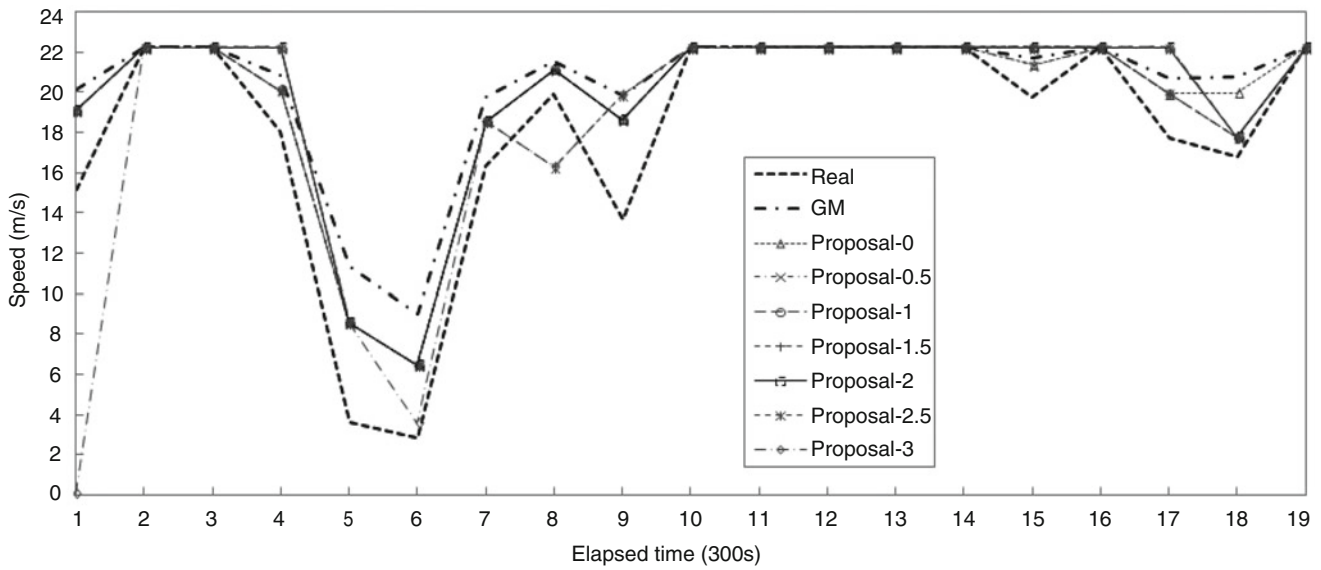


Fig. 4 Comparison between estimated speed and real speed versus elapsed time on Road-2 (ID=-24535297♯ 2,length=210m, max.speed=22m/s, 1 lane, two downstreams with 2 lanes, no traffic lights equipped at the exit intersection).

change in traffic volume in a timely manner, while the proposed model is also sensitive to such change under properly selected ε. In some cases, the proposed model becomes not sensitive to the sudden change in the traffic conditions if the value of ε is too large. Last, analysis of the estimation results for the 12 roads suggests that, besides the tuning parameter ε, the traffic profile on the concerned road also has strong impact on the accuracy of the estimation model. However, factors such as the existence of traffic light at the exit intersection, the length of the concerned road and its downstreams, the number of lanes do not demonstrate obvious impact. Due to the space limit of this paper, we only selectively demonstrate the time series of the estimation results on two typical roads in Fig. 3 and 4.

6 Conclusions

Greenshield's Model (GM) has been widely used in transportation research community to estimate average speed from measurable data, e.g. traffic flow, on highway. In recent years, this model is also applied to estimate speed from traffic volume on urban roads. However, due to the difference between highways and urban roads, we believe that GM may not a good candidate for urban traffic study. Correspondently, we have proposed an effective speed estimation model for urban roads by incorporating the impedance effect of exit intersections, which is strongly related to the traffic conditions on downstream roads. The evaluation results confirmed the effectiveness of the proposed modeling approach, as the estimation errors by the proposed model are constantly lower than that by GM. By properly selecting the tuning parameter, the proposed model successfully reduces estimation error by more than 50% in the studied urban traffic network. On the other hand, the proposed model also satisfies the practical requirement of speed estimation, as its relative errors are constantly lower than 20%. Furthermore, we have noticed that GM has the tendancy of overestimation, especially when traffic volume is high. Since speed is often used as an indicator of urban traffic congestion by road authorities to make control decision, the overestimation of speed could give a distorted picture of the traffic conditions and thus impede effective traffic control and management in urban areas. This study suggests that we should be carefull when we apply existing models to new scenarios. In the future work, we intend to extend the proposed model to cover the following situations: 1)roads with multiple lanes; 2)traffic conditions when accident happens; 3)application of the proposed model to traffic control systems.

References

1. Jayakrishnan, R., Tsai, W.K., Chen, A.: A dynamic traffic assignment model with traffic-flow relationships. Transportation Research-C 3 (1995) 51-72.
2. Balakrishna, R.: Calibration of the demand simulator in dynamic traffic assignment system. Master Thesis, MIT (2002).
3. Liang, Z., Wakahara, Y.: Real-time urban traffic amount prediction models for dynamic route guidance systems. EURASIP Journal on Wireless Communications and Networking (2014).
4. Daganzo, D.F.: Queue spillovers in transportation networks with a route choice. Transportation Science 32 (1998) 3-11.
5. Greenshields, B.D.: A study in highway capacity. Highway Research Board, Proceedings 14 (1935) 458.
6. Dhamaniya, A., Chandra, S.: Speed prediction models for urban arterials under mixed traffic conditions. Procedia-Social and Behavioral Sciences 104 (2013) 342-351.
7. Pan, T.L.; Sumalee, A.; Zhong, R.X.; Indra-Payoong, N., "Short-Term Traffic State Prediction Based on TemporalSpatial Correlation," IEEE Transactions on Intelligent Transportation Systems 14 (2013) 1242-1254.
8. Sivaraman, S.; Trivedi, M.M.,: Integrated Lane and Vehicle Detection, Localization and tracking: a synergistic approach. IEEE Transactions on Intelligent Transportation Systems 14 (2013) 906-917.
9. Choi, J.M.: Multi-touch based standard UI design of car navigation system for providing information of surrounding areas. Design, User Experience, and Usability, User Experience in Novel Technological Environments. Lecture Notes in Computer Science 8014 (2013) 40-48.
10. Krajzewicz, D., Erdmann, J., Behrisch, M., and Bieker, L.: Recent development and applications of SUMO Simulation of Urban MObility. International Journal on Advances in Systems and Measurements 5 (2012) 128-138.
11. "TAPAS Cologne" Scenario. http://sumo-sim.org/userdoc/Data/Scenarios/TAPASCologne.html
12. Gazis, D.C.: Optimum control of a system of oversaturated intersections. Oper. Res. 12 (1964) 815-831.
13. Hazelton, M.L.: Estimating vehicle speed from traffic count and occupancy data. Journal of Data Science 2 (2004) 231-244.
14. Flores, B.E.: A pragmatic view of accuracy measurement in forecasting. Omega Int. J. of Mgmt Sci. 14 (1986) 93-98.
15. Hyndman, R.J.: Another look at forecast accuracy metrics for intermittent demand. International Journal of Applied Forecasting 2006 (2006) 43-46.
16. Pan, J., Popa, J.S., Zeitouni, K., and Borcea, C.: Proactive vehicular traffic rerouting for lower travel time. IEEE Trans. Veh. Tech. 62 (2013) 3551-3568.
17. Greenberg, H.: An analysis of traffic flow. Operations Research 7 (1959) 79-85.
18. Underwood, R.T.: Speed, volume, and density relationships: quality and theory of traffic flow. Yale Bureau of Highway Traffic (1961) 141-188.
19. Rakha, H., Growther, B.: Comparison of Greenshields, Pipes, and Van Aerde car-following and traffic stream models. Transportation Research Record 1802/2002 (2007) 248-262.
20. Hensher, D.A., Button, K.J.: Handbook of trnasport modelling [Chapter 10]. Elsevier, UK, 2008.
21. Castillo, J.M.: Three new models for the flow-density relationship: derivation and testing for freeway and urban data. Transportmetrica 8 (2012) 443-465.
22. Liang, Z., Wakahara, Y.: City traffic prediction based on real-time traffic information for intelligent transportation systems. Proc. of ITST (2013) 378-383.

Superpixel based semantic segmentation for assistance in varying terrain driving conditions

Ionut Gheorghe, Weidong Li, Thomas Popham, and Keith J. Burnham

1 Introduction

In pursuit of enhanced drivability, maneuverability and safety, modern vehicles are made aware of the surrounding environment using a vast array of sensors. Vehicle perception has been directly linked to ADAS (advanced driver assistance systems) in recent years [8]. Despite the availability of a wide range of sensors, video cameras are still preferred to laser and T-O-F (time of flight) [5], [13] as a prerequisite for vehicle perception due to their cost. This work investigates the problem of environment perception focusing on driving conditions incurred by various terrain surfaces. A good recognition system would allow the already existent subsystems within the vehicle to be reconfigured autonomously. Throttle response, suspension stiffness, direction response, differential locking, to name a few, are all subject to terrain changes given the driving path of a vehicle.

In this work a holistic approach is taken to predict all image superpixels from a forward driving perspective. The following classes will be under scrutiny: {grass, tree, sky, tarmac, dirt, gravel, shrubs}. The images have been recorded with a stereo camera (Bumblebee2) thus allowing future investigation into the 3D realm. At this stage however, only colour images are used (Fig. 1).

Previous work associated with terrain classification has been mostly related to the robotics community using colour, stereo camera and laser [4], [12], [15], [16], [17].

The literature addressing road/terrain changes using colour camera is scarce [6], [7], [9], [12], and [22]. In previous work, road classification was investigated for regions of interest within colour images [9], [22]. Moving away from the problem domain, there is an abundance of techniques describing how to segment an image more generally. State of the art results in semantic image segmentation are typically obtained by predicting image parts (e.g. superpixels) using a classifier and then reinforcing some spatial constraints under a probabilistic graphical model framework (e.g. MRF, CRF) [3], [11], [18], [19], [26]. These constraints often translate to statements like "sky is above tarmac, tree is under the sky" and form the basis for a structured prediction problem. While there is no doubt that spatial constraints are useful in practice, there is also the risk of enforcing structures, thus adding prior knowledge might lead to wrong predictions. For now at least, superpixel predictions will be made independently of one another and cast as semantic segmentation.

2 Superpixel extraction

Simple linear iterative clustering (SLIC) [1] is a very efficient superpixel algorithm that exhibits good adherence to image boundaries (Fig. 2, left). It uses a k-means clustering approach in the CIELAB colour space. Superpixels correspond to clusters found by k-means in the *labxy* colour image plane space after seeding k centroid points.

$$C_k = [l_k, a_k, b_k, x_k, y_k]^T \tag{1}$$

However, the typical pixel assignment to certain clusters is done based on a formulated distance metric that normalizes colour and spatial proximity in their respective ranges. Furthermore it and allows the selection of a relative importance between spatial and colour consistency.

I. Gheorghe (✉) • W. Li • T. Popham • K.J. Burnham
Control Theory and Applications Centre, Coventry University, Coventry CV1 5FB, UK

Jaguar & Land Rover Research, University of Warwick, Coventry CV4 7AL, UK
e-mail: gheorghi@uni.coventry.ac.uk; aa3719@coventry.ac.uk; ctac@coventry.ac.uk; tpopham@jaguarlandrover.com

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_98,

Fig. 1 Images recorded using Bumblebee2 stereo camera

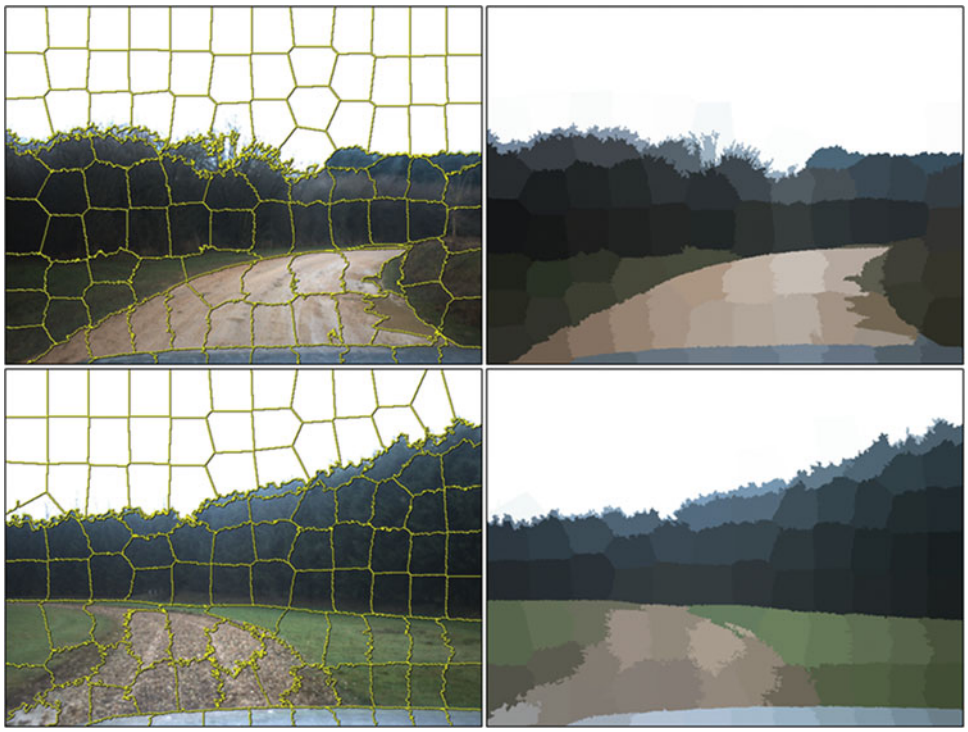


Fig. 2 SLIC and average superpixel colour. Dirt and gravel have different texture patterns

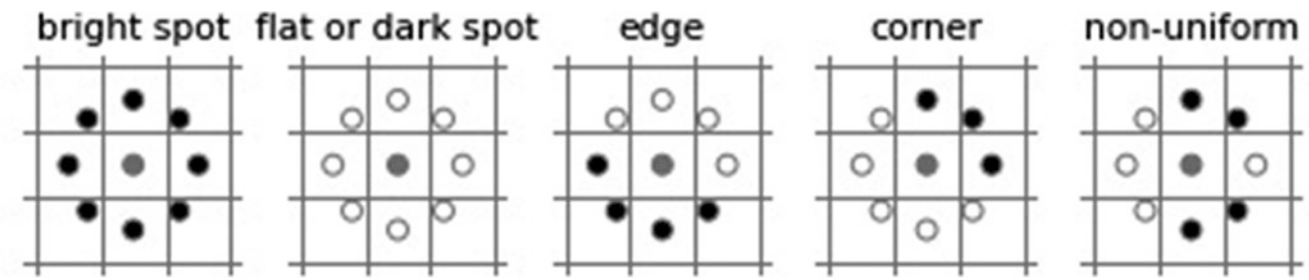


Fig. 3 The typical unit radius, 8 points LBP with {0, 1, origin} as {black, white, gray} circles

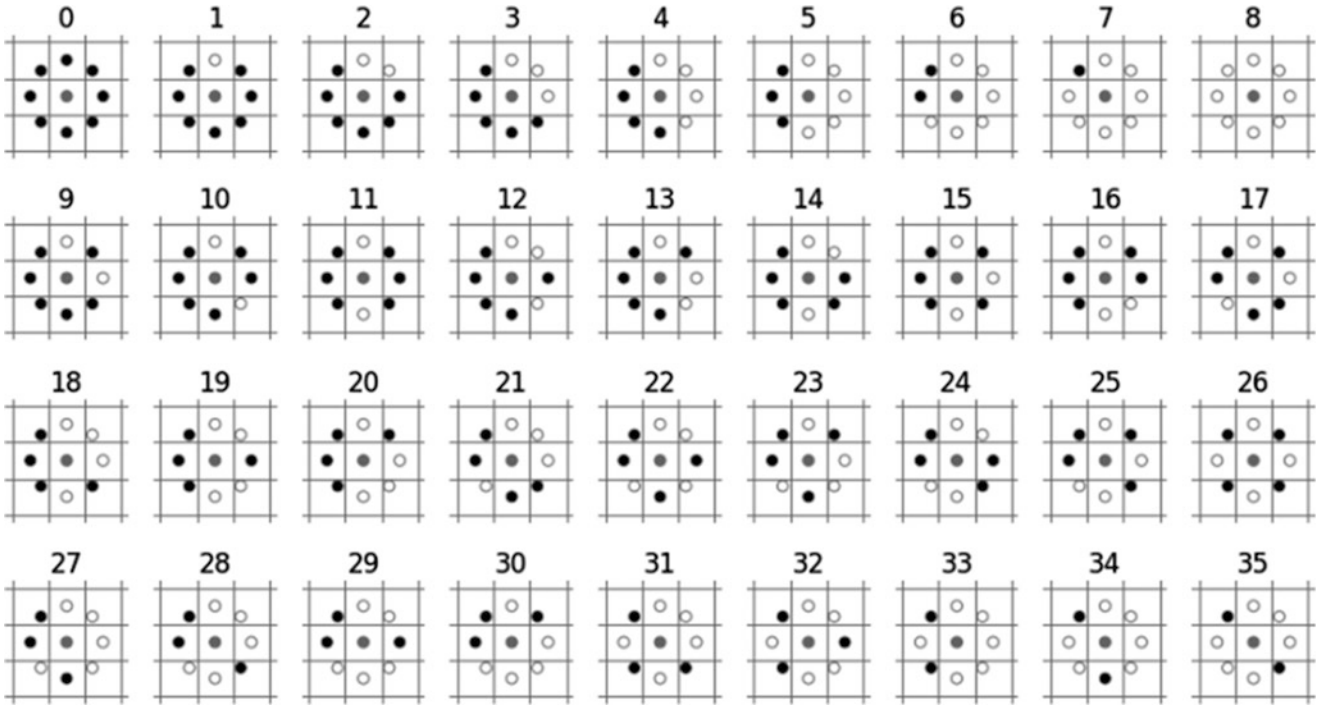


Fig. 4 Unique rotation invariant patterns; first 9 patterns are uniform and the rest are non-uniform

3 Feature extraction

The primary lenses through which the learning algorithm sees the world of this particular scenario are the extracted features. They are formed via a concatenation of average colour with gradient based descriptors whose purpose is to act as texture discriminants between similarly coloured classes.

3.1 Average superpixel colour

After the segmentation step each superpixel is a locally coherent collection of pixels. Their average RGB colour is computed in an attempt to extract three of the feature dimensions.

$$R_a,\ G_a,\ B_a = \frac{1}{n} \sum\nolimits_{i \in \text{ superpixel}} R_i,\ G_i,\ B_i \qquad (2)$$

3.2 Local binary pattern (LBP)

Local binary patterns [20] are texture descriptors that have been successfully used in many classification problems. Good performance has been reported for problems like face and facial expression recognition [2], [21], human detection [25], foreground detection [14], pavement crack detection [10], texture analysis [20], [27] and even visual terrain classification [15], [16] albeit for robotic applications, to name a few. They are computationally efficient and can be made rotation invariant (Fig. 4) using a simple vector quantization technique [20].

In its simplest form, the LBP compares each pixel with the neighbors P on a circle of radius R and stores the result as follows:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) * 2^p,\ s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad (3)$$

Since the number of neighbors and radius are given (e.g. $P = 8, R = 1$) what really decides the value of $LBP_{P,\ R}$ is a sequence of P binary values, allowing for a total of 2^P possible outcomes. A histogram can be constructed at this point to perform recognition tasks. However, if the texture rotates just slightly around the reference pixel (and this is very likely in practice) the output will be different. To overcome this problem and obtain rotation invariance each pixel's binary sequence is shifted akin to a 'rotary dial' such that a maximum number of bits are 0, starting from the most significant one g_{P-1}. This gives rise to a total of 36 unique rotation invariant patterns (for $P = 8$). It has been observed [20] that some of these patterns account for more than 90 %

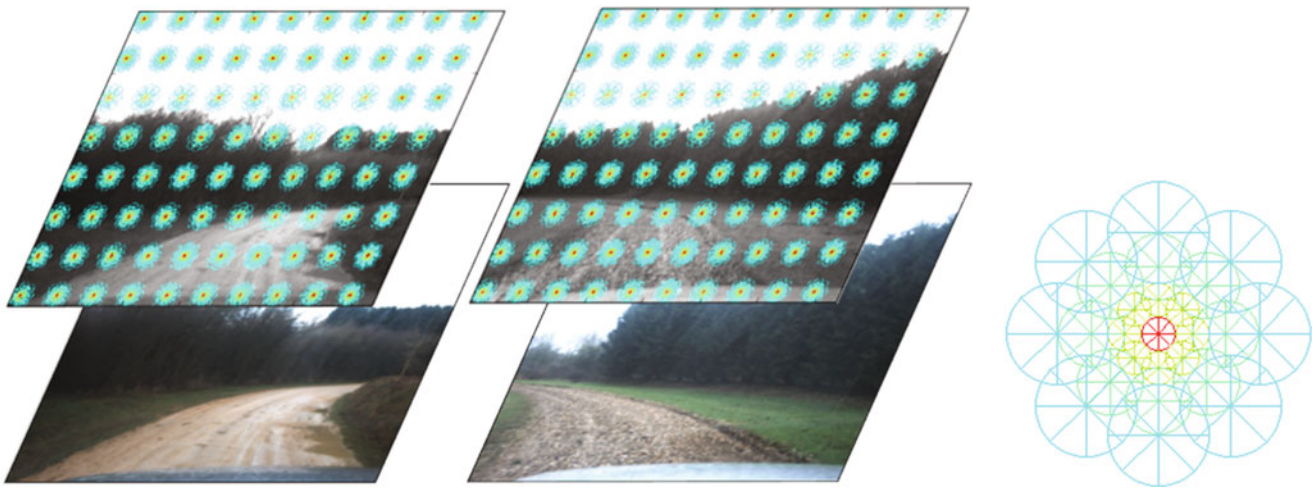


Fig. 5 Scarce DAISY descriptors overlaid on grayscale images for visualization

of the fundamental properties of textures, namely the uniform patterns (Fig. 4, line 1). By uniform pattern it is meant that there are at most two transitions between 1 and 0.

$$U(LBP_{P,R}) = \left|s(g_{p-1} - g_c)\right| + \sum_{p=1}^{P-1}\left|s(g_p - g_c) - s(g_{p-1} - g_c)\right| \quad (4)$$

$$LBP_{P,R}^{riu2} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c) & \text{if } U(LBP_{P,R}) \leq 2 \\ P+1 & \text{otherwise} \end{cases} \quad (5)$$

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (6)$$

Hence each descriptor is going to be quantized by a code (e.g. from 0 to 9). The uniform patterns will be represented by a number from 0 to 8 and the non-uniform patterns by the number 9.

3.3 Dense DAISY pixel descriptors

DAISY descriptor (Fig. 5) has been initially introduced for wide baseline stereo matching to overcome strong appearance changes [23]. It has also been used for face recognition tasks [24]. There are many similarities to SIFT descriptor, but DAISY can be computed densely which makes it suitable for this semantic labelling of pixel sets. Like SIFT, it too relies on gradient orientation histograms, but according to [23] it can be computed 66 times faster in the fine density setting.

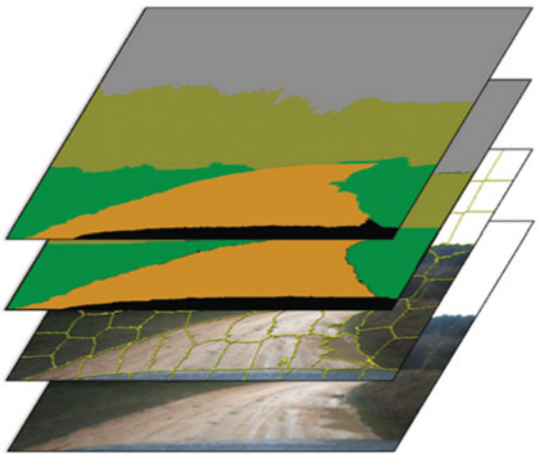
Orientation maps are defined in certain directions of the image as the positive values of the image gradient norm $M_o = \left(\frac{\partial I}{\partial o}\right)^+$. Convolving M_o with Gaussian kernels of increasing $\sum$ gives rise to different convolved orientation maps. They are fast to compute because larger kernel convolutions can rely on smaller consecutive ones. Each descriptor consists of a vector formed using convolved orientation maps by taking the values from concentric circles around the pixel location. The circle radius indicates the amount of Gaussian smoothing present in the convolved orientation map. It encloses the region for histogram calculation on a circular grid, as opposed to the regular grid used in SIFT. The result is a concatenation of normalized histograms on pixel neighborhood from all orientation maps.

4 Method discussion

4.1 Bag of words

The superpixels generated using SLIC on the raw image span a number of DAISY and LBP descriptors each. The bag of words strategy is taken in order to produce a feature vector of constant length that will work as a texture cue for each superpixel individually.

Firstly, DASY pixel descriptors are extracted on a coarser level from each image in the dataset. Secondly, a number of 300 prototypes are found using k-means on all descriptors. Such a procedure is typically referred to as codebook generation. Superpixels can now be represented in the form of a histogram, whereby each of the 300 bins holds occurrences of descriptors similar to each prototype. This vector quantization technique takes the original 200 dimensional



R, G, B	label
0, 128, 0	grass
128, 128, 0	tree
128, 128, 128	sky
128, 64, 128	tarmac
192, 128, 0	dirt
128, 64, 0	gravel
0, 192, 0	shrubs
0, 0, 0	void

Fig. 6 Colour-coded classes of interest

pixel descriptor into a simple bin count. In this case the final superpixel feature has 303 dimensions, 300 from DAISY and 3 from the average colour. With LBP there is no need for codebook generation, as each descriptor can only take on a value from 0 to 9. Each superpixel is represented by a histogram and likewise, the final 13 dimensional feature is formed by concatenating the average colour.

4.2 Ground truth

Evaluation of this method is done on a superpixel level and labels must be specified individually. Having the manually annotated ground truth image overimposed on SLIC image, each superpixel receives a label by a majority vote. Assuming that a superpixel spans across a region with several labels, only the most predominant subset would take over as label (Fig. 6). Superpixel annotations are done in a similar fashion to MSRC 21-class dataset.

5 Machine learning applied to superpixel features

As a machine learning solution, a linear kernel SVM has been used for multi-label classification via one-vs-one predictions.

In the one-vs-one approach there are a number of $\frac{n(n-1)}{2}$ classifiers each fitted for a pair of classes. At test time all classifiers vote for a class. The final prediction is given by the class receiving the most votes. To check the impact of DAISY on the classification accuracy, the total number of images was split into 80 % and 20 % random image sets for training and testing respectively. This partitioning process was repeated in order to investigate the LBP texture descriptor. Splitting whole images into sets rather than superpixels for cross-validation resulted in different superpixel distributions (for training and testing) in each experiment.

6 Experimental results

The original images of 1024x768 as captured by the Bumblebee2 camera have been rescaled to a resolution of 640x480 to speed up the processing times. Using SLIC resulted in the generation of 5937 superpixels.

A classification accuracy of 87 % was attained on the test set using DAISY and 76 % using LBP (Fig. 7). Although, this work presents the results obtained on an unbalanced data set, by far the most easily distinguishable class is “sky” (Fig. 8). At the opposite end, “shrubs” are often predicted as “tree” but since no real world height information is provided yet this would hold true even from a conceptual point of view.

7 Conclusion

A method to address the issue of semantic segmentation with an emphasis on terrain classification for possible future automotive applications has been presented. DAISY descriptors have been shown to perform better than LBP for texture discrimination in this particular scenario. By splitting images into superpixels using a state of the art algorithm and classifying them individually good

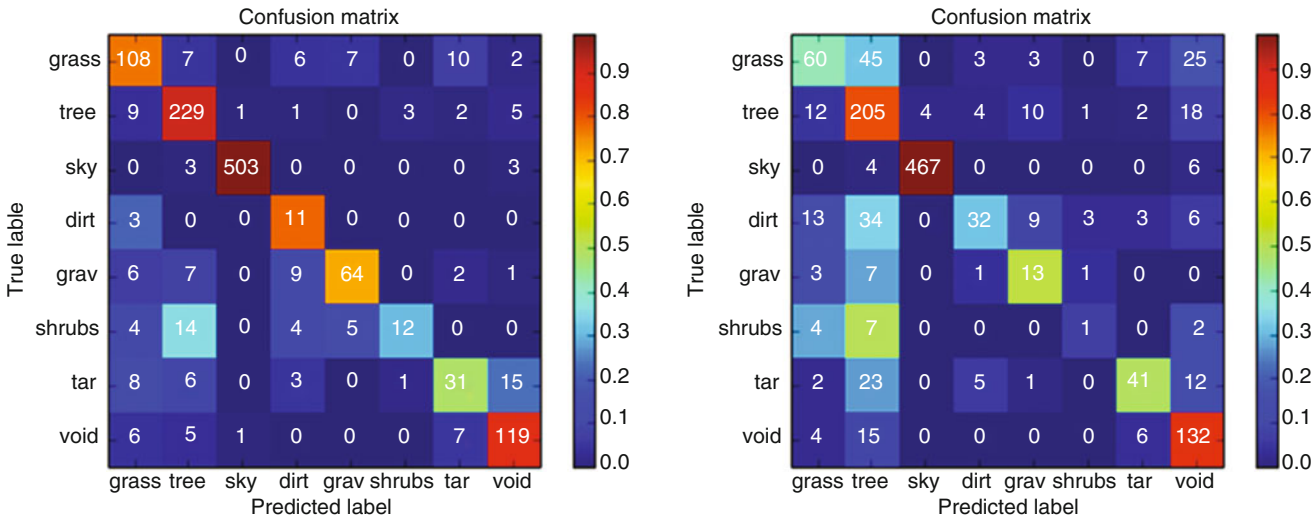


Fig. 7 Linear kernel SVM on DAISY+R_a, G_a, B_a (left), and LBP+R_a, G_a, B_a (right)

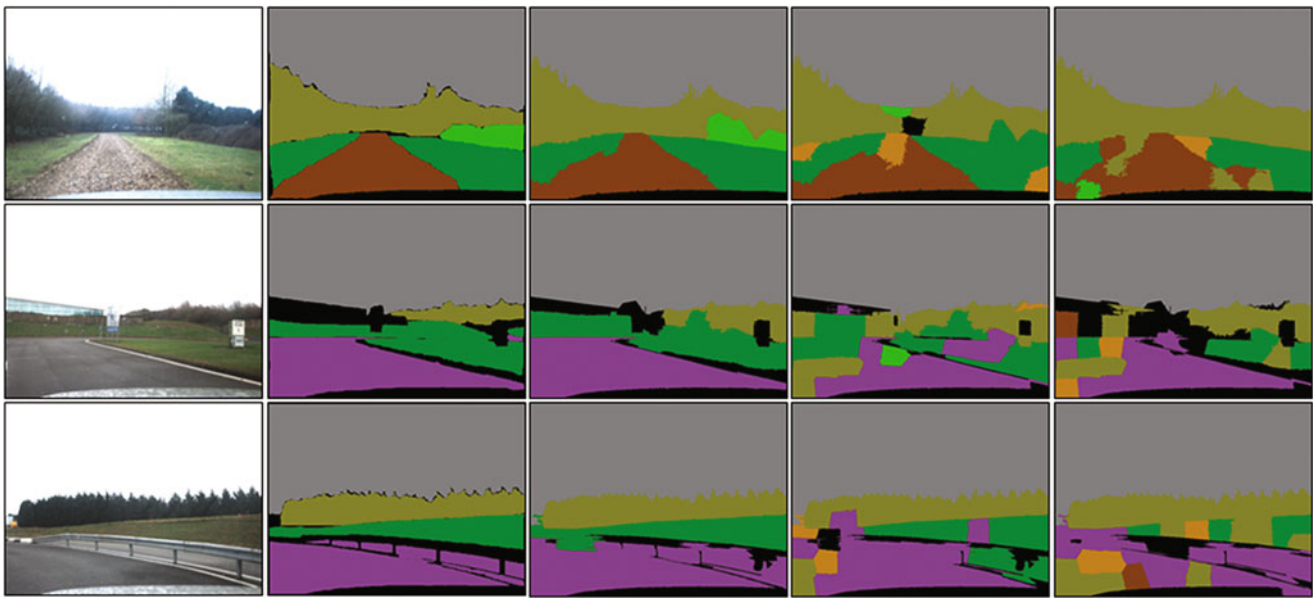


Fig. 8 From left to right: original image, {manually annotated, superpixel} ground truth, {DAISY, LBP} +R_a, G_a, B_a

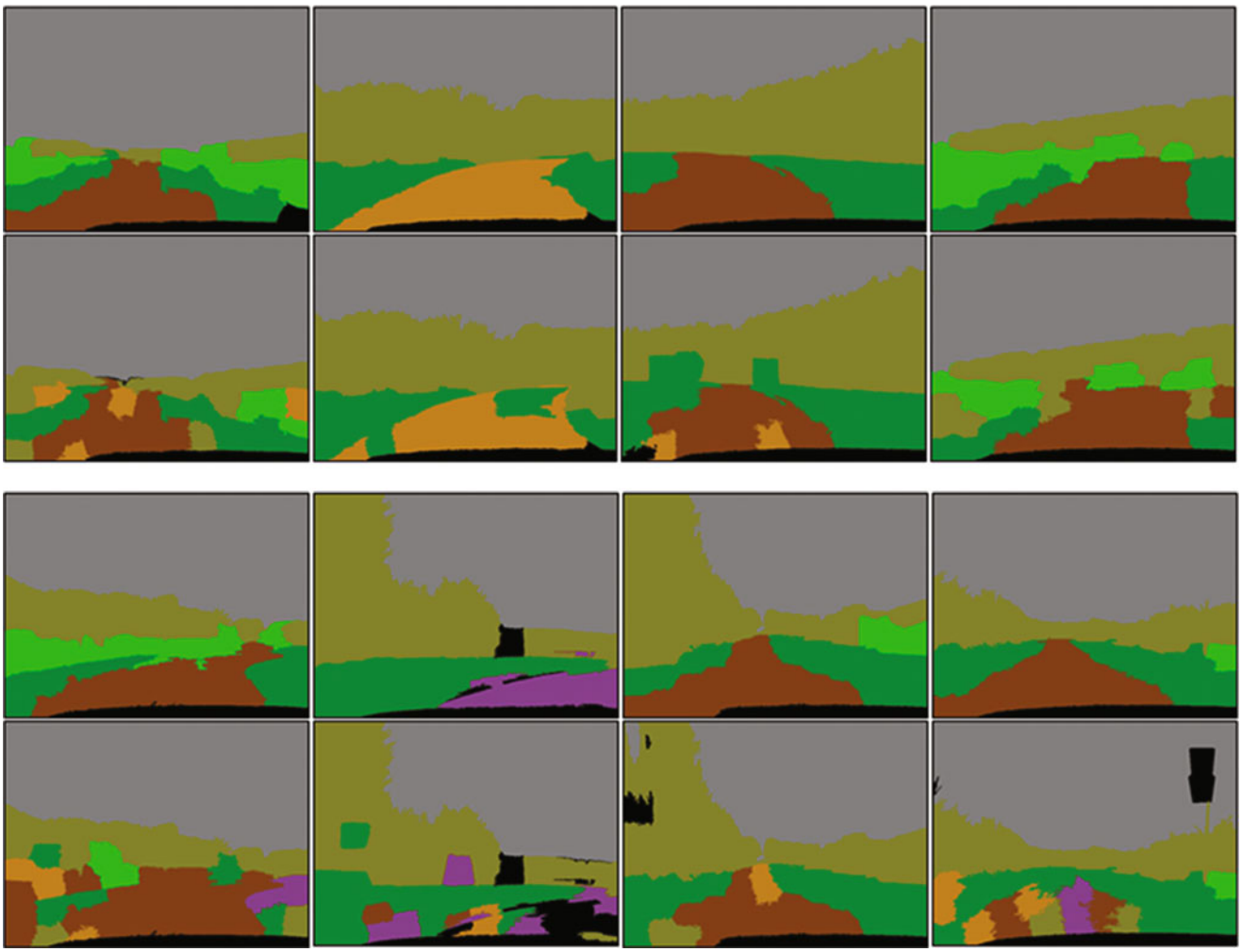


Fig. 9 Good segmentation results based on DAISY's superior texture discrimination

performance was achieved without enforcing any spatial priors (Fig. 9). As future work, classification would benefit from 3D stereo information, particularly where height and point normal distributions hold enough discriminative power.

Acknowledgements This work has been funded by the EPSRC (Engineering and Physical Sciences Research Council) in collaboration with Jaguar Land Rover.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. & Susstrunk, S. 2012, "SLIC superpixels compared to state-of-the-art superpixel methods", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274-2282.
2. Ahonen, T., Hadid, A. & Pietikäinen, M. 2004, "Face recognition with local binary patterns" in *Computer vision-eccv 2004* Springer, pp. 469-481.
3. Alvarez, J.M., LeCun, Y., Gevers, T. & Lopez, A.M. 2012, "Semantic road segmentation via multi-scale ensembles of learned features", *Computer Vision–ECCV 2012. Workshops and Demonstrations* Springer, pp. 586.
4. Angelova, A., Matthies, L., Helmick, D.M. & Perona, P. 2007, "Fast Terrain Classification Using Variable-Length Representation for Autonomous Navigation", *CVPR*.
5. Badino, H., Huber, D. & Kanade, T. 2011, "Integrating LIDAR into Stereo for Fast and Improved Disparity Computation", *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011 International Conference on*, pp. 405.
6. Beucher, S. & Yu, X. 1994, "Road recognition in complex traffic situations", *Proc. 7th IFAC/IFORS Symp. Transp. Syst.: Theory Appl. Adv. Technol.*, pp. 413-418.
7. Fernandez-Maloigne, C. & Bonnet, W. 1995, "Texture and neural network for road segmentation", *in Proc. Intell. Veh. Symp.*, pp. 344–349.
8. Geronimo, D., Lopez, A.M., Sappa, A.D. & Graf, T. 2010, "Survey of pedestrian detection for advanced driver assistance systems", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 7, pp. 1239-1258.
9. Gheorghe, I., Li, W., Popham, T., Gaszczak, A. & Burnham, K. 2014, "Key Learning Features as Means for Terrain Classification"

in , eds. A. Grzech, & J.M. Tomczak, Springer International Publishing, pp. 273-282.
10. Hu, Y. & Zhao, C. 2010, "A local binary pattern based methods for pavement crack detection", *Journal of pattern Recognition research*, vol. 1, no. 20103, pp. 140-147.
11. Ibrahim, M.S. & El-Saban, M. 2011, "Higher order potentials with superpixel neighbourhood (HSN) for semantic image segmentation", *Image Processing (ICIP), 2011 18th IEEE International Conference on* IEEE, pp. 2881.
12. Jansen, P., van der Mark, W., van den Heuvel, J.C. & Groen, F.C.A. 2005, "Colour based off-road environment and terrain type classification", *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*, pp. 216.
13. Jiejie, Z., Liang, W., Ruigang, Y. & Davis, J. 2008, "Fusion of time-of-flight depth and stereo for high accuracy depth maps", *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1.
14. Kertész, C. 2011, "Texture-Based Foreground Detection.", International Journal of Signal Processing, Image Processing & Pattern Recognition, vol. 4, no. 4.
15. Khan, Y.N. 2013, "Visual Terrain Classification for Outdoor Mobile Robots".
16. Khan, Y.N., Komma, P., Bohlmann, K. & Zell, A. 2011, "Grid-based visual terrain classification for outdoor robots using local features", *Computational Intelligence in Vehicles and Transportation Systems (CIVTS), 2011 IEEE Symposium on* IEEE, pp. 16.
17. Manduchi, R., Castano, A., Talukder, A. & Matthies, L. 2005, "Obstacle detection and terrain classification for autonomous off-road navigation", *Autonomous Robots*, vol. 18, pp. 81-102.
18. Micusik, B. & Kosecka, J. 2009, "Semantic segmentation of street scenes by superpixel co-occurrence and 3d geometry", *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on* IEEE, pp. 625.
19. Müller, A.C. & Behnke, S. 2013, "Learning a Loopy Model For Semantic Segmentation Exactly".
20. Ojala, T., Pietikainen, M. & Maenpaa, T. 2002, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971-987.
21. Shan, C., Gong, S. & McOwan, P.W. 2009, "Facial Expression Recognition Based on Local Binary Patterns: A Comprehensive Study", *Image Vision Comput.*, vol. 27, no. 6, pp. 803-816.
22. Tang, I. & Breckon, T.P. 2011, "Automatic Road Environment Classification", *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 2, pp. 476-484.
23. Tola, E., Lepetit, V. & Fua, P. 2010, "Daisy: An efficient dense descriptor applied to wide-baseline stereo", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 5, pp. 815-830.
24. Velardo, C. & Dugelay, J. 2010, "Face recognition with DAISY descriptors", *Proceedings of the 12th ACM workshop on Multimedia and security* ACM, pp. 95.
25. Wang, X., Han, T.X. & Yan, S. 2009, "An HOG-LBP human detector with partial occlusion handling", *Computer Vision, 2009 IEEE 12th International Conference on* IEEE, pp. 32.
26. Zhang, C., Wang, L. & Yang, R. 2010, "Semantic segmentation of urban scenes using dense depth maps" in *Computer Vision–ECCV 2010* Springer, pp. 708-721.
27. Zolynski, G., Braun, T. & Berns, K. 2008, "Local binary pattern based texture analysis in real time using a graphics processing unit", *VDIBERICHT*, vol. 2012, pp. 321.

Emotion Estimation using Geometric Features from Human Lower Mouth Portion

P. Shanthi and A. Vadivel

1 Introduction

Human emotion is the important dimension of cognitive science, which is the conscious experience, emerged from the interaction between various cognitive processes. Emotions differ from thoughts in several other aspects and its responses involve the body as well as the brain. This response is due to the physical changes such as alterations in body temperature, hormonal secretion, heart rate and muscle tension [1, 2]. Models are proposed to consider feedback from the body as an important factor while processing the emotions [3]. It is known that human face is considered as basic component and is used for interpersonal communication. Certain changes in the facial regions produced by the brain can be used to understand the current attention level of individuals in the current environment. In general facial expressions are caused by movements of underneath muscles in certain salient regions and are called as facial features. Facial Action Coding System (FACS) [4] manually interprets the facial expression. This approach contains facial muscle movements with 46 Action Units (AU) and a combination of these action units are used to identify various expressions. There are six common facial expressions such as angry, disgust, fear, happy, sad and surprise across various cultures [5]. Researchers from industry and academia have employed a wide range of approaches to automate and recognize emotions by studying changes in the facial region. Broadly, it is classified as appearance based and geometric based methods. The appearance based model measure the changes in the facial regions using texture and shape. The entire face is processed as a whole or divided into a number of sub regions. The Active Appearance Model (AAM) uses the whole appearance of the face and an Active Shape Model (ASM) uses local appearance of image information from all subspace. In ASM, also known as smart snakes [6], a model has been constructed to describe the mean shape which is expressed as the sum of a mean shape and a linear combination of mean shapes with the help of the Point Distribution Model (PDM). The AAM is similar to the template matching, where a model is generated by combining shape variations with appearance variations from the labeled training sample. Even though, AAM represents most of the variation in facial images of training sample in a compact way, the distance between the appearance vectors of different people with similar expression is found to be irrelevant. The limitations of AAM are addressed in [7] and provide a solution by proposing person invariant expression representation. This makes the traditional AAM to recognize expression of the unknown person using basic classifier. In contrast, the geometric based approach, concentrates only on actual and relative location of the facial features based on the priori information. The geometric relationship can be used to find the similarities between different persons with a similar expression. A mathematical analysis has revealed that structure present in images with facial expressions is sufficient, in principle, to generate some of the structure of the emotion categories that humans perceive [8]. In this paper, we consider a frontal face in static images to find the deformation using neutral state as the reference to recognize some basic expressions of the lower face (mouth). The geometric features are characterized by the combination of basic transformation without manual intervention in all phases. Even though, information required to discriminate the basic expressions equally distributed in all facial components, our aim is to estimate some basic expression from minimal details by assuming all other facial components are under occlusion. Associative Recurrent Neural Network (ARNN) is used for classification and it is observed that the proposed approach achieves the better recognition rate.

The rest of this paper is organized as follows: We review the related works and its limitations in Section 2. In Section 3, the proposed approach is presented. Experimental results

P. Shanthi (✉) • A. Vadivel
Cognitive Science Research Group, Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamil Nadu, India
e-mail: shanthicse@mamce.org; vadi@nitt.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_99,

explain the effectiveness of our approach with comparative results in Section 4 and finally we conclude the paper in the last section of the paper.

2 Related works

The presence of irrationality and complexity in emotion makes the cognitive parameter trivial in the earlier study of cognitive science. Even though, the need for incorporating emotions into the cognitive theory was pointed out earlier, it has been noticed that there are only few attempts made to study emotion within the cognitive modeling framework. First emotion recognition theory related with facial expression in man and animal is proposed by [9]. This theory explains the similarity of muscle movement between man and animals in certain situation such as concentration, anger and efforts of memory. Recognizing emotion through facial sign language from the observer point of view is the key idea behind the facial expression analysis. A recent study [10] have proposed a hybrid method, which consider appearance features from salient patch based Gabor and geometric features by calculating distance after patch matching process. One of the drawbacks of this method is that wrong selection of patches for matching leads to low recognition rate. Human visual cortex like encoding system using Gabor filters has been proposed [11] to generate a global representation of facial expression with the help of local classifier and also investigate issues of partial occlusion. The optimal value of some design parameters used in this method is very difficult to fix and more efficient method is required to integrate local feature. The shape and texture based method has been introduced [12] and uses SMOM- ESTM algorithms to represent a facial expressive model of the input image based on the probability of occurrence of pixel value of the training data. However, the recognition rate of this approach is low, in case, the expression details are intermingled with other expression.

Instead of considering the information from regions of salient facial components, geometric based method exactly considers the location of facial components and the relative geometric relationship among other feature points. The geometric feature based tracking system has been proposed [13] and uses the minimal subset of nodes from Candide grid model to recognize the basic expression. One of the drawbacks to this approach is that the initial model fitting takes more time and manual intervention is needed for the process. This work is extended by using shape and texture feature and named as Discriminant Non-negative Matrix Factorization [14] to investigate how the partial occlusion affects the recognition accuracy of basic facial expression. Another geometric based model has been proposed to capture the temporal information of facial action using particle filtering with factorized likelihoods tracking method [15]. Four corner points are used to cover entire mouth portion and capture subtle changes to the lower portion of the face for recognizing expression like disgust, sadness and anger. The graph based feature point tracking method to represent the basic expression using the optimal graph node structure has also been proposed [16]. Even though, it reduces the number of nodes to represent the expression, it requires a very large training sample to achieve a good recognition rate. Another geometric based tracking method [17] use image ratio features, which consists of both facial animation parameter and skin deformation parameter to handle the illumination variations.

The variation in the level of the psychological states of the individuals in a particular context causes same emotion as person variant. This issue is handled by optical flow based tracking method which is one of the alternative ways to recognize expressions. But this approach is usually affected by the environmental change. Based on the study by [18], expression recognition from masked mouth is more difficult compared to masked eyes. The Pyramid of Local Binary Pattern (PLBP) [19], has been used to represent the spatial layout of local texture information from the salient feature regions identified through psycho-visual experiment. The ensemble feature based expression classification approach [20] uses appearance and motion vector features with Support Vector Machine (SVM). However, its recognition result is low in some expression as it is mixed with other expression. A method to track the smiling expression using mouth visual feature is proposed [21]. This tracking method is based on the a priori information on the relative location of mouth with respect to the eye. Some of the factors such as illumination variation, occlusion with glass and head orientation may affect this a priori information. Combination of appearance and motion feature descriptor based technique to capture the dynamic nature of facial deformation is introduced [22] to recognize the basic expression. Since, it requires a different kind of descriptor, it takes more time to calculate different features. Recent study [23], perform the exhaustive search for the static images using non-rectangular emotional mask specifically designed to detect smiling expression. However, it is found that the mask creation is a time consuming process. Subset feature selection based on estimating distribution algorithm [24], reported that feature selection process improves the accuracy and complexity in terms of number of features used for classification. Image registration error can also be avoided using subspace learning technique and it is demonstrated with geometrically enriched training samples created using transformation technique [25].

Based on the above discussion, it is observed that the computational cost for model fitting, training and feature dimension is high. Morphing technique can be used for

measuring the structural similarity for estimating the emotion without human intervention. As a result, we have proposed a geometrical model to estimate facial expression from the lower portion of the face (mouth).

3 The proposed approach

The proposed approach consists of three subsystems, namely, facial landmark localization, landmark tracking and classification. The first subsystem is mainly used for detecting face region and mouth and extracting low-level feature. Subsequent two subsystems are used to measure the deformation of facial components by tracking the variations in the feature point position in expressive image with respect to neutral image. The geometric transformation based feature vectors from previous stage are used for final classification of the expressions. Facial salient regions are isolated to recognize an expression, where the evolved expression forms a deformation over the face. The face is localized and detected for eliminating some unwanted portion or noisy pixel information from the input image like background, hair and neck. To reduce the searching time of the mouth, ¾ of the lower portion of the detected face is considered as Region of Interest (ROI) for the process. The low-level feature of the mouth is extracted to find four corner points and with these reference points more points are estimated. Initially, the effect of illumination is eliminated by stretching the contrast to highlight the mouth area of the remaining facial skin portion. The flood fill operation is performed on the image to bring the intensity values of dark areas that are surrounded by lighter areas to the same intensity level as surrounding pixels. After thresholding, Morphological open operation is performed to remove stray foreground (lip) pixels in background (other facial skin). Four corner points and a center point are identified from the minimum and maximum coordinate value (x, y) of the contour boundary. As a result, five points are obtained and elliptic model for lip is constructed with 16 points and it is described as follows, $P = \{p_0, p_1, \cdots, p_{15}\}$ and p_1, p_9, p_5, p_{13} and p_0 are the left, right, top, bottom and center point, respectively. Upper and lower lip ellipses are constructed using an elliptic parametric equation with linearly spaced 9 points generated between each pair of four corner points and all points on the ellipse edge are connected to the center point to calculate the angle difference during deformation. This elliptical model can be applied to any expressive mouth and small orientation changes can be handled by preprocessing the face image using primitive transformation to wrap that into the nearly frontal face. Mathematically, the following equations explains the parameters used in lip model construction.

$$\{p_1, p_9, p_5, p_{13}\}_{contour} = \{\{x_{min}, y), (x_{max}, y), (x, y_{min}), (x, y_m ax)\}_{contour} \quad (1)$$

$$Centerpoint\{p_0\} = \left(\frac{x_1 + x_9 + x_5 + x_{13}}{4}, \frac{y_1 + y_2 + y_5 + y_{13}}{4}\right) \quad (2)$$

And three points introduced between each pair of corner points are generated based on the following parametric equations

$$(X, Y) = (x_0 \quad y_0) + \left\{ \begin{bmatrix} a \\ b \end{bmatrix} \begin{bmatrix} cos\,\alpha & -sin\,\alpha \\ cos\,\alpha & sin\,\alpha \end{bmatrix} \begin{bmatrix} cos\,\beta & sin\,\beta \\ sin\,\beta & -cos\,\beta \end{bmatrix} \right\} \quad (3)$$

Where a is the semi major axis (common for upper and lower lip), b is the semi minor axis (different for upper and lower lip), α is the angle, which varies from 0 to 2π and β is the angle between x- axis and major the major axis ellipse. The following formulas are used to calculate semi major axis (a), semi minor axis for upper (b_{up}) and lower (b_{low}) lip.

$$a = \frac{x_1 - x_2}{2} \quad (4)$$

$$b_{up} = \frac{y_1 + y_2}{2} - Y_5 \quad (5)$$

$$b_{low} = Y_4 - \frac{y_1 + y_2}{2} \quad (6)$$

The sample lip model and the steps involved in feature point extraction process are shown in Fig. 1. The displacement of the feature points on the expressive image is measured. The neutral state of the mouth is considered as reference and the difference between points on the deformed face is calculated. Deformation of any shape can be viewed as the combination of various transformation functions in which each point in the shape, possibly change the size, orientation and position. Neutral state mouth points $f(x,y)$ are displaced and mapped on to the image of the same subject with expression. Displacement of points on the ellipse $g(x,y)$ is characterized by the composite transformation (translation, rotation and scaling) and quantitatively measured by finding the difference between $f(x,y)$ and $g(x,y)$.

$$G_j^k = (g(x,y)_j - f(x,y))^k \quad (7)$$

Where G_j^k is the feature vector of the k^{th} subject with j^{th} expression. This difference is expressed in terms of distance, angle, scaling and aspect ratio. In other words,

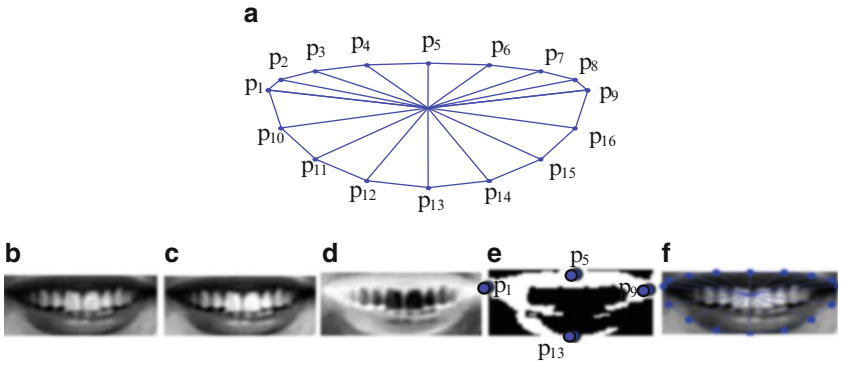


Fig. 1 Low- level Feature point extraction steps. (a). Lip model (b). Input image (c). Contrast stretched image (d). Complemented filled image (e). Contour with four corner points (f). Happy lip model for an input image

$$G_j^k = \left\{ \triangle d, \triangle \theta, \triangle (s_x, s_y), \triangle AR \right\} \tag{8}$$

As we know, the position of the points on both neutral (p) and expressive image(p'), its corresponding displacement in terms of distance (d) and rotation angle (θ) between pair of points ($p_{i,i+1}$), scaling factor in x and y direction (sx, sy)and aspect ratio(AR) is mathematically expressed as follows,

$$\triangle d = \{\|x_j^{'} - x\| + \|y_j^{'} - y\|\}^k \tag{9}$$

$$\triangle \theta = (\|\theta_{(p_{i,i+1}^{'})_j} - \theta_{p_{i,i+1}}\|)^k \tag{10}$$

$$\triangle (s_x, s_y) = \left\{ x_j^{'} - x_0^{'}/x - x_0, y_j^{'} - y_0^{'}/y - y_0 \right\}^k \tag{11}$$

$$AR = f_4(BB) = \left\{ (BB_{width}/BB_{heigth})_j - (BB_{width}/BB_{heigth}) \right\}^k \tag{12}$$

Where i is the number of points and x_0 and y_0 are center point of upper and lower lip (p_0). Rotation angle θ is the angle between the pair of points calculated between pairs of points with center point from the lip model. Aspect ratio (AR) is calculated as the ratio of withe widthnd height of the minimum Bounding Box (BB) used to cover the mouth contour. Collective normalized value of these features from Eq. (9), (10), (11) and (12) are used to construct the feature vector. The geometric transformation and related features obtained is now integrated into the feature matrix for a global representation of each facial expression. The machine learning algorithm plays an important role to improve the accuracy. Associative memory based dynamic recurrent neural network is used in this stage to recognize expressions from feature set. Associative memories can be implemented using networks with or without feedback. The function of an associative memory is to recognize based on previously learned input vectors, even in the case, where some noise has been added. The Hetero associative network is the kind of associative networks, which map m input vector $F = \{X_1, X_2, \ldots, X_m\}$ in n-dimensional space to m output vectors $Y = \{Y_1, Y_2, \ldots, Y_m\}$ in k-dimensional space. The goal of such a network is to identify the 'name' of the input pattern. In general, all inputs to a threshold neuron are combined into a single number, z, using the following weighted sum:

$$z_j = \sum_{i=1}^{m} W_{ji} x_i + \mu_j \tag{13}$$

Where w_i is the weight associated with i^{th} input (attribute) x_i, μ is the bias term and j is the number of nodes in the output layer. Using the mapping or activation function$f(.)$, output perceptron z_j produce the outputsy_j.

$$y_j = f(z_j) = \begin{cases} 1 & if z_j > 0 \\ 0 & if z \leq 0 \end{cases} \tag{14}$$

$$x_i(n) = \begin{bmatrix} y_j(n-1) \\ x_i(n) \end{bmatrix} \tag{15}$$

Using the feedback connection, it can remember previous stage training sample for subsequent mapping by grouping external input $x_i(n)$ with the delayed output$y_j(n-1)$, where n is the current stage. Using this combined input, perceptrons calculate the weight and bias using Eq. (14) and final mapping is carried out using Eq. (15). Preprocessed input vector contains the most frequently associated features, which explain the geometric deformation similarity of the people's particular expressions. Most dominant features identified from the previous stage are then given to the ARNN for training and testing purpose. With the help of associative memory based learning nature, it classifies the input vector with the knowledge learnt from previous few training samples.

4 Results and Discussion

To evaluate the proposed work, we use two publicly available datasets JAFFE database and Yale facial expression dataset. JAFFE consists of 10 subject's grayscale images with different emotional intensity of seven distinct expressions (happy, angry, disgust, fear, sad and surprise) including neutral. From Yale, we have used 11 male and one female gray scale static images with three basic expressions (happy, surprise and sad). From these two datasets, images with apex state expression and its corresponding neutral state images are considered in estimating the expressions. After detecting the mouth feature points, geometrical transformation difference between the neutral and expressive faces are determined using Eq. (9), (10), (11) and (12). All the geometry information is integrated and grouped together based on expression using Eq. (8). A decision based feature selection is performed to determine more contributing parameters so that search space and computational cost of the final prediction algorithm is greatly reduced and found that the following variables which is shown in Fig. 2. are important for estimating happy, surprise and sad where T is the threshold value selected on that column vector to increase the homogeneity of the target class

Collection of parameters in decision rules for which more than 70 % of records classified correctly is considered for final prediction. The expression such as disgust, fear and anger not only depends on the mouth region alone, since the geometric structural variations are not same for all people. Further, the distance between the nose and mouth and displacement of the nose are important factors for disgust expression. Likewise, for fear and anger expression, eye brow displacement and relative distance between mouth and eyebrow is essential. Geometric information of these (disgust, fear and anger) expression are scattered over the other three expressions (happy, surprise and sad) region. This is due to the fact that the mouth deformation for the above said expression is very small, compared to happy and surprise. Because of the low geometric difference and lack of features needed to discriminate those expressions, for experiments, we mainly focused on happy, sad and surprise only. The classification is carried out by randomly partitioning the data, 50 % for training and 50 % for testing. The proposed approach achieves 75.5 % and 88.2 % recognition rate for the local features for JAFFE and Yale data set, respectively. The recognition rate of surprise and sad expression in JAFFE dataset is low. Because, for sad expression, changes in feature point in lower face are very low compared to upper portion. Similarly, in Yale database, 20 % of happy sample is misclassified as surprise due to geometric similarity (lip apart). The result can be improved by considering the geometrical features of other facial components and increasing the training sample. The recognition rate has increased to 96.2 % in the same data set by considering the global features. The local and global feature recognition results individual expressions are shown in Table.1.

From the above result, it is observed that geometric transformation features improve the recognition rate of three expressions while using local features, the happy achieves 100 % accuracy and other emotions are considerably low. The surprise achieves good classification accuracy, moderate and good for sad and happy in the Yale data set. While using global features, for which the result is combined, the classification accuracy for surprise and sad is very good and appreciable for happy. This recognition rate can be further improved if we consider geometrical information from the remaining facial component with the support robust machine learning approach.

$$
\begin{aligned}
&if\ sy_2 > T_{sy_2} \& sx_6 > T_{sx_6} \& d_7 \le T_{d_7} then\ Happy\\
&if\ sy_2 > T_{sy_2} \& sx_6 \le T_{sx_6} \& t_1 > T_{t_1}\ then\ Surprise\\
&if\ sy_2 \le T_{sy_2} \& sy_3 > T_{sy_3} \& d_1 \le T_{d_1}\ then\ Sad
\end{aligned}
$$

Fig. 2 Classification rules for three expressions

5 Conclusion

In this paper, we propose a geometric transformation feature based approach to analyze some facial expressions of the lower face. The deformation to the mouth feature points due to the expression is represented in terms of translation, rotation, scaling and aspect ratio. The most contributing

Table 1 Confusion Matrix of JAFFE and Yale with local and global features

	Experiment I (local feature)						Experiment II (global feature)		
	JAFFE Predicted (%)			YALE Predicted (%)			JAFFE & YALE Predicted (%)		
Observed	Happy	Surprise	Sad	Happy	Surprise	Sad	Happy	Surprise	Sad
Happy	100	0	0	80	0	16.7	87.5	0	0
Surprise	0	75	50	20	100	0	12.5	100	0
Sad	0	25	50	0	0	83.3	0	0	100

factors are identified from the input data using the feature selection process to reduce the final prediction algorithm search space and computational time. Since the features from mouth alone is considered, we focus only three expressions such as happy, surprise and sad. The reduced feature set identified from JAFFE and Yale dataset effectiveness is tested using ARRN which gives a good recognition rate. This recognition model accuracy can be improved further if we consider the more geometric features to find the similarity exist among other facial component and by increasing training sample with more variations in each expression. Our future work will focus to consider all facial features to recognize more expression, intensity levels and propose suitable object modeling techniques for extracting expression from video sequences.

Acknowledgement This work is supported by a research grant from the Indo-US 21st century knowledge initiative programme under Grant F. No/94-5/2013 (IC) dated 19-08-2013.

References

1. Levenson, R.W.: Autonomic nervous system differences among emotions. Psychol. Sci. 3, 23 –27 (1992).
2. Witvliet, C.V.O., Vrana, S.R.: Psychophysiological responses as indices of affective dimensions. Psychophysiology, 32, 436 – 443 (1995)
3. Friedenberg, J., Silverman, G.: Cognitive Science: An introduction to the study of mind. SAGE Publications, London (2006)
4. Ekman, P., Friesen, W.V.: Facial Action Coding System Investigator's Guide, Consulting Psychologist Press, Palo Alto, CA, (1978)
5. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol., 17, 124–129 (1971)
6. Cootes, T., Taylor, C.J.: Active shape models—smart snakes. In: British machine Vision Conference, pp. 266– 275, Springer-Verlag (1992)
7. Soladie, C., Stoiber, N., Seguier, R.: A new invariant representation of facial expressions: Definition and application to blended expression recognition. In: 19th IEEE International Conference on Image Processing (ICIP), pp. 2617 – 2620, Orlando, FL (2012)
8. Calder, A.J., Burton, M.A., Miller, P., Young, A.W., Akamatsu, S.: A principle component analysis of facial expression. Vision Research, 49, 1179-1208 (2001)
9. Darwin, C.: The Expression of the Emotions in Man and Animals. J. Murray, London (1872)
10. Zhang, L., Tjondronegoro, D.: Facial Expression Recognition Using Facial Movement Features. IEEE Transactions On Affective Computing, 2 (4) (2011).
11. Gu, W., Xiang, C., Venkatesh, Y.V., Huang, D., Lin, H.: Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. Pattern Recognition 45, 80–9 (2012)
12. Xie, X., ManLam, K.: Facial expression recognition based on shape and texture. Pattern Recognition, 42, 1003 – 1011 (2009)
13. Kotsia, I., Pitas, I.: Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines. IEEE Transactions On Image Processing, 16 (1) (2007)
14. Kotsia, I., Buciu, I., Pitas, I.: An analysis of facial expression recognition under partial facial image occlusion. Image and Vision Computing, 26, 1052–1067 (2008).
15. Valstar, M.F., Pantic, M.: Fully Automatic Recognition of the Temporal Phases of Facial Actions. IEEE Transactions On Systems, Man, And Cybernetics—PART B: CYBERNETICS, 42 (1) (2012)
16. Zafeiriou, S., Pitas, I.: Discriminant Graph Structures for Facial Expression Recognition. IEEE Transactions On Multimedia, 10 (8), 1528-1540 (2008)
17. Song, M., Tao, D., Liu, Z., Li, X., Zhou, M.: Image Ratio Features for Facial Expression Recognition Application. IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, 40 (3), 779-788 (2010)
18. Li, Y., Zhao, Y., Ji, Q.: Simultaneous Facial Feature Tracking and Facial Expression Recognition. IEEE Transaction on Image Processing, 22 (7), 2559-2570 (2013)
19. Khana, R.A., Meyer, A., Konik, H., Bouakaz, S.: Framework for reliable, real time facial expression recognition for low resolution images. Pattern Recognition Letters 34, 1159–1168 (2013)
20. Tariq, U., Lin, K.H., Li, Z., Zhou, X., Wang, Z., Le, V., Huang, T. S., Lv, X., Han, T.X.: Recognizing Emotions From an Ensemble of Features. IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, 42 (4) (2012)
21. Geetha, A., Ramalingam, V., Palanivel, S., Palaniappan, B.: Facial expression recognition – A real time approach. Expert Systems with Applications 36, 303–308 (2009)
22. Ji, Y., Idrissi, K.: Automatic facial expression recognition based on spatiotemporal descriptors. Pattern Recognition Letters, 33, 1373–1380 (2012)
23. Danisman, T., Bilasco, I.M., Martinet, J., Djeraba, C.: Intelligent pixels of interest selection with application to facial expression recognition using multilayer perceptron. Signal Processing, 93 (6), 1547–1556 (2013)
24. Dornaika, F., Lazkano, E., Sierra, B.: Improving dynamic facial expression recognition with feature subset selection. Pattern Recognition Letters 32, 740–748 (2011)
25. Maronidis, A., Bolis, D., Tefas, A., Pitas, I.: Improving subspace learning for facial expression recognition using person dependent and geometrically enriched training sets. Neural Networks 24, 814–823 (2011)

Cognitive Based Sentence Level Emotion Estimation through Emotional Expressions

S.G Shaila and A. Vadivel

1 Introduction

Emotions have been widely studied in psychology and behavior sciences, as they are considered as an important element of human nature. It represents the psychological state of a person which is normally based on internal factors such as mental and physical status of a person and external factors say, social sensory feeling [15]. Identifying emotions from natural language texts has drawn the attention of several information processing communities since, it plays a vital role in human intelligence, decision making, social interaction, awareness, learning, creativity, etc,. Analysis of the emotional content in text, determines opinions, attitudes, evaluations and inclinations. This has focused on recognizing positive and negative orientation of a person with respect to various topics. Also, researchers have focused in the field of human computer interaction namely facial expressions studies, recognition of emotions using sensors, opinion mining and market analysis, etc. Recent online chat systems and blogs are considered as information repository of text with emotional contents. Future human-computer interaction is expected to emphasize the naturalness and effectiveness by integrating the models of human cognitive capabilities that includes emotional analysis and generation. Several efforts have been made by the natural language processing researchers to identify emotion at different level of granularities say word, sentence or document [5][1][2] using reviews, news, question answering, information retrieval, etc. A model is proposed in [9] to estimate the emotions in text by considering the relations among words in a sentence and uses symbolic clues as well as natural language processing techniques for word/phrase/sentence level analysis. In [8], both supervised and unsupervised machine learning classification techniques has been proposed on blog data for comparative evaluation. Here, Support Vector Machine(SVM) has been used to identify the intensity of the community mood. In [11], a corpus of short stories, which are manually annotated with sentiment tags has been used for automatic emotion based classification of sentences. The above literatures focus on the genre of fiction with only sentence-level emotion annotations and they do not identify emotion indicators within a sentence. In [6], an approach is proposed by considering semantics in the text to identify emotions at the sentence level using real-world knowledge from a commonsense knowledge base. The sentences that contain some emotional information are extracted from the knowledge base. Later, this information is utilized in building emotional models of text, which are used to label each sentence with a six-tuple that corresponds to Ekman's[4] six basic emotions. Identifying emotion understanding the importance of verbs and adjectives has been proposed in [14], which is topic and genre independent. Here, each post from a blog has been classified as objective, subjective-positive and subjective-negative. Yahoo! Kimo Blog has been used as corpora in [3] to build emotion lexicons. Emoticons were used to identify emotions associated with textual keywords. A system has been proposed for classifying news articles according to the reader's emotions [7]. Emotion classification task on web blog corpora using SVM and CRF machine learning techniques is carried out. It has been observed that the CRF classifiers outperform SVM classifiers in case of document level emotion detection. In [10], characterization of words and phrases according to their emotive tone has been described. The system classifies the reviews into two types, namely recommended and not recommended using the semantic orientation of the phrases in the review. However, in many domains of text, the values of the individual phrases may bear little relation to the overall sentiment expressed by the text. In [2], emotions are extracted based on WordNet Affect list and dependency relations using intensities. The SVM based supervised framework is employed by incorporating different word and context level features. In

S.G. Shaila (✉) • A. Vadivel
Multimedia Information Retrieval Group, Department of Computer Applications, National Institute of Technology, TamilNadu, India
e-mail: shaila@nitt.edu; vadi@nitt.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_100,

[1], emotion analysis on blog texts has been carried out on the English SemEval 2007 affect sensing corpus containing only news headlines. Conditional Random Field (CRF) based classifier has been applied for recognizing six basic emotion tags for different words of a sentence. A score based technique has been adopted to calculate and assign tag weights to each of the six emotion tags. Since, emotion is subjective entity and a sentence may have multiple emotions, classifying the sentence based on the mood is a hard task and above mentioned approaches in sentence classification achieve only modest performance in this domain. Most of the above discussed machine learning based models have considered sentence as their basic key constituent whereas our proposed approach deals with word and phrase in sentences for fine grained pattern analysis.

Based on the above discussion, it is observed that the words in sentences play an important role in tracing the emotions and to find the cues for generating such emotions. However, in many text domains, the phrases are given less weightage in the sentences. In our approach, like words, phrases are considered as the semantic units for emotional expressions and are used in identifying emotional patterns at sentence level. We mainly focus on the characteristics of Emotional triggered (*ET*) terms and the role of co-occurrences of *ET* term in the phrase for sentential emotion recognition and patterns that effectively contributes for positive and negative emotions in a sentence. Here, the proposed approach considers the POS features of *ET* terms and its co-occurrence terms. A supervised framework is employed for classifying the sentences into positive and negative emotional patterns. The proposed approach performs well and achieves encouraging results in obtaining emotional expressions, positive and negative emotion patterns on benchmark dataset. The rest of the paper is organized as follows. The proposed work is presented in the next section and the experimental results are presented in Section 3. The paper is concluded in the last section of the paper.

2 Proposed Work

Emotion analysis is considered as a pattern identification problem at sentence level. The main objective of the proposed approach is to identify the patterns of emotions with respect to positive and negative orientation at the sentence level. The sentences are constructed with large number of terms and only certain terms represent the emotions. These terms project the degree of emotional constituents along with other surrounding related hints and referred as emotional triggers (*ET*). We consider emotion triggers and Part Of Speech (POS) tags such as adverbs (*RB*), adjectives (*JJ*), verbs (*VB*), nouns (*NN*), intensifiers (*INTF*), negations (*NEG*), interjections (*IJ*) and conjunctions (*CJ*) as the baseline for our work.

2.1 Extracting Emotional Triggered Sentences

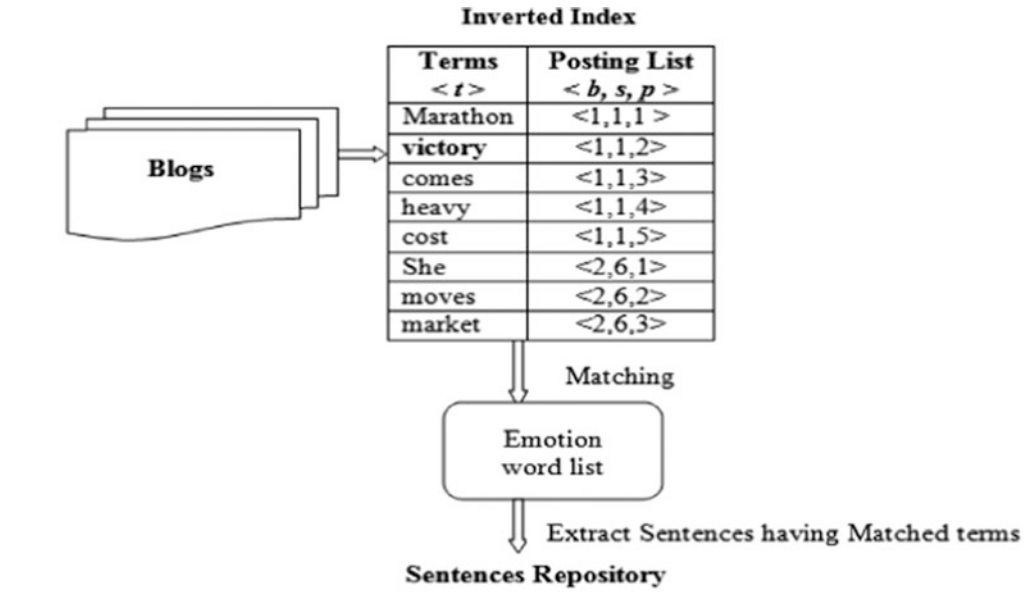
We have considered six basic emotions proposed by Ekman [4] in our work such as *happy, sad, anger, disgust, surprise* and *fear*. These emotions are also represented in the form of facial expressions. First, we have generated a list of seed words commonly used for six basic emotions. The emotion word lists like WordNet Affect lists[13] and word dictionary based thesaurus in English have been utilized as a resource for analysis. These lists are used to extract the emotion triggers (*ET*) present in the expressions that in turn contribute for identifying emotion patterns at sentence level. The dataset of emotional text shared task on news headlines at SemEval 2007 [12] is used for analysis. The corpus has news headlines that are extracted from news web sites such as Google news, CNN and other newspapers. The training dataset with 253 sentences are considered for standard preprocessing steps, which includes tokenizing, stemming and removal of stop words. The terms are stored in the inverted index and for each term $<t>$ in the inverted index, there is a posting list that contains sentence *id* and frequency of occurrence $<s, f>$. Let S be a set of sentences and T be a set of terms present in S. This may be treated as a labeling approach denoted as follows.

$$l : T \times S \rightarrow \{True, False\} \tag{1}$$

The inverted index consists of *ET* terms as well as other terms. From Eq. (1), it is assumed that a term $t \in T$ present in a sentence $s \in S$, if $l : (t, s) = True$. A sample posting list in retrieval applications is extracted from the inverted index and is in the form of $< t, b, s, p >$, where p is position of term t in the sentence s in blog b. Since, a term can be physically appearing in sentences of blogs, given *ET*, such that $ET \subseteq T$ and is defined as the relationship of $< ET, b, s, p >$ as follows and is represented in Eq. (2).

$$C^D(ET) = \{< ET, b, s, p > \mid b \in B, ET \subseteq T \text{ and } l : (ET, b) = True\} \tag{2}$$

Identifying an emotional sentence containing a single emotion is easier than identifying a sentence having mixed emotions. Hence, the positional information about terms is considered to identify emotions in long expressions. The terms in the inverted index are matched with emotional word list and the terms that match are considered. The corresponding sentences are extracted from the corpus and referred as *ET* sentence. Sentences which do not have emotional expressions are referred as neutral sentence. We consider only *ET* sentences and sentence repository is built, which have sentences that belongs to various emotions. The Fig. 1 shows inverted index and sentence



Ekman's Emotions	Emotional Triggered (*ET*) Sentence
Happiness	Marathon victory comes at heavy cost.
Surprise	Game on! London exhibition celebrates the history of video games.
Sadness	Sri Lanka clashes kill 129 troops.
Fear	Combative Bush Releases Parts of Terror Study.
Disgust	Sony Hates Europeans, Will Prevent the Importing of PS3.
Anger	Riot warning for France suburbs.

Figure. 1 Inverted index and sentence repository

Table 1 POS tagged sentences

POS tagged Sentences
Marathon/NN victory/NN comes/VBZ at/IN heavy/JJ cost/NN ./.
Sri/NNP Lanka/NNP clashes/NNS kill/VBP 129/CD troops/NNS ./.
Sony/NNP Hates/VBZ Europeans/NNS ./, Will/MD Prevent/VB the/DT Importing/NN of/IN PS3/NNP ./.

repository and the input is Web blogs. The content of inverted index and the output of the sentences with *ET* terms are depicted.

The *ET* sentences of repository are passed through the Stanford Parser (http://nlp.stanford.edu/software/tagger.shtml), a probabilistic lexicalized parser containing 45 different POS tags from the Pen Treebank tag set. It contains 36 POS tags and 12 other tags. Table. 1 presents the POS tagged sample sentences.

The size of sentence repository is huge and is difficult to understand the POS tagged sentences patterns. Each sentence pattern has different meaning based on the context with respect to emotions. The *ET* sentences are considered for sentence level classification based on their POS features as shown in Fig. 2.

In our approach, we have considered four significant POS tags such as adverbs, verbs, adjectives and nouns, which can hold *ET* terms in the sentences. These POS tags are considered as base POS tags and their extensions such as comparatives, superlatives, etc, are considered to belong to the base POS tags. For instance, adjective comparative (JJR), adjective superlative (JJS) belongs to their base POS tag adjective (JJ). Along with them, intensifiers, negations, interjections, conjunctions are also used. Using these Tags, the sentence classification is done in various levels and is depicted in Fig. 2. In the first level of classification, the POS feature of *ET* tokens are considered, which can be appeared as noun/verb/adjective/adverb in the sentences. In the second level, immediate co-occurrence terms of the *ET* tokens ($ET \pm 1$) are considered and verified for the presence of intensifiers (very, really, so etc) since, these terms adds force to the meaning of verbs, adjectives and adverbs by modifying exclusively. Intensifier enhances and gives additional emotional context to the word it modifies. (Eg. *It's* ***absolutely*** *amazing*). In the third level, co-occurrence terms of the *ET* tokens ($ET \pm n$) are considered for the presence of negations (not, neither, none, etc,). Negation words are specific words that express a negative idea to influence reader. Presence of negation type words greatly influence other associated words in the sentence (Eg. *I was happy so much that I could* ***not*** *control*). In the fourth level, co-occurrence terms of the *ET* tokens ($ET \pm n$) are considered for the presence of conjunctions (and, or, but, etc) in the sentence. Conjunction words joins the sentence parts and they can

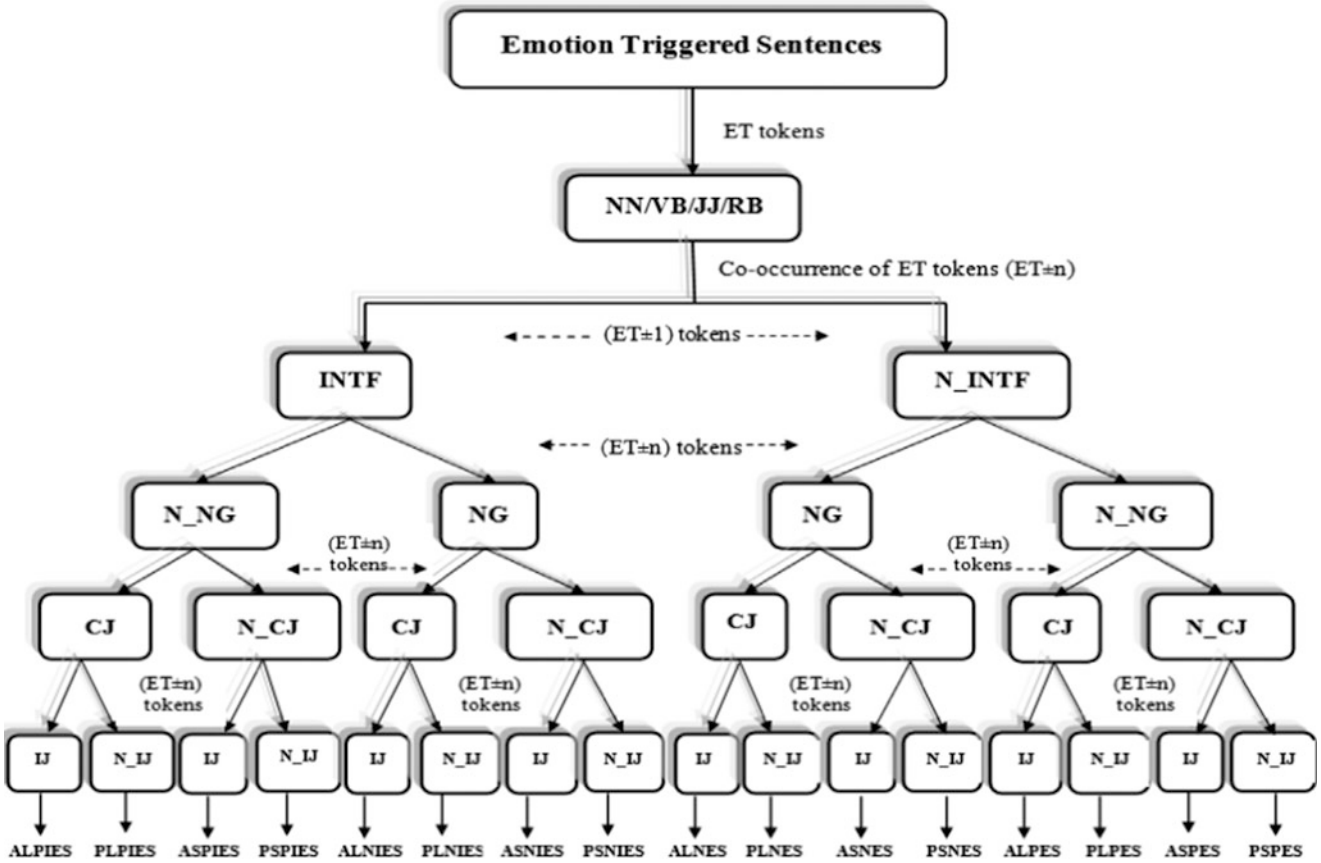


Figure. 2 Sentence classification based on emotion

appear anywhere in the sentence. In general, conjunctions are well used in the long expressions that have mixed emotions (Eg. *she was surprised,* ***but*** *not happy about the gift*). In the last level, interjections are considered for the classification. Interjections are exclamatory words that express emotions (Eg. ***wow****! Look at the sunset!*) These words can be placed before or after a sentence ($ET \pm n$) followed by exclamation mark or punctuation mark. Finally, the proposed Tag based approach hierarchically classifies the sentences into 16 classes and their name is represented in Table 2. In case of long expressions sentences having mixed emotions, we consider the phrase of first *ET* token of the sentence for classification.

2.2 Assigning degree of intensities to emotional patterns

It is also important to analyze the positive and negative emotions, which plays vital role in analyzing the psychology of a person. The positive and negative orientation in the expression occurs due to the effect of the emotions. The degrees of emotion contents in these patterns are captured by suitably categorizing them as classes. For Instance, the intensity of *ALNIES* class is *Negative Emotions Very High (N_EVH)*. The patterns of this class have long expressions (*CJ*) consisting of phrases that gives additional force (*INTF, IJ*) to emotions(*ET*) and exposes negative orientation(*NEG*) of the expression. The instance here, explains mixed

Table 2 Types of Sentence classes

Name of the Classes	Acronym
Active Long Positive Intensified Emotional Sentence	*ALPIES*
Passive Long Positive Intensified Emotional Sentence	*PLPIES*
Active Short Positive Intensified Emotional Sentence	*ASPIES*
Passive Short Positive Intensified Emotional Sentence	*PSPIES*
Active Long Negative Intensified Emotional Sentence	*ALNIES*
Passive Long Negative Intensified Emotional Sentence	*PLNIES*
Active Short Negative Intensified Emotional Sentence	*ASNIES*
Passive Short Negative Intensified Emotional Sentence	*PSNIES*
Active Long Negative Emotional Sentence	*ALNES*
Passive Long Negative Emotional Sentence	*PLNES*
Active Short Negative Emotional Sentence	*ASNES*
Passive Short Negative Emotional Sentence	*PSNES*
Active Long Positive Emotional Sentence	*ALPES*
Passive Long Positive Emotional Sentence	*PLPES*
Active Short Positive Emotional Sentence	*ASPES*
Passive Short Positive Emotional Sentence	*PSPES*

Table 3 Intensities distribution for negative and positive emotions patterns.

Classes	Negative(N)Emotions Patterns	Intensities
ALNIES	*ET + INTF + NEG + CJ + IJ*	*Negative Emotions Very High(N_EVH)*
PLNIES	*ET + INTF + NEG + CJ*	*Negative Emotions High (N_EH)*
ASNIES	*ET + INTF + NEG + IJ*	*Negative Emotions Medium (N_EHM)*
PSNIES	*ET + INTF + NEG*	*Negative Emotions Medium (N_EM)*
ALNES	*ET + NEG + CJ + IJ*	*Negative Emotions Low (N_EL)*
PLNES	*ET + NEG + CJ*	*Negative Emotions Very Low (N_EVL)*
ASNES	*ET + NEG + IJ*	*Negative Emotions Poor (N_EP)*
PSNES	*ET + NEG*	*Negative Emotions Very Poor (N_EVP)*
Classes	Positive(P)Emotions Patterns	Intensities
ALPIES	*ET + INTF + CJ + IJ*	*Positive Emotions Very High(P_EVH)*
PLPIES	*ET + INTF + CJ*	*Positive Emotions High (P_EH)*
ASPIES	*ET + INTF + IJ*	*Positive Emotions Medium (P_EHM)*
PSPIES	*ET + INTF*	*Positive Emotions Medium (P_EM)*
ALPES	*ET + CJ + IJ*	*Positive Emotions Low (P_EL)*
PLPES	*ET + CJ*	*Positive Emotions Very Low (P_EVL)*
ASPES	*ET + IJ*	*Positive Emotions Poor (P_EP)*
PSPES	*ET*	*Positive Emotions Very Poor (P_EVP)*

emotion which concludes the expression in negative emotion. Likewise, the intensities are interpreted for other categories and their patterns for negative and positive emotions and is shown in Table 3.

3 Experimental results

For experiments, SemEval dataset is considered for performance evaluation, since, it consists of both training and testing benchmark dataset. The details of the dataset is presented below in Table 4, where the number *ET* sentence is identified in the training dataset. The *ET* sentence contains both single and mixed emotions. The mixed emotions are sorted from single emotions by recognizing multiple *ET* tokens and conjunction POS tags.

We used manual annotation to judge the patterns into 16 classes which is a tedious and vital process. These classification annotations are done by a group of graduate and research students using NLP tool. Later, the training dataset is learned by Artificial Neural Network (ANN) by giving the POS features of *ET* sentence along with their intensities as input. The output of the ANN gives 16 classes of emotion patterns which contains various degrees of emotions in their phrases/expressions. The richness of the positive and negative emotion content are represented by the patterns. The outliers are obtained due to the presence of mixed emotions in the sentence, which conflicts and misleads the classifier during classification. We observed that difference in the classification of classifier and manual annotation is less, which lies in the range of 2-7 % for each pattern. The classification accuracy is calculated to estimate the performance of the classification based on the proposed approach using Eq. (3) and the results are given in Table 5. It is observed that classification accuracy of the classifier and human annotated classification match above 75 % for a sample dataset.

Table 4 Details of the SemEval dataset

Emotions	# *ET* Sentences
Happy	43
Surprise	8
Sad	50
Fear	20
Disgust	4
Anger	10
Mixed	15
Total	150

$$Classification\,Accuracy_n = \frac{Sentences\ correctly\ classified\ by\ the\ classifier\ to\ class\ type\ n}{Human\ annotated\ sentences\ of\ class\ type\ n} \quad (3)$$

Later, we used testing dataset to evaluate the performance of the proposed approach for measuring the positive and negative emotions. The 10-fold cross validation is used to evaluate the precision, recall and F1-measure for the patterns. The average results of positive and negative emotion classes is evaluated and compared with other approaches proposed by [1] and [2] and represented as shown in Table. 6.

Based on the results presented in both Table 5 and 6, it is observed that the proposed sentence level pattern model captures the emotional information effectively to analyze the persons psychology. The performance of the proposed approach is encouraging when compared with other similar approaches.

Table 5 Classification accuracy between human annotation and neural network

Manual Annotation		Neural Network		
Positive emotion classes	Sentence patterns (%)	Positive emotion classes	Sentence patterns (%)	Classification accuracy (%)
ALPIES	8.17	*ALPIES*	6.4	73.56
PLPIES	17.04	*PLPIES*	12.03	70.5
ASPIES	3.5	*ASPIES*	3.0	85.7
PSPIES	15.69	*PSPIES*	12.78	81.5
ALPES	13.45	*ALPES*	10.34	76.8
PLPES	19.73	*PLPES*	15.22	77.1
ASPES	7.17	*ASPES*	6.12	81.4
PSPES	14.34	*PSPES*	12.09	70.36
		Outliers	22.02	
Negative emotion classes	Sentence patterns(%)	Negative emotion classes	Sentence patterns (%)	Classification accuracy (%)
ALNIES	5.44	*ALNIES*	3.2	58.8
PLNIES	12.85	*PLNIES*	10.11	78.6
ASNIES	5.0	*ASNIES*	4.2	84.0
PSNIES	19.74	*PSNIES*	16.2	82.0
ALNES	10.07	*ALNES*	9.89	98.2
PLNES	20.29	*PLNES*	16.34	80.5
ASNES	4.01	*ASNES*	6.0	75.1
PSNES	22.21	*PSNES*	19.4	87.3
		Outliers	15.00	

Table 6 Performance evaluation(%) for SemEval dataset using 10-fold cross validation

	Positive emotions			Negative emotions		
Approaches	Prec	Recall	F1	Prec	Recall	F1
Proposed approach	82.76	81.09	81.91	85.97	82.36	83.47
Das and Bandyopadhyay (2009b)	61.00	62.9	61.94	68.44	52.68	58.93
Das and Bandyopadhyay (2010)	76.3	74.5	75.65	74.12	72.65	73.12

4 Conclusion

The proposed approach identifies patterns of emotions based on POS features of emotion triggered terms and its co-occurrence terms in the expression. The expressions are classified based on the POS patterns. The generated patterns of classification are analyzed and grouped into the positive emotions and negative emotions. Later, the intensities are assigned for capturing the degree of emotions that exist in semantic of expression. Further, neural network is used as machine learning tool to learn the patterns of positive and negative emotions which captures the psychology of a person. The performance of the proposed approach is encouraging when compared with other similar approaches.

Acknowledgement The work done is supported by research grant from the Indo-US 21st century knowledge initiative programme under Grant F. No/94-5/2013(IC) dated 19-08-2013.

References

1. Das, D., and Bandyopadhyay, S.: Word to Sentence Level Emotion Tagging for Bengali Blogs. In: ACL-IJCNL.pp. 149–152. Singapore(2009b)
2. Das, D., and Bandyopadhyay, S.: Sentence Level Emotion Tagging on Blog and News Corpora. J. Intelligent System. 19(2), 125–134 (2010)
3. Ekbal, A., and Bandyopadhyay., S.: Web-based Bengali News Corpus for Lexicon Development and POS Tagging. POLIBITS. 37, 20–29(2008).
4. Ekman, P.: An Argument for Basic Emotions. Cognition and Emotion. 6, 169–200 (1992).
5. Ku, L.,-W., Yu,-T., L., and Chen, H.,-H.: Opinion extraction, summarization and tracking in news and blog corpora. In: AAAI. pp. 100–107(2006)
6. Liu,H., Lieberman,H.,Selker, T.: A Model of Textual Affect Sensing using Real-World Knowledge. In: 8th international conference on intelligent user interfaces, ACM, (2003).
7. Lin, K., H,-Y., Yang, C., and Chen, H.,-H.: What Emotions do News Articles Trigger in Their Readers? In: SIGIR, pp.733–734 (2007).

8. Mishne, G., and Rijke, M., de.: Capturing Global Mood Levels using Blog Posts. In: AAAI, Symposium on Computational Approaches to Analyzing Weblogs. pp. 145-152(2006)
9. Neviarouskaya, A., Prendinger, H., and Ishizuka, M.: Narrowing the Social Gap among People Involved in Global Dialog: Automatic Emotion Detection in Blog Posts, In: Intl. Conf on Weblogs and Social Media, ICWSM.pp. 293-294(2007)
10. Peter, D., Turney.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: 40th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 417- 424(2002)
11. Read, J.: Recognizing affect in text using pointwise mutual information. Master's thesis, University of Sussex (2004)
12. Strapparava, C., and Mihalcea, R.: SemEval-2007 Task 14: Affective Text. In: 4th Intl. Workshop onSemantic Evaluations, ACL. pp. 70-74(2007)
13. Strapparava, C., and Alessandro, V.: WordNet-Affect: an affective extension of WordNet. In: 4th Intl. Conf on Language Resources and Evaluation, LREC, Lisbon, pp. 1083-1086(2004)
14. Vincent, B., Xu, L., Chesley, P., and Srhari, R., K.: Using verbs and adjectives to automatically classify blog sentiment. In: Symposium on Computational approaches to analyzing Weblogs, AAAI-CAAW. pp-27-29(2006).
15. Zhang, Y., Li, Z., Ren, F., and Kuroiwa, S.: A preliminary research of Chinese emotion classification model. IJCSNS, 8(11),127-132 (2008)

Hybrid Multilingual Key Terms Extraction System for Hindi and Punjabi Text

Vishal Gupta

1 Introduction

Key terms represent overall theme of any text document. By looking these key terms we can easily determine the nature of the document which is having these key terms. A number of statistical and linguistic based approaches are present for finding the key terms. This paper describes a hybrid approach for finding key terms from multilingual Hindi and Punjabi documents. The technique used is the hybrid of both the statistical and linguistic based features for finding the key terms from Hindi and Punjabi text. It is first time in history that this hybrid system has been proposed for Hindi and Punjabi languages. Regarding statistical features two features are used: 1) Entropy based metric [1] for finding relevant terms and 2) TermFreq-InverseLineFreq metric for finding key terms [2]. Regarding linguistic features 03 features are used i) treating nouns and verbs as key terms ii) treat terms belonging to lines in quotation marks [7], bold/italics/underlined fonts and with more font size as key terms iii) treating title words with maximum coverage [7] as key terms. For Hindi and Punjabi very less number of language resources are existing. This multi lingual key terms extraction system for Hindi and Punjabi will be very much helpful for developing other NLP resources in Hindi and Punjabi like: automatic text summarization, document classification, document clustering, question answering and topic tracking etc [9].

V. Gupta (✉)
University Institute of Engineering and Technology, Panjab University Chandigarh, Chandigarh, India
e-mail: vishal@pu.ac.in

2 Hybrid Multilingual Key Terms Extraction from Hindi and Punjabi

This system of key terms extraction is hybrid of statistical, linguistic and position based approaches for finding the key terms from Hindi and Punjabi text. Different types of features are explained below:

2.1 Statistical Feature for Key Terms Extraction

We have applied entropy metric [1] as statistical measure for extracting key terms from Hindi and Punjabi documents.

Let our multilingual Hindi and Punjabi text is having length L which can be calculated as number of terms in that text i.e. L Terms and segment this text into S sections. For each term of any type t, the measure of probability for a section $P_k(t)$ is given as below:

$P_k(t) = Freq_k(t) / \sum_m Freq_m(t)$ Where k and m and $= 1$ to S sections of text.

In the above equation $Freq_k(t)$ is considered as relative occurrence freq for a term t in k^{th} section of text.

For this distribution we can calculate entropy given by Shannon as follows:

$E(t) = S(w) = -1/\ln(S) \sum_k P_k(t) \ln(P_k(t))$

Moreover for any term with absolute frequency n for any random text can also be shown as: $E_{random} = 1 - (S - 1)/(2n \times \ln(S))$

For removing frequency dependency we can normalize the Entropy $E_{normalied}$ as follows: $E_{normalied} = (1 - E(t)) / (1 - E_{random}(t))$

If value of $E_{normalied}$ is greater for any term then we can say that that term is more relevant. The value of $E_{normalied}$ is very much dependent on frequency of sections of text in which we have to segment the source text. we have considered using natural segmentation of text like: paragraphs, sections and chapters etc. This feature is totally language independent as is treated as Feature1.

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_101,

The other statistical feature used is to use a metric TermFreq-InverseLineFreq i.e. TF-ILF metric [2, 3]. TermFreq is actual frequency of a term in a particular line and ILF is inverse line frequency which is frequency of lines containing the given term. We know that terms present in multiple lines in a single text document is not considered to be useful for the purpose of segmentation of topic. This metric is applied for finding relevance of any term in in a text document on the basis of its frequency of occurrence in a line & distribution of it over all lines in a given document. For calculating this feature first requirement is to delete all the stop words from Hindi and Punjabi text. Stop words are those terms which have very high frequency but are not the key terms like is, of, on, at, and, or for English language. We have created two separate lists of stop words for Hindi and Punjabi text. Before calculating the second feature all the occurrences of Hindi and Punjabi stop words are removed from source input text. Examples of Hindi stop words are: हैं, को, पर, इस, होता, कि, जो, कर. Examples of Punjabi stop words are: ਹੈ, ਦੇ, ਨਾਲ. The value of this second TF-ILF based feature is calculated as follows:

TF-ILF(t,l) = TF(t,l)* ILF(t)

Here TF(t,l) is frequency of occurrence of a term t in a line l.

Inverse line frequency ILF(t) = log(|L|/ LF(t)) here line frequency LF(t) is the frequency of lines containing term t.

Terms having higher value of TF-ILF are considered as more relevant. This feature is denoted as Feature2.

2.2 Linguistic Features for Key Terms Extraction

Before applying linguistic features, we have applied stemmers for Hindi and Punjabi for converting terms into their root form. Regarding Punjabi we have applied Punjabi stemmer as given by Gupta and Lehal (2011) [4]. For Hindi we have applied a stemmer given by Ramanathan & Rao (2003) [5].

Nouns and Verbs are always important and are considered as key terms in our system. This feature is a Boolean feature having values 0 or 01. If a term is noun or verb then flag for this feature will be set to 01 otherwise its value will be 0. For Punjabi, we are using Punjabi dictionary for finding if a term is noun or verb in Punjabi [3][8]. For Hindi we are using Hindi WordNet [6] available at IIT Bombay website for finding if a given term is noun or verb in Hindi. This feature is first linguistic feature and is represented as Feature3. Examples of Hindi Nouns and Verbs are: कार, दोस्त and लखिना etc. Examples of Punjabi Nouns and verbs are: ਕਾਰ, ਅੱਖ, ਚਲਾਂਦਾ and ਬੋਲਦੇ etc.

Second linguistic feature is calculated by considering terms which belong to lines containing quotation marks, having underlined/bold/ italics fonts or having more font size than the regular text. For this feature, we can set Boolean flag = True for those terms which belong to such lines which are containing quotation marks, having underlined/ bold/ italics fonts or having more font size than the regular text. This feature is denoted as Feature4.

Third linguistic feature is is calculated by considering terms which belong to title lines after excluding stop terms and should have maximum coverage [7] in different paragraphs. Title terms can be considered as key terms only if they are having maximum coverage in the whole text. It is calculated on the basis of presence of a title term in different locations of text. Different locations are determined based on number of paragraphs present in the text. The score for this feature is calculated as ratio of paragraphs frequency in which a given title word occurs to the total frequency of paragraphs lying in given text document. This feature is denoted as Feature5.

$$\text{Score}_{\text{For Each Term t}}\ (\text{Feature 5}) = \frac{\text{Frequency of Paragraphs Containing Term t}}{\text{Total frequency of paragraphs}}$$

2.3 Calculation of Final scores of Words

In this sub phase scores of different terms for five features discussed above are added using the equation:

Final_Score (Each Term t) = $\sum_f \text{Score}_t\ (\text{Feature}_f)$ Where f = 1 to 5 features.

Here each term is represented with t. Top scored n terms are treated as key terms where n is number of key terms for a particular input text.

2.4 Hybrid Algorithm for Multi lingual Key Terms Extraction from Hindi Punjabi

This hybrid algorithm starts by deleting all the occurrences of Hindi and Punjabi stop words from input text and then it applies stemming for Hindi and Punjabi words for converting them into root words. For each term t in the input text (after removing stop terms) follow steps I to V:

Step I: Calculate Score of Entropy feature ($E_{normalied}$ is calculated) for each term t in input text. This feature (i.e. Feature1) is calculated as given in Section 2.1. If value of $E_{normalied}$ is greater for any term then we can say that that term is more relevant and can be treated as key term.

Step II: Calculate Score of TF-ILF feature i.e. Feature2 for each term t. Term frequency is actual frequency of a term in a particular line and ILF is inverse line frequency which is frequency of lines containing the given term.

TF-ILF(t,l) = TF(t,l)* ILF(t)

Here TF(t,l) is frequency of occurrence of a term t in a line l.

Inverse line frequency ILF(t) = log(|L|/ LF(t))

Here line frequency LF(t) is the frequency of lines containing term t.

Step III: Set Boolean Score of terms in Hindi and Punjabi which are Nouns or Verbs. For finding Punjabi Nouns or Verbs Punjabi dictionary is used which is developed by Punjabi University Patiala. For finding Hindi Nouns or Verbs Hindi-Word-Net is consulted which is developed by IIT Bombay. This feature is the third feature i.e. Feature3.

Step IV: Set Boolean flag = True for each term which belongs to those lines which are containing quotation marks, or having underlined/bold/ italics fonts or having more font size than the regular text. This feature is denoted as Feature4.

Step V: Calculate score of fifth feature by taking ratio of paragraphs frequency in which a given title term occurs to the total frequency of paragraphs lying in given text document. This feature is denoted as Feature5.

$$\text{Score}_{\text{For Each Term t}}\ (\text{Feature 5}) = \frac{\text{Frequency of Paragraphs Containing Term t}}{\text{Total frequency of paragraphs}}$$

Step VI: Calculate final score of each term t by adding score values of five features.

Final_Score (Each Term t) = $\sum_f \text{Score}_t\ (\text{Feature}_f)$

Where f = 1 to 5 features. Here each term is represented with t. Top scored n terms are treated as key terms where n is number of key terms for a particular input text.

Algorithm Input:

घर को अच्छी तरह चलाने के लिए हर घर के कुछ असूल होते हैं। हर घर की अपनी मर्यादा होती है। अगर घर के सभी जीव अपनी अपनी मर्यादा में रह कर अपनी उम्र और रश्ते मुताबकि करेंगे तो ही कोई घर अच्छी तरह चल सकता है और एक आदर्श घर बन सकता है। बेशक हर घर दा आपणा आपणा तौर तरीका हुंदा है पर इह जरूरी नही कि इक घर दे असूल दूसरे घर विच वी उतने ही कामयाब नतीजे देण जितने पहले घर विच मिले हन कयोंकि घर विच मनुॅख वॅसदे हन अते मनुॅखी मन बहुत संवेदनशील हुंदा है। कसि इनसान दा कई वारी लॅख रुपऐ ते वी मन नही डोलदा पर दूसरे ही पल उही इनसान दे पैसे दी नगिणी जही चीज ते ही डोल जांदा है।

Algorithm Output (Assuming n = 10)

घर
असूल
मर्यादा
रश्ते
घर
तरीका
नतीजे
मनुॅख
डलदा
इनसान

As we can see input and Output 10 high scored key terms four in Hindi and six in Punjabi are extracted for n = 10. Although this output seems to be satisfactory but one problem is some common key terms in Hindi and Punjabi will appear two times for example here घर and ਘਰ are denoting to same key term 'House' but it will appear two times once for Hindi and once for Punjabi.

Table I Test data set size.

	Hindi documents	Punjabi documents
Number of sentences	490 sentences	532 sentences
Number of words	4537 words	6216 words
Total number of sentences	1022 sentences in Hindi and Punjabi	
Total number of words	10753 words in Hindi and Punjabi	

3 Results and Discussions

The algorithm for Hybrid Multilingual Key terms extraction has been tested on 100 multilingual Hindi and Punjabi documents which were taken from Popular Hindi and Punjabi Websites like: www.likhari.org, http://www.amarujala.com/, www.bhaskar.com/, www.ajitjalandhar.com/ and epaper.punjabitribuneonline.com etc.

The size of test data set is given in Table I.

The Precision and Recall for this hybrid multilingual key terms extraction system are calculated as follows:

$$\text{Precision} = \frac{\text{Number of correctly extracted key terms by our system}}{\text{Total frequency of extracted key terms by system}}$$

$$\text{Recall} = \frac{\text{Number of correctly extracted key terms by our system}}{\text{Total frequency of correct key terms which should be extracted}}$$

Based on dataset given in Table I. the overall Precision and Recall for this hybrid multilingual key terms extraction system are 90.78 % and 87.43 % respectively which are reasonably good. Feature wise values of Precision and Recall are given in Table II and are shown in Fig1.

We can analyze the results given in Table II and Fig.1 and can conclude that with increase in features from 1 to 05, values of Precision and Recall are also improving. For statistical features results are not good but with addition of linguistic features for Hindi and Punjabi the values of Precision and Recall are improving considerably well.

Table II Results of multi lingual hybrid key terms extraction for Hindi and Punjabi.

Features	Precision	Recall
Feature 1	71.87 %	70.65 %
Feature 1+ Feature 2	74.34 %	72.78 %
Features 1+ Feature 2 + Feature 3	85.98 %	83.34 %
Features 1+ Feature 2 + Feature 3 + Feature 4	88.45 %	85.21 %
Features 1+ Feature 2 + Feature 3 + Feature 4+ Feature 5	90.78 %	87.43 %

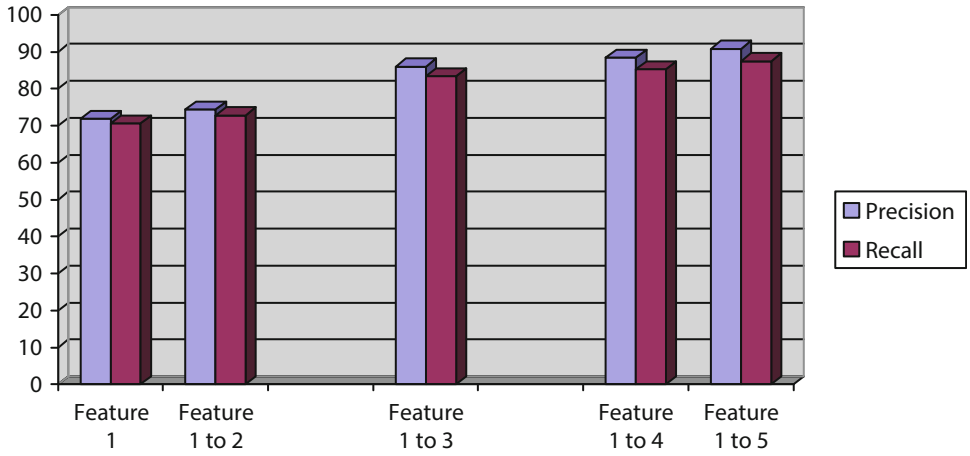


Fig. 1 Precision and recall for hybrid multilingual key terms extraction system

4 Conclusions

We can conclude that this hybrid multilingual key terms extraction system for Hindi and Punjabi is proposed first time. This system is hybrid of statistical and linguistic features of text in Hindi and Punjabi. Results of this system are improving with increase in features from 1 to 05. Statistical features alone are not performing well but addition of linguistic features has improved the precision and recall values considerably well. This system can further be improved by adding more statistical or linguistic features of text. This multi lingual key terms extraction system for Hindi and Punjabi will be very much helpful for developing other NLP resources in Hindi and Punjabi like: automatic text summarization, document classification, document clustering, question answering and topic tracking etc. The accuracy of this system in terms of precision and recall is reasonably good.

References

1. C. Carretero-Campos, P. Bernaola-Galván, A.V. Coronado, P. Carpena, "Improving statistical keyword detection in short texts: Entropic and clustering approaches", Elsevier's Physica A 392 (2013) 1481–1492.
2. Neto, J.L., et al.: Document Clustering and Text Summarization. In: Proceedings of 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining, London, pp. 41-55. (2000)
3. Gupta, V., Lehal, G.S.: Automatic Keywords Extraction for Punjabi Language. In: International Journal of Computer Science Issues, vol.8, pp. 327-331. (2011)
4. Vishal Gupta and Gurpreet Singh Lehal," Preprocessing Phase of Punjabi Language Text Summarization", International conference on Information Systems for Indian Languages Communications in Computer and Information Science ICISIL2011, Volume 139, Part 2, 250-253, Springer-Verlag Berlin Heidelberg, (March 2011).
5. Ramanathan, A., Rao, D.D.: A Lightweight Stemmer for Hindi. In Proceedings of Workshop on Computational Linguistics for South-Asian Languages, EACL (2003)
6. http://www.cfilt.iitb.ac.in/wordnet/
7. Carly W.Y. Wong, Robert W.P. Luk *, Edward K.S. Ho, "Discovering title-like terms", International Journal of Information Processing and Management, Elsevier, 41 (2005) 789–800.
8. Kaur, K., Gupta, V.: Keyword Extraction for Punjabi language. In: Indian Journal of Computer Science & Engineering (IJCSE), vol. 2, pp. 364-370. (2011)
9. Kaur, J., Gupta, V.: Effective Approaches For Extraction Of Keywords. In: International Journal of Computer Science Issues, vol. 7, pp. 144-148. (2010)

Sentiment and Emotion Prediction through Cognition: A Review

T. Vetriselvi and A. Vadivel

1 Introduction

Social media is a source of enormous amount of data with sufficient information about the product suggestions and opinions. This information is helpful for domain specific event prediction applications. Tweets or review write-ups are unstructured in nature and preprocessing is essential so that the information is understandable and useful. Emotions have been widely studied in psychology and behavior sciences, as they are considered as an important element of human nature. It represents the psychological state of a person, which is normally based on internal factors such as mental and physical status of a person and external factors say social sensory feeling. Identifying emotions from natural language texts has drawn the attention of several information processing communities since, it plays a vital role in human intelligence, decision making, social interaction, awareness, learning, creativity, etc. Analysis of the emotional content in text, determines opinions, attitudes, evaluations and inclinations. This has focused on recognizing positive and negative orientation of a person with respect to various topics. Also, researchers have focused in the field of human computer interaction, namely facial expression studies, recognition of emotions using sensors, opinion mining and market analysis, etc. Online chat systems and blogs are considered as an information repository of text with emotional contents. Future human-computer interaction is expected to emphasize the naturalness and effectiveness by integrating the models of human cognitive capabilities that includes emotional analysis and generation. In recent times, suitable classification algorithms are applied to predict the future event information and however these approaches may not predict correctly. This issue can be handled by using cognitive theory based approaches such as inductive inference, computational and intuitive theories, which helps to predict the event pattern more accurately. The inductive inference is involved in generalizing sparse data and can be applied along with Bayesian or Deep Belief Network (DBN) to enhance the performance of the existing sentiment prediction models. In this paper, we review some of the recently proposed approaches, which captures the sentiments from the cognition view point. The rest of the paper is organized as follows. The review is presented in the next section and we conclude the paper in the last section of the paper.

2 Review on Cognitive based sentiment

Automatic crime prediction approach has been proposed from Twitter posts using events extractor [5].The twitter data is used for automatic crime prediction by using NLP tool and it is observed that the predictive power is boosted by social media information and Symbol Role Labeling (SRL). The multiple events associated with a day is extracted and Latent Dirichet Allocation (LDA) technique is used to extract the topics from events. Also, it reduces the dimensionality of documents to lie within the K-dimensional space of topics. Generalized Linear regression Model (GLM) is used as prediction model. The performance of this approach is evaluated using hit-and-run crimes with the ground truth record. A crime modeling is proposed [14] and it uses prior criminal incidents for training. However, the predicted rate of this approach is poor while the past hitting is used. Similarly, Symbol Role Labeling (SRL) is a verb, noun, event based topic extraction approach for day to day events. However, the prediction rate and accuracy is not encouraging while the past histories are presented to the model. This model has failed to consider both Geographic information systems and Demographic information repository, which is mostly involved in the previous crime prediction models.

T. Vetriselvi • A. Vadivel (✉)
Cognitive Science Research Group Department of Computer Applications, National Institute of Technology, Tiruchirappalli, TamilNadu, India
e-mail: vetriselvi09@gmail.com; vadi@nitt.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_102,

A lexicon model for deep sentiment analysis and opinion mining applications [8] have designed a model, which classify the sentiment by verb, noun, adjective and attitude. People always interested to post their opinion, subjectivity or objectivity over a product/event in a social media web. However, each user differs in describing their opinion and expressions in text level. The Part of Speech(PoS) in NLP has been used for determining the opinion and later it is felt that this alone may not be sufficient to determine the opinion. As a result, a method is added to consider means of a word and relate it with others for determining sentiment. The semantic categories are identified as noun classes, verbs categorization and adjectives. It is well-known that the self-expression of a speaker/writer is subjective and it is essential to be captured it during the design. The subjectivity is handled in the form of lexicon categories such as speaker and character subjectivity. While the speaker subjectivity is handled toward others and character subjectivity is handled towards the individual. The subjectivity is identified in the lexicon.

The cornetto data set is used to combine two different semantic organizations such as the dutch word net and dutch reference. In lexicon model, nouns are categorized as animal, man, physical condition, a period of time, location etc., and verbs are categorized into types like action, experience, process state. These categories are used to evaluate the speaker/writer attitude from their general opinions. The accurate attitude is obtained by considering the composition and samples such as frequency, polysem, superset, size, measure, etc. However, the proposed lexicon model is suitable only for Dutch language and may fail to understand the comment in English language. Fuzzy deep belief network for semi-supervised sentiment classification [23] is a model for embedding prior knowledge into the learning structure**.** Web data is unlabeled and a semi-supervised learning model is proposed for handling it. DBN is one of the semi-supervised learning approaches and is used for sentiment classification. The DBN structure contains directed belief-net with hidden layers and follows a greedy as well as layer wise unsupervised learning. It is found that various documents are not precise and there exists ambiguity and vagueness. This issue is handled by capturing the vagueness using suitable fuzzy membership function and Fuzzy Deep Belief Network(FDBN) is developed. The FDBN has fuzzy classification ability with back propagation strategy based on fuzzy set. The DBN has initially trained with abundant unlabeled reviews and few labeled reviews. The FDBN architecture trained using all reviews and their corresponding membership value. This model fails to add multi -view distance metrics, which will improve the model performance.

Theory based Bayesian models of inductive learning and reasoning is a mode l ,which provide a frame work for learning and reasoning [1]. Inductive inference of a human makes them to understand what they visually see and perceive. Most of the children categorize table and chair by their talent and this is due to the fact that the generalized knowledge is extracted from a specific object. This is possible because the information is previously labeled by their parents. Properties of objects, their cause effect relations and domain knowledge, is to create the ability to generalize the sparse data. Another important phenomenon of learning is how the child understood the word meaning. An intuitive theory was thought with the set of casual laws, structural constraints for specific domain. Intuitive theory learns the domain specific knowledge and is used to generalize the sparse data. Abstract domain principles can come from more than one step. In the first step, the learner has to observe the world and create knowledge about unobserved data. Second one is structured probabilistic model with desired domain knowledge gives the correct prediction. few principles followed are taxonomy principle, contrast principle, competent and cooperative speaker with randomly sampled examples. This intuitive theory has been applied with Bayesian models for inductive learning and reasoning. This model describes the inductive learning and reasoning ability of human in computational terms. This focuses on inductive generalization and identification of the similarity. It finds association and correlation among the objects, with others based on domain knowledge and also analyzes the independent and statistical mechanisms of inference.

Combining social cognitive theories with linguistic feature for multigenre sentiment analysis [13] is proposed and the authors have used working principles of cognitive theories for sentiment classification on linguistic approach. The sentiment classification can be applied on any domain specific target. Each target supported by specific issue and sentiment analysis is categorized into target dependent and target independent. The preprocessing is done for normalizing the informal documents and each document is tokenized with PoS. The State-of-art English (SoE) entity extraction system is used to detect and extract target and target related issues. The Support Vector Machine (SVM) classification technique is used to design the supervised learning model, which detects the sentiment. The error analysis has shown that currently available sentiment lexicons and various linguistic features alone are not sufficient for sentiment classification. Social structures have its own impact on target and issue. The model is enhanced by combining three hypotheses based on social cognitive theory and incorporates these hypotheses into framework for propagating consistency across documents. Further, a hypothesis discussed based on social cognitive theories such as one sentiment per indicative target-issue pair (impression formation theory), one sentiment per indicative target-target pair(social categorization process) and is one sentiment per user-target issue during a short time(social

balance theory). These three cognitive theories valid for 90 % of instances and this model notified issues not a property of a single document. It depends on the labels for each document that mentions the target-issue pair. The baseline results are corrected by using three methods such as confident sentiment propagation, majority voting and weighted majority of voting. However, it is observed that the presence of sarcasm tend to misclassify and some Domain –Specific Latent Sentiments need deep mining. Also, this approach has failed to consider multiple sentiments and should enhance with global logic inference.

Based on the above review, it is observed that most of the approaches follow a similar sequence of tasks for emotion estimation. The information present in the social media is considered as input and requires a well-defined preprocessing approach for removing the noises. Some of the preprocessing activities done by replacement such as replacing emotions with sentiment polarity, negations with NOT tag and sequence of repeated characters by maximum three character [6].The large volume of data are reduced using instance reduction techniques namely filter methods and wrapper methods [25]. In general, PoS approach of NLP is used for tokenizing the word so that the input of this phase is amenable for a domain specific application. The pattern and associate pattern is identified for the application domain, which gives the degree of emotional content and suitable classifier is formulated for classifying the pattern. The classification approaches namely SVM, Bayesian classifiers and Artificial Neural Network considers the cognitive aspect for training and learning purpose for predicting the emotion pattern.

3 Concluding remarks

From the vast data available in social media, sentiment and emotion of the speaker or writer can be identified. Suitable preprocessing approaches are used for preprocessing the sentence. Most of the approaches have used machine learning approaches for training the patterns for predicting the emotional content. Based on our review, it is concluded that the sentiment and emotion classification with normal and existing classification algorithms may not provide effective result. The cognitive theories namely computational cognitive and intuitive theory can improve the sentiment and emotion prediction. We also suggest that in future the personal record of the writer/speaker can be integrated with compositional semantics for improving the sentiment and emotion prediction. Further, it is noticed that the language knowledge based on geographical location plays a major role for expressing the opinion of the writer and thus it can be considered as one of the parameters for predicting the sentiments.

Acknowledgement The work done is supported by research grant from the Indo-US 21st century knowledge initiative program under Grant F. No/94-5/2013(IC) dated 19-08-2013.

References

1. Joshua B. Tenenbaum, Thomas L. Griffiths, Charles Kemp. “Theory-based Bayesian models of inductive learning and reasoning” TRENDS in Cognitive Sciences, Vol.10 No.7 July 2006.
2. Megan A. Boudewyn, Debra L. Long, Tamara Y. Swaab. “Cognitive control influences the use of meaning relations during spoken sentence comprehension” Neuropsychologia, Volume 50, Issue 11, September 2012.
3. Michael J. Cole, JacekGwizdka, Chang Liu, Nicholas J. Belkin, Xiangmin Zhang. “Inferring user knowledge level from eye movement patterns” Information Processing &Management,Volume 49, Issue 5, Pages 1075-1091, September 2013
4. Jamie C. Gorman Peter W. Foltz Preston A. Kiekel Melanie J. Martin. “Evaluation of latent semantic analysis-based measures of team communications content” Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE JOURNALS, October 2003
5. Xiaofeng Wang, Matthew S. Gerber, and Donald E. Brown. “Automatic Crime Prediction using Events Extracted from Twitter Posts” 5th International Conference, SBP 2012, College Park, MD, USA, April 3-5, 2012. Proceedings
6. ApoorvAgarwalBoyiXie Ilia Vovsha Owen Rambow Rebecca Passonneau. “Sentiment Analysis of Twitter Data” LSM ’11 Proceedings of the Workshop on Languages in Social Media, Association for Computational Linguistics Stroudsburg, PA, USA ©2011
7. Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, Shrikanth Narayanan. “A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle” ACL ’12 Proceedings of the ACL 2012 System Demonstrations, Association for Computational Linguistics Stroudsburg, PA, USA ©2012
8. Isa Maks, PiekVossen. “A lexicon model for deep sentiment analysis and opinion mining applications”, Decision Support Systems, ELSEVIER, November 2012
9. Shaila S.G, et al, “Constructing Event Corpus from Inverted Index for Sentence Level Crime Event Detection and Classification” 3rd Joint International Semantic Technology (JIST) conference, November 28-30, 2013
10. Kenneth A. Norman, Sean M. Polyn, Greg J. Detre, James V. Haxby. “Beyond mind-reading: multi-voxel pattern analysis of fMRI data” TRENDS in Cognitive Sciences, 2006 Sep;10(9):424-30
11. Krauss Jonas, Nann, Stefan, Simon Daniel, Fischbach Kai. “Predicting movie success and academy awards through sentiment and social network analysis” ECIS, page 2026-2037. (2008)
12. Stefam Th. Gries, “Corpus-based methods and cognitive semantics: The many senses of to run”
13. Hao Li,2012, “Combining Social Cognitive Theories with Linguistic Features for Multi-genre Sentiment Analysis” Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, Nov 2012, Pages 127 -136
14. Wang X., Brown, D.E.: “TheSpatio-temporal generalized attitude model for criminal incidents” Security Informatics, vol. 1, 02/2012
15. VasileiosHatzivassiloglou, Kathleen R. McKeown. “Predicting the semantic orientation of adjectives” ACL ’98 Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, Pages 174-181

16. Peters BO, Pfurtscheller G, Flyvbjerg H. “Mining multi-channel EEG for its information content: an ANN-based method for a brain-computer interface” Neural Netw. 1998 Oct;11 (7–8):1429-1433
17. NikolausKriegeskorte, Rainer Goebel, Peter Bandettini , ”Information-based functional brain mapping” Proceedings of the The National Academy of Sciences of the USA, 2006
18. LA. Zadeh, “Fuzzy Sets”, Information and Control 8, 338-353, 1965
19. Vallabhaneni A[1], He B., “Motor imagery task classification for brain computer interface applications using spatiotemporal principle component analysis” Neurol Res. 2004 Apr;26(3):282-7.
20. Alison Gopnik and Andrew N. Meltzoff. “Words, Thoughts, and Theories”, MIT Press, 1997
21. Moshe Bar. “The proactive brain: Using analogies and associations to generate predictions” TRENDS in Cognitive Sciences Vol.11 No.7 , Jun 4, 2007
22. Chris Thornton. “Renewing the link between cognitive archeology and cognitive science” Journal of Archaeological Science, July 2012, Pages 2036–2041
23. Shusen Zhou, Qingcai Chen, Xiaolong Wang. “Fuzzy deep belief networks for semi-supervised sentiment classification” Neurocomputing, 5 May 2014, Pages 312–322
24. Jay Friedenberg, Gordon Silverman. “Cognitive Science An Introduction to the Study of Mind”, SAGE Publications, 2012
25. Pooja, Saroj Ratnoo "A Comparative Study of Instance Reduction Techniques" Proceedings of 2nd International Conference on Emerging Trends in Engineering and Management, ICETEM 2013

A Short Review for Mobile Applications of Sentiment Analysis on Various Domains

M. Sivakumar and U. Srinivasulu Reddy

1 Introduction

Sentiment analysis in recent times become a powerful complement to traditional business intelligence. Social networking gets the attention of business developers in various sectors such as marketing, health care, education and finance etc to improve their business from the people opinion shared on the different websites like facebook, twitter and other bolgs. Opinions available in the social networking sites either in the form structred or unstructured. The structuredness is found to be from questionaire and the unstructureness is from text box as free text. So it is necessary to process those unstructred text to detect what people trying to say about something by using effective machine learning techniques. Naive bayes classifier, support vector machines, rule based classifier and genetic algorithms are the best machine learning techniques which can be used to find the polarity of the reviews [7]. The polarity classification can be done at three different levels such as document level sentiment analysis, sentence level sentiment analysis and aspect level sentence analysis [19]. Based on requirement any one of these type of sentiment analysis can be used for polarity classification. Usually document level sentiment classification does not give much effect on finding the polarity of the people opinion. So either sentence level sentiment analysis or aspect level sentiment analysis can be used to identify the polarity of the opinons. Most of the work done so for on sentiment analysis at the sentence level and document level only. Sentence level sentiment analysis fails to infer congnitive and affective information associated with each sentence. So the concept level sentiment analysis overcome these issues [8]. Presently feature based sentiment analysis is also used mostly in product review, which gives more classification accuracy [13].

Sentiment analysis in the marketing field is used to analysis the product reviews, movie review contents and reviews about the services offered. It will help the business intelligence people to know the brand status of their product or services for further development. This technology is rapidly used in the field of education to help the students and to improve the infrastructure of the institutions by analyzing opinions from different people through online web portals. E-Health is becoming a growing area on various assistance to support patients and doctors using sentiment analysis techniques and tools effectively. This can also be started using in the election process to make a survey of opinionabout different candidates. So sentiment analysis techniques can be used effectively in all the fields. Our objective is to find the issues present in the different sentiment analysis applications and explore how mobile based applications can be created for those applications.

In this paper, we performed a preliminary study on how different types of sentiment analysis were used on various application areas and different machine learning techniques used on those type of sentiment analysis. Also we identified some sentiment analysis techniques which can be specially used on the education domain. The following section describes sentiment analysis used for diffrent domains.

2 Review on Applications of Sentiment Analysis

2.1 Sentiment Analysis for Product Review

Feature Based Sentiment Analysis on Customer Feedback [1] is an approach,where opinions are categorized based on product features. Here opinions are crawled from social networking sites, such as twitter, facebook and other blogs. The product features are collected either from manufactures or internet. The feature based opinion categorization is

M. Sivakumar (✉) • U.S. Reddy
Department of Computer Applications, National Institute of Technology, Tiruchirappalli, TamilNadu, India
e-mail: siva.mcs@gmail.com; usreddy@nitt.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_103,

achieved as a four step operations. Initially using WordNet [21] the synonyms for each features of a product is collected. Since customer may use different words to comment about featuresof the product. Then with the help of POS tagging, the words of opinions are tagged as either noun or adverb or adjective etc. In this approach only nouns are assumed to be the features of a product. So it shows the explicit features of a given product. Then the words tagged as noun are compared with the features collected from manufactures. Now, nouns in the sentences, matches with features of a product collected from manufactures, are clustered as bag of related sentences. From these bags of related sentences the semantic score of adjectives and adverbs are calculated using SenticWordNet [21]. The semantic scores are then used to identify the semantic orientation of each sentence. Sentences having greater than zero semantic scores are identified as positive sentiment orientation and rest of the sentences are identified as negative sentiment orientation. In this work the accuracy is calculated by reading all reviews manually from output file. This becomes a time consuming process of analyzing reviews. This work fails to extract implicit features of the products from the opinion of customer. Our focus is to explore this work on mobile application by elimination the issues found on it.

2.2 Sentiment Analysis for E-Health

Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality [2] is a used to measuring health of a patient.A tool called Patient Reported Outcome Measures (PROMs) was proposed to measure the patient Health Related Quality of Life (HRQoL). This model helps the patients to evaluate their health status through both structured (fixed questionnaire) and unstructured (free text) way and also it reduces the gap between the structuredness and unstructuredness of the patient data. It produces the semantics and sentics from the patient opinionated data as output. A set of topics and the polarity associated with it is extracted from the patient data at semantic level. Hence the patient opinions are fed as input to this tool and semantics and sentics as outputs. This patient opinion mining engine is a combination of four components such as a preprocessing module, a semantic parser, the ConceptNet module and the AffectiveSpace module. The affective valence indicators present in the opinion text are interpreted by preprocessing module. Then these texts are deconstructed into concepts by semantic parser using a lexicon and create a Small Bag of Concepts (SBoC). These bags of concepts are analyzed by ConceptNet and AffectSpace tool. The ConceptNet produces semantics from each SBoC by projecting the concepts on the matrix. This matrix helps to calculate the semantic relatedness of each concept and the degree of different class. On the other hand the AffectiveSpace module produces sentics using dimensionality reduction techniques. Finally cumulative polarity of both semantics and sentics are calculated in an effective manner. This application is lacking on conducting on-field usability tests for different case-mixes and detection of spam patient reviews. We are planning to enhance this model using Principal Component Analysis (PCA) technique [22] with mobile implementation to help patients through mobile phone by the practitioners.

A Hybrid System for Online Detection of Emotional Distress [11] is a model to detect the distress through online. This model helps to identify depressed people to provide help and follow-up services from the emotions expressed in the social networking sites and other web media. Initially the emotional text contents are retrieved through some blog search engines using meta-search approach [20]. These contents are made available for further analysis with automated process. Now the contents are classified using Hand-crafted model and supervised machine learning technique called support vector machine. These classifiers generate prediction scores for each blog contents. The final score is calculated with aggregation of the scores of both classifiers. The classified data published in blog, which is reviewed by clinical psychologists. In this depression detection system some standard evaluationmeasures such as precision, recall and F-measures to evaluate the performance of affect analysis techniques.It is far way from detecting more sensitive words related to depression of a user from social networking sites. This system can be implemented as a mobile based model to help the depressed users by sending some energetic tips.

2.3 Sentiment Analysis for Education

In [3] a framework was proposed to collect the exact feedback from students.The authors discussed about different existing approaches for collecting student feedback about a lecturer using handheld devices such as clickers and mobile phones. Usually it is found that using these devices the teaching methods can be improved from student feedback. Clicker is a real time interactive technology for education to asses teaching and learning process and to solve teaching problems. In this technology the students can response to the different multiple choice questions posted on a screen using a remote transmitter which contains buttons for ‘yes’ and ‘no’ options. These results are collected instantly and tabulated for instructor to act accordingly. Another method is the use of mobile phones for getting feedback either through mobile applications or SMS. Though these methods help the instructors to know the effectiveness of a session

from the student feedback, there is a possibility for misusing the mobile phones in the classroom and breaking the transmitter. So everyone is looking for better solution which helps the instructor to get feedback of a session without disturbing the current session.

So the authors suggested that making use of social media for collecting feedback about the session from students could be a better solution that will not disturb the present teaching session. Since the social networking sites like twitter, facebook and other micro blogging sites are used as a tool for getting opinion about different products, movies, treatment and other interesting areas to make some decision for the further improvement. People share their opinions about different thing either if they get benefitted or get affected to make other to be aware of those things. So in this paper the authors proposed a new architecture for getting student feedback using sentiment analysis techniques. In this the social networking sites are used as an interface for students to give feedback about each session on a particular time slot. Then using sentiment analysis techniques such as naïve bayes, maximum entropy and support vector machine the extracted free text from those sites are analyzed and results are sent to instructor to make further decision. So the instructor act according to those feedbacks and the session can be improved in a much better way. Thus the authors concluded that the sentiment analysis system can be used in the education sector for improving teaching learning methodologies. Student feedback posted through online portal sometimes in vagueness form,so fuzzy logic techniques can be used to handle the vagueness content. Our aim is to use aspect based sentiment analysis method to analyze the emotional contents of the students and alerting lectures via their mobile phones.

In [4] an Opinion Mining Framework was designed to collect opinion about Malaysian university and it analyze the online opinionsabout various Malaysian universities. The objective of this framework is achieved by three processes such as extraction, text processing and polarity classification. The extraction process collects opinions in an unstructured text format from various online sources. In includes a set of operations such as preprocessing and tagging. Using web crawling, HTML (Hyper Text Markup Language) parsing and Part Of Speech (POS) tagging the opinionated sentences are extracted from top ranked URL's (Unified Resource Locator). These sentences are segmented using regular expression and tokenized for tagging process. Here the Brown Corpus was chosen to tag the words in the tokenization process. Then stemming and stop word removal process is performed on those words using Named Entity Recognition (NER). Thus a corpus is constructed with words relevant to education domain. So these entities are parsed manually to SentiWordNet tool to find polarity. Thus polarities of various Malaysian universities are calculated and shown in graphical pattern.This framework covers only specific geographic area, so it can be extended to all kinds of institutions available in the world.

The Implementation of Social Networking as a Tool for Improving Student Participation in the Classroom [13] is a model to create an effective classroom. Though the effectiveness of the classroom teaching can be improved in different ways, it is not possible to use feedback posted on the social networking site instantly while taking classes. The idea behind this work is to use the feedback of the students posted on the social networking sites can be viewed by the lecturers on their current teaching slides without interrupting the session. So this will enable the students to shout their real time questions to their lectures during the session. Therefore student hesitate to post the question can post their question through online social networking sites using this approach. When the queries posted by the students get increased, the lecturer will get interrupted. So time slots could be allotted for posting queries to lecturer. The anonymous feedback become can be detected and sentiment analysis techniques can be used to analysis the feedbacks automatically.

2.4 Sentiment Analysis on Mobile

Mobile Sentiment Analysis [5] proposed to evaluate the sentiment of a person through mobile applications. In this approach the sentiment analysis is performed on the mobile phones locally. Here SMS messages and social media review content seen on the mobile phones are used as the input to the system. Hence this system calculates how much positive or negative messages read on a mobile phone using sentiment analysis techniques. Here the sentiment analysis is done at sentence level. English language is considered for SMS messages to be analyzed. The sentence level sentiment analysis consists of two major tasks such as subjectivity classification and sentiment classification. The assumption was made that the sentences contains an opinion of single opinion holder. To perform these tasks of classification the naïve base classifier, regression model and support vector machines are used.

In this work the sentiment analysis was performed using the principles of the SentiCorr sentiment analysis engine. Adaptive boosting algorithm was used to perform subjectivity classification. Rule Based Estimation Model (RBEM) was used for polarity classification in which eight different rules were emerged from eight different pattern groups. This model collects patterns and its associated rule is applied that match with sentiment in the messages. After collecting the match patterns from messages those eight rules are applied in a correct order. Rule1 Setting Stops sets stops on all the left flips and stop patterns. Rule2 Removing Stops removes the stop in the continuator pattern. Rule3 positive sentiment emission calculates an emission value for each positive pattern. Rule5 amplifies sentiment amplifies the sentiment

emitted by positive or negative sentence. Rule6 Attenuating Sentiment reverses the Rule5 operations. Rule7 and Rule8 for right flipping and left flipping sentiment to ignore the patterns contains right flip and left flip. Finally the polarity of the message is calculated by adding the emission values of each element. Sentence having greater than zero polarity value become positive sentence and less than zero polarity value become negative sentence. The rest of the sentence is categorized as unknown sentences.

3 Conclusion

In this paper, preliminary study of various sentiment analysis applications and its techniques was performed. We also found that there is a scope for using sentiment analysis techniques in a much better way in the field of education to help students, faculties and management to resolve the issues through feedback posted on the web portal. Feature based sentiment analysis can be a better option to categorize sentiment contents posted by students, employees and public about a specific institution. This will help the management of those institutions to act accordingly to take the institution in a right path. It is also possible to develop a mobile based application which incorporates the discussed techniques. It will become easiest way for the student, employees and management to share their emotions and get the remedial actions instantly. Finally we conclude that feature based sentiment analysis will be a better option to improve the students quality in education domain. In the future work we are planning to apply these techniques in various domains with cloud computing implementation. And also there is a scope for improving accuracy using computational intelligence techniques.

References

1. Avani Jadeja, Indr Jeet Rajput "Feature Based Sentiment Analysis on Customer Feedback" Journal of Information, Knowledge And Research in Computer Engineering, Nov 09 to Oct 10, Volume – 01, Issue – 01
2. Altrabsheh, Nabeela, Gaber, M. and Cocea, Mihaela SA-E: Sentiment Analysis for Education, 5th KES International Conference on Intelligent Decision Technologies, 2013-06-26 - 2013-06-28, Sesimbra
3. A. M. H. Elyasir, K. S. M. Anbananthen "Opinion Mining Framework in the Education Domain" International Journal of Social, Human Science and Engineering Vol:7 No:4, 2013
4. Lorraine Chambers, Erik Tromp, Mykola Pechenizkiy, Mohamed Medhat Gaber "Mobile Sentiment Analysis" Proceedings of the 16th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, 2012-09-10 - 2012-09-12, San Sebastian
5. Erik Cambria, Tim Benson, Chris Eckl, Amir Hussain "Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality" Expert Systems with Applications 39 (2012) 10533–10543
6. Novak Jeremy P., Cowling, Michael A. "The Implementation of Social Networking as a Tool for Improving Student Participation in the Classroom" ISANA International Academy Association Conference in Hobart, Tasmania, from 29 November to 2 December 2011 at The Wrest Point Conference Centre
7. Rudy Prabowo, Mike Thelwall, "Sentiment Analysis: A Combined Approach" Journal of Informetrics, Volume 3, Issue 2, April 2009, Pages 143–157
8. Erik Cambria, "An Introduction to Concept-Level Sentiment Analysis" 12th Mexican International Conference on Artificial Intelligence, MICAI 2013, Mexico City, Mexico, November 24-30, 2013
9. Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, Meeyoung Cha, "Comparing and Combining Sentiment Analysis Methods" COSN '13 Proceedings of the first ACM conference on Online social networks
10. Minsu Park, Chiyoung Cha, Meeyoung Cha "Depressive moods of user portrayed in Twitter" HI-KDD '12, August 12, 2012, Beijing, China
11. Tim M.H. Li, Michael Chau, Paul W.C. Wong, Paul S.F. Yip "A hybrid system for online detection of emotional distress" Pacific Asia Workshop, PAISI 2012, Kuala Lumpur, Malaysia, May 29, 2012
12. Kenneth Forbus, Jeffrey Usher, Andrew Lovett, Kate Lockwood, and Jon Wetze "CogSketch: Open-domain sketch understanding for cognitive science research and for education" International Conference, Spatial Cognition 2010, Mt. Hood/Portland, OR, USA, August 15-19, 2010
13. Novak, Jeremy P, Cowling, Michael A "The implementation of Social Networking as a Tool for improving student participation in the classroom" 22nd ISANA International Education Association Conference Proceedings, held at Wrest Point Conference Centre, Hobart, 29 Nov-2nd Dec 2011
14. Adnan Duric, Fei Song, "Feature Selection for Sentiment analysis based on Content and Syntax models" Decision Support Systems, Volume 53, Issue 4, November 2012, Pages 704–711
15. Subhabrata Mukherjee, Pushpak Bhattacharyya, "Feature specific sentiment analysis for product reviews" 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012
16. Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, Shrikanth Narayanan, "A system for real-time twitter sentiment analysis of 2012 U.S. Presidential Election Cycle" ACL '12 Proceedings of the ACL 2012 System Demonstrations Pages 115-120
17. Minsu Park, David W. McDonald, Meeyoung Cha, "Perception difference between the depressed and non-depressed users in twitter" 7th International AAAI Conference on Weblogs and Social Media (ICWSM), Boston 2013
18. Godfrey Winster Sathianesan and Swamynathan Sankaranarayanan, "Sentiment Analysis and Opinion Mining: A Survey" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012
19. Mohammad Sadegh, Roliana Ibrahim, Zulaiha Ali Othman, "Opinion Mining and Sentiment Analysis: A Survey" International Journal of Computers & Technology, Volume 2 No. 3, June, 2012
20. Michael Chau, Michael Chau, Ivy Chan, Hsinchun Chen, "Redips: Backlink Search and Analysis on the web for business intelligence analysis" Journal of the American Society for Information Science and Technolog, Volume 58, Issue 3, 1 February 2007
21. Hussam Hamdan, Frederic Béchet, Patrice Bellot, "Experiments with DBpedia, WordNet and SentiWordNet as resource for sentiment analysis in micro-blogging" Joint Conference on Lexical and Computational Semantics, 2013
22. Jeevanandam Jotheeswaran, Loganathan R., Madhu Sudhanan B., "Feature Reduction using Principal Component Analysis for Opinion Mining" International Journal of Computer Science and Telecommunications, Volume 3, Issue 5, May 2012

Human Cognition and Vision Based Earlier Path Determination System for Indoor Mobile Robot Path Planning

N. Nithya and D. Tamil Selvi

1 Introduction

Vision is the powerful sensing method for autonomous mobile robot navigation. Computer vision is a process that mimicking the human perception system and helps to acquire the knowledge of the environment and taking decisions autonomously. Mobile Robot Navigation had three important skills such as localization, path planning and map building. Vision helps these three fundamental tasks. For example, Autonomous vehicle localization by object monitoring using scale invariant feature transform method and trajectory deviation errors were recovered using 3D space transformation and calibration line on the detected object[1]. Global navigation done by detecting or locating natural landmarks like doors, walls, and floor for indoor environment [2]. Detect static landmarks and moving human by single camera tracking and recognition system [3]. Cognition allows an autonomous mobile robot to get increased autonomy in matters of learning, knowledge about its environment. Cognitive science processes are explained in terms of functionalities which inside people's head like to perceive, store, recall, taking decisions. An important aspect of cognitive based human like perception and decision making of mobile robots being able to safely move in it is environment like the human brain and its thought process.

N. Nithya (M.E) (✉)
Department of Computer Science and Engineering, Thiagarajar college of Engineering, Madurai-15, India
e-mail: er.nithyacse@gmail.com

D. Tamil Selvi
Department of Information Technology, Thiagarajar college of Engineering, Madurai-15, India
e-mail: dtamilselvi@tce.edu

1.1 Cognitive Perception

Perception or sensing an environment is a fundamental state of cognition. Computer vision and its algorithms give robust sensing capabilities to robot needs to perceive the world similar to human. Object detection is used to detect and estimate the location of landmarks [2], obstacles, and goal points [4] in an image and video frames without prior knowledge of its location information. However, in a perception there is typically sensory degradation or lack of perceptual cues affects the cognition. The visual systems are suffer from low image resolution, poor lighting conditions, pose deformation, occlusion, and scale variation. Occlusion is one of the major issues in an indoor clutter environment [5] for navigable space extraction and object localization in vision based path planning. For example, in automated video surveillance system human object partially occludes each other in crowd scenes [6]. Detection of partially occluded doors in landmark based navigation using data-driven Markov chain Monte Carlo(DDMCMC) [2]. Usually multiple objects are situated along the navigation path, in the indoor environment. Due to view point variation of robots, the required goal point may be partially occluded by other objects. In this way vision sensor is more compatible in robot navigation compare than other sensors like laser range finders and ultrasonic sensors. Shape feature is often sufficient cue for object detection. Many techniques have been developed for shape feature based object detection [7–11] is to identify and locate the target object in the environment image. Contour-based methods are more simple and effective for object detection and they work well in partial occlusion. The occluded parts or missing parts of an object, which result in the changes of object shapes, and it greatly decreases the true positive rate of the detection algorithm. Numerous techniques have been proposed for detection of occlusion in shape based and template based detection methods. Shape reconstruction is an efficient methodology to recover the missed portions or edge curves of object shape in the edge map. It computes which contour

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_104,

part fills the missing portion of an object shape. This inspiration comes from the reconstruction ability of human vision, which is quickly recovered a whole shape when it is partially occluded with the help of prior knowledge object shape. Kimia et al [12] proposed Euler spiral method use a variety of french style curve portions to fill the gaps in object boundaries using the minimum length in the tangent bundle as model but it is suitable for simple and smooth contours. Mumford [13] proposed Elastica based curvature inpainting method uses Bayesian based extension of curves and it is not optimal in rich local texture shapes. The above two methods not suitable for severe occlusion, but it can fill the gap continuity in object contours. X.Ron et al [14] proposed a Probabilistic model for contour completion using local geometric based curve for different scale structure objects but it fails to support for large occlusion. Venkatesh et al [16] states Symmetric shape completion method gives efficient results in handling a large portion of occlusion of shape. Kanizsa et al [16] used Amodal completion method performs stratification of the shrinkage effect of the object shape. Reinsink et al [17] used Early completion method fills the occluded edges by linking of associated regions of shapes. To reconstruct missing object contours with the help of common object shape templates to fill the missed portion of the object shape and getting original complete object shape with higher shape similarity.

1.2 Cognitive Based Robot Mapping

Map building is an important task in mobile robot navigation. The map is a data structure (or) geometric representation of the robot moving environment. It must contain information about the surface area of environment, efficiently locate object entities present in the real world, and facilitate finding shortest path and extents the map when the robot move one place to another place. The maps are classified into two types [19], metric map and topological map. The grid based metric map is commonly used for 2D representation of the environment. Topological map represents the relationship between the two places. In addition to that one effective map is called cognitive map [19, 20, 22] which gives ability to the robots to take decisions like human. Edward Tholman [19] introduces the cognitive maps alias mental maps. It serves the robots to acquire, store and recall information about spatial environment. In the example, classifying and recognition of different places in the office environment by searching objects which are uniquely present in the environment, and to update the map information with revisited places [20, 25]. Construct global localization hybrid map with objects and its spatial layout of the entire office environment [21] using stereo vision camera. It contains direction information which is angled deviation value to reach from one room to other rooms. Haptic virtual environment map with orientation changing information to navigate known and unknown environment for blind person's guidance system [23]. Compact internal representation(CIR) create an effective pathway for time evolving navigation environment using animal cognition method [24]. The proposed system supports to estimate decision making factors for vision based mobile robot path planning in an indoor clutter environment before start its navigation like an office, kitchen, and manufacturing plant. The earlier determination of travelling path system works like the human eye perception system and interpret with its environment in the cognitive map to identify the goal amidst the occluded obstacles along the way.

2 Earlier Path Determination System

The proposed earlier path determination system gives control to the robot to locate partially occluded target object in the environment. It estimates the distance away from its initial location and orientation value such as an angle deviated from the straight line focus. These values are used for direction changes and how long it will move forward for mobile robot path planning. This system imitates the sensing characteristics of the human eye and analyzing characteristics of human brain before start its walk along the path.

2.1 Goal Object Detection with Partial Occlusion

The Earlier path determination system detects the target object in the environment. Goal object search has found the location of the target object that present in the environment image if it may occlude by other objects. This system is based on one of the template matching method called Optimal Curve Segment (OCS) method used to detect objects and fix the location in the perceived environment image. Due to partial occlusion of object shape reduces the detection performance in the most of the shape matching methods. In Fig. 1 Shows occlusion arises in the view of mobile robots. The objects located at different positions in the environment. In robot view stool is partially occluded by the yellow ball.

Here the environment image capture by the robot camera is given as input. It enhances the object contours by OCS method and gradually reduces the shape dissimilarity value between the target object template and reconstructed object shape. The template selection is a step, that mimicking the human visual experiences because, If a partially occluded object contour or incomplete shape is observed, the rest of the shape can imagine, from the knowledge of object shapes by visual experiences. The statistical model of an incomplete

Fig. 1 Normal and robotic view of objects in the indoor environment.

shape (p.shape) comes from a certain complete shape (c.shape) templates formulated as Maximum a Posteriori (MAP) problem in Bayesian theory in formula (1). Where, p(p.shapelc.shape) is the posterior probability that Partial shape comes from the complete shape. Next segment or partition the template image into multiple partitions and compute the hausdroff similarity value between partial object shape and its segmented template.

$$\text{maxp}(\text{p.shape} \mid \text{c.shape}) = \text{maxp}(\text{c.shape} \mid \text{p.shape}).\text{p}(\text{c.shape}) \tag{1}$$

Hausdorff distance is a method to estimate the resemblance measurement between the two point sets (p.shape) and (c.shape"). Hausdorff distance is defined as follows,

$$\begin{aligned} &\text{H}(\text{p.shape}, \text{c.shape"}) \\ &= \max(\text{h}(\text{p.shape}, \text{c.shape"}), \text{h}(\text{c.shape"}, \text{p.shape})) \end{aligned} \tag{2}$$

$$\text{h}(\text{p.shape}, \text{c.shape"}) = \max(\min \text{d}(\text{c.shape"}, \text{p.shape})) \tag{3}$$

The function h (p.shape, c.shape") is a value in which each point in partial shape (p. shape) is near to some point in the segmented shape template (c.shape"). The Hausdorff distance is the maximum among h(p.shape, c.shape") and h (c.shape", p.shape). Based on this distance value OCS can fix on which part of segmented templates reduce its similarity value within the rate of 0.1 - 0.15. This template is assigned as missing object part and it will be reconstructed with the help of shape template.

2.2 Conical Cognitive Map Building with Human Vision

The robot map building is a process of converting sensor values perceived from the real environment to the suitable spatial coordinates that is image coordinates are transformed into space coordinate system. The proposed system builds the cognitive map of visibility coverage for environment

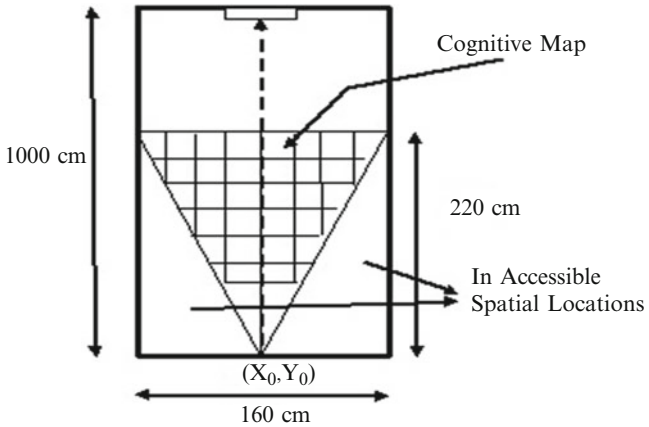


Fig. 2 Conical cognitive map structure for indoor environment.

based on human vision system and starts its movement like a human cognition using a single image. In human eye the coverage of surface area is differed where distance is increased towards the forward direction. The cognitive map building phase quantifies how many pixels occupied floor area of navigable space of the environment in the image in the straight line view. The coordinates of cognitive maps are represented as (X_i, Y_i). It computes the number of pixel covers for single grid cell which is 20 cm X 20 cm on the floor area. Every 20 cm distance increased towards forward direction the number of the grid increases gradually. In Fig 2 shows the human vision system based conical cognitive grid map model of the indoor environment. This conical cognitive geometric map is suitable for environment focused by straight line view and global search phase. It has knowledge about the entire configuration space and also locate the target object effectively upto a certain distance. This map structure makes path planning effectively by means of minimizing the search on the grid based metric environment because the object beyond the conical view has not been perceived by the camera, therefore it is not required to map that floor area into the cognitive map. Before estimating the distance value, find the bottom center (x_c, y_c) position value of the detected target object. The goal object search phase (section 2.1) returns the

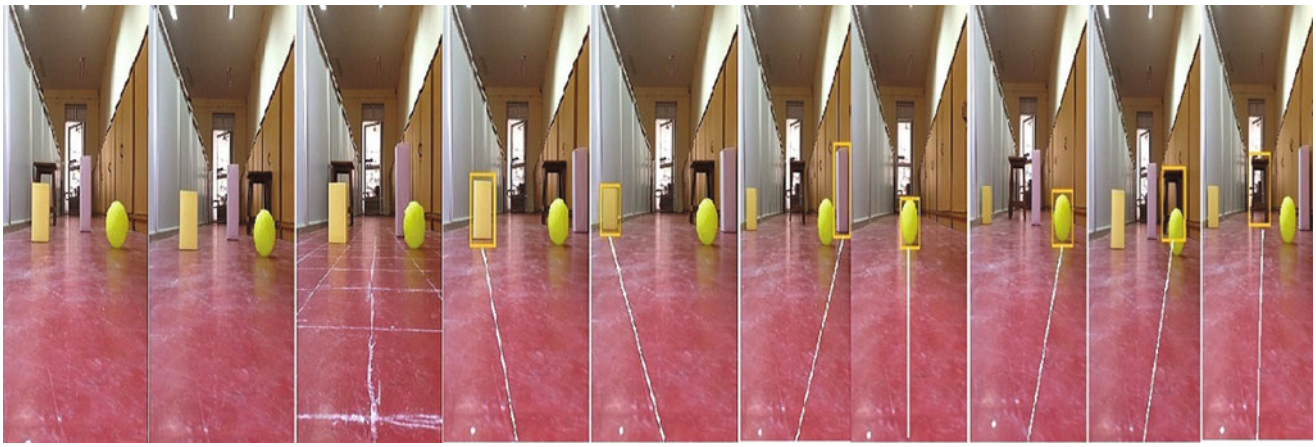


Fig.3 The partially occluded target object Stool and white box and sample results of proposed system.

starting pixel coordinates (x_0,y_0) of detected target object in the image. The below formula (4), (5) used to estimate the bottom center(x_c, y_c) value of the detected target object. Where, H, W is the height and width of the object respectively. Now pixel coordinates (x_c, y_c) mapped into the cognitive map. The coordinate values of the cognitive map for target object is represented as the X_t and Y_t.

$$x_c = x_0 + H \tag{4}$$

$$y_c = y_0 + (W/2) \tag{5}$$

After completing the mapping process to estimate the distance D_t between the robot and target object using the Euclidian distance method.

$$D_t = \sqrt{(X_t - X_0)^2 + (Y_t - Y_0)^2} \tag{6}$$

Here (X_0, Y_0) is the initial location of the robot. Now calculate the angle of deviation value of the target object from the straight line of focus. The angle can be calculated using the formula,

$$\Theta = tan^{-1} \frac{\sqrt{(X_t - X_i)^2 + (Y_t - Y_i)^2}}{\sqrt{(X_i - X_0)^2 + (Y_i - Y_0)^2}} \tag{7}$$

Here (X_i, Y_i) is the location intercept of the target object on the straight line. The earlier path determination system gives the results as cognitive map coordinates of target object (X_t, Y_t), distance to be traveled and angle of deviation values required to change the direction like left, right or move straight. The robot has decided the change of direction based on the value. This direction changes give as controls like turn left, turn right and move straight. These controls work like movements of the human eye.

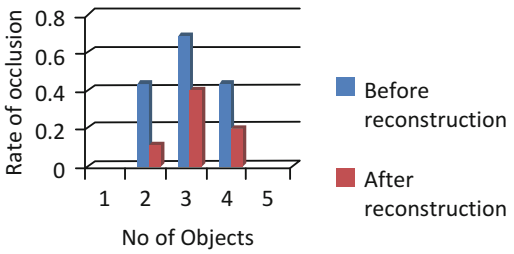


Fig.4 Reduction of shape dissimilarity after reconstruction for objects stool and white box.

3 Experimental Results

The images for our experimental evaluation were collected from the lab environment with a size of 1000 cm x 160 cm. The robot camera placed at the height of 20 cm from the floor level. The camera covers the entire width (160 cm) of the floor from length of 180 cm to 220 cm. Human made objects of four different shapes such as ball, yellow rectangle box, white rectangular box, and stool are placed on the floor at different locations. During object search phase the objects stool and white box were partially occluded by other objects by view of robots. OCS method effectively detects these objects and returns the location values efficiently in the image. In Fig. 3. Shows the stool, white box was occluded by yellow rectangular box, ball and sample resultant images for detection and path determination of four input objects. Compare the shape similarity value between the reconstructed object with its templates, the dissimilarity was drastically reduced after reconstruction process. It is shown in Fig 4. In the same way, detect remaining objects and find its locations.

Table 1. Shows the rate of dissimilarity value of white box and stool object before and after reconstruction. The dissimilarity values are reduced after reconstruction. It shows the similarity value is increased rapidly.

In cognitive map construction the pixel boundaries for each grid cells are estimated by analyzing 110 images. It covers 40 cm length and 40 cm width of the floor occupies 30 % of image from bottom to top. In this manner, the coverage area increased drastically. The entire width of the area (160 cm) covers a length of 180 cm from robot's initial location. The computation of number of pixels counting for 20 cm x 20 cm square area on the floor accurately estimated up to 220 cm long from its initial location in the view of straight line. The bottom center values (x_c, y_c) are mapped into the cognitive grid map and estimate the distance and angle values. If the objects are placed away from 220 cm long the mapping cannot be done accurately.

Table 1 The dissimilarity value of partially occluded objects before and after reconstruction.

	Dissimilarity Value	
Object	*Before Reconstruction*	*After Reconstruction*
White box-ball	0.4419	0.1205
Stool-yellow box	0.699	0.4175
Stool-ball	0.4391	0.2011

In Table 2 depicts comparison of coordinate values and path determination factors for object yellow ball, stool, yellow box, white box in real environment and proposed system. The comparison between these values results the following. The direction doesn't change for object white box where the X-axis value changed in cognitive map. The distance value is increased where the Y-axis value is changed in cognitive map for object yellow ball.

The object yellow rectangle box both the control parameters distance, angle and direction were changed in cognitive map by displacement of X, Y axis values and the distance value is reduced where Y-axis value and angle value is increased in cognitive map. The one detection result of stool object estimated as infinity due to misclassification of OCS method. The distance value is reduced where X-axis value is reduced in row 2. In Fig 5 shows the comparison of coordinate's values between the real environment and simulation results of cognitive map. The X axis and Y axis of the graph represents length and width of the environment. The X and Y coordinates are representing the location of object in real environment and simulated conical map. Some of the locations of objects were misplaced in the conical cognitive map.

Table 2 Comparison of Coordinate values and path determination factors for real world and cognitive map for input objects.

	Real environment					Cognitive Map				
Objects	X axis	Y axis	D (Cm)	Θ^0	Direction	X axis	Y axis	D (Cm)	Θ^0	Direction
Stool	1	11	220	15.25	Left	1	11	220	15.25	Left
	8	11	240	19.98	Right	7	11	220	15.25	Right
	3	11	220	5.19	Left	∞	∞	∞	∞	Can't Determine
Yellow ball.	6	6	120	18.43	Right	6	6	120	2	Right
	6	7	140	15.93	Right	6	8	160	14.03	Right
	4	6	120	0	Straight	4	6	120	0	Straight
Yellow box	2	10	200	11.30	Left	2	9	180	11.30	Left
	2	8	160	16.69	Left	2	8	160	16.69	Left
	3	4	80	30.96	Left	4	3	60	34.99	Straight
White box	8	10	200	16.69	Right	8	10	200	16.69	Right
	5	6	120	21.80	Straight	5	6	120	21.80	Straight
	4	10	200	5.71	Straight	5	10	200	0	Straight

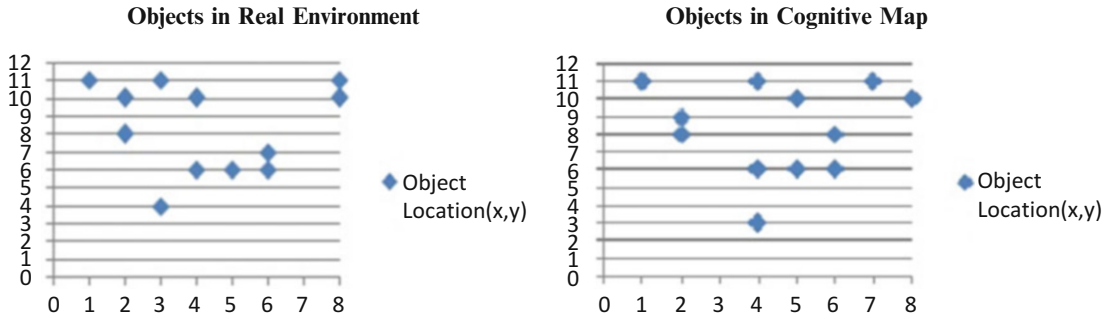


Fig. 5 Comparison of location values between real environment and estimated cognitive map.

4 Conclusion

The earlier path determination system makes path planning effectively by means of minimizing the search location on the grid based metric environment. The OCS object search algorithm acquires good detection results of partially occluded objects within 0.45 occlusion rate. The conical cognitive map coordinate values X and Y are decide the control parameters of mobile robots. Y axis values decide number of rotation required to move towards target and X axis values determines the change of direction whether left or right. In future, this human vision based path planning system can be extended to agriculture and domestic used robots like grass pruning, fruit picking and automated floor cleaning robots. It effectively detect the fruits occluded by leafs and branches.

References.

1. Kuan-Chieh Chen and Wen-Hsiang Tsai.: Vision-Based Autonomous Vehicle Guidance for Indoor Security Patrolling by a SIFT-Based Vehicle-Localization Technique, IEEE transactions on vehicular technology, vol. 59, no. 7, (2010).
2. Zhichao chen., Yinxiao Li., Stanley T. Brich field.,:Visual detection of lintel-occluded doors by integrating multiple cues using a data-dreiven markov chain monte carlo process, Robotics and Autonomus System,No-59,pp 966-976,(2011).
3. Widodo Budiharto, Djoko Purwanto and Achmad Jazidie,: A robust obstacle avoidance for service robot using Bayesian approach, International Journal of Advanced Robotic Systems, Vol. 8, No. 1, ISSN 1729-8806, pp 37-44, (2011).
4. Takeshi Saski, Drazen Brscic, Hideki Hashimoto,:Human-Observation-Based extraction of path patterns for mobile Robot Navigation, IEEE Transaction on Industrial Electronics, Vol-57, pp-1401-1410, (2010).
5. Emmanuel menegatti, Alberto Scarpa, Enrico pagello,:Omni directional vision scans matching for robot localization in dynamic environments, IEEE transaction on robotics Vol-22, No-3, pp: -523-535,(2006).
6. Lu Wang, Nelson hon Ching yang, :Three dimensional model based human detection in crowded scenes IEEE intelligent transportation system, Vol. 13, No-2, (2012).
7. Garage belong, Jithenra Malik, Jone puzicha,:Shape matching and object recognition using shape context, IEEE transaction on pattern analysis and machine intelligence, Vol-24, No-24, (2002).
8. Laporte. C., Brooks. R., and Arbel.T.,: A fast discriminant approach to active object recognition and pose estimation. I*CPR*, (2004).
9. Garage belong, Jithenra Malik, Jone puzicha,:Shape matching and object recognition using shape context, IEEE transaction on pattern analysis and machine intelligence, Vol-24, No-24, (2002).
10. Riemenschneider. H., Dacer. M., Bischof H.,: Using partial edge contour matches for efficient object category localization, European Conference on Computer Vision., (2010).
11. Srinivasan. P., Zhu.Q., and Shi.J., :Many-to-one contour matching for describing and discriminating object shape, IEEE Conferences. Computer Vision. And Pattern Recognition, (2010).
12. Kimia.B., Frankel.I., and Popescu.A.,:Euler spiral for shapecompletion, International. Journal on. Computer. Vision, vol. 54, pp. 157–180, (2003).
13. Mumford. D., "Elastica and computer vision,:Algebra. Journal on Geometric Application. pp. 491–50, (1994).
14. Ren. X., Fowlkes C., and Malik J., :Learning probabilistic models for contour completion in natural images, International Journal of Computer Vision., vol. 77, no-10, pp. 47–63, (2008).
15. Venkatesh., And.M. V., Cheung.S.S., :Symmetric Shape Completion under severe Occlusions, IEEE International Conference on Image processing, pp.709-712,(2006).
16. Kanizsa.G, Gerbino. W., Beck. J., :Amodal completion: Seeing or thinking," in Organization and Representation in Perception, (1982).
17. Reinsink.R. and Enns.J., :Early completion of occluded objects, journal of Visions. Research., Vol-38, pp-2489-2505,(1998).
18. Sebastian Thrun," Learning metric-topological maps for indoor mobile robot navigation" Science direct, Artificial Intelligence 99, pp. 21-71,(1998).
19. A.K.Reid., J.E.R.Staddon., :A reader for the cognitive map, Journal of Information Sciences 100 (1–4) 217–228, (1997).
20. E.C. Tolman., :Cognitive maps in rats and men, Psychological Review, 55 pp.189–208, (1948).
21. Soonyong Park., Soohwan Kim., Mignon Park., Sung-Kee Park.,: Vision-based global localization for mobile robots with hybrid maps of objects and spatial layouts", Information Sciences 179, pp. 4174–4198,(2009) ELSEVIER.
22. Michael P. Wellman,:Inference in cognitive maps, Mathematics and Computers in Simulation 36, pp. 137-148,(1994) ELSEVIER.
23. Orly Lahav., David Mioduser., :A blind person's cognitive mapping of new spaces using a haptic virtual environment", Journal of Research in Special Educational Needs, Volume 3, Number 3, pp. 172–177,(2003).
24. José Antonio Villacorta-Atienza., Valeri A., Makarov.,: Neural network architecture for cognitive navigation in dynamic environments, IEEE transactions on neural networks and learning systems, vol. 24, no. 12, pp. 2075-2087,(2013).
25. Shrihari Vasudevan., Stefan Gachter., Viet Nguyen., Roland Siegwart.,: Cognitive maps for mobile robots an object based approach, Robotics and Autonomous Systems 55, pp. 359–371, (2007),ELSEVIER.

Special Session: Nature-inspired Computational Methods and Applications

Teaching Learning Based Optimization (TLBO) Based Improved Iris Recognition System

Shikha Agrawal, Shraddha Sharma, and Sanjay Silakari

1 Introduction

The increasing rate of security threats increases the application of biometric security system. Biometric security system is most secured security systems in the current age of information technology. The variants of biometric security system are face recognition, finger print recognition, voice recognition and iris recognition system. In these entire biometric security systems, iris recognition system is more reliable and secured. Feature extraction and feature optimization is an essential area of research in the field of iris recognition system [1, 6]. The extraction of feature from iris image is a very difficult task due to lower content of the feature. In image processing concept, the iris is basically divided of three types such as color, texture and shape and these are considered as features of iris. There are various algorithms which can be applied to extract the features in feature extraction process and pattern matching processes. These algorithm uses texture features of the iris. Texture analysis based methods are considered as more suitable method to solve the problem of iris recognition. The texture feature extraction method is based on the fractal dimension which was first proposed to estimate length of coastline. There are several algorithms which have been developed to calculate the fractal dimension of a 2-D image [10]. In the iris recognition system the fractal dimension can be used to examine the texture of iris images efficiently and it has been accepted to illustrate the Gabor filtered images.

In the feature extraction process, Gabor wavelet transform function performs better than other methods which are used for extraction of texture feature from the iris image [16]. The feature space consists of two parts: one is relevant feature and other is unnecessary feature. It is very tedious task to reduce the unnecessary feature in feature space. To reduce such types of features,feature optimizationn techniques are used. Now a days, various methods for feature optimization are proposed such as PSO, Genetic Algorithm, and Neural Network by many researchers. Swarm Intelligence (SI) based techniques are low cost, accurate and efficient. Recently in 2011, a new SI based algorithm called Teaching Learning Based Optimization algorithm (TLBO) [3] have been developed by Prof. R V Rao in 2011. We use this algorithm to process the task of feature optimization in feature space and generate an optimal template for code generation for Iris Recognition System.

Organization of the paper is as follows: Section II presents TLBO based iris recognition system. Experimental results are discussed in section III and finally section IV gives conclusion.

2 TLBO Based Iris Recognition System

Iris feature optimization is a challenging task in the field of iris recognition. Now, optimization processes of iris image need a feature set of iris image data. In current years, various methodologies are available for iris feature optimization such as artificial neural network, genetic algorithm, particle swarm optimization etc. We proposed a TLBO based architecture for iris recognition system as shown in Fig. 1.

In the TLBO Based Iris Recognition System, feature extraction is optimized using using TLBO algorithm so as to enhance the efficiency of the system as a whole. The algorithm is as follows:

1. Extract the features from iris image dataset and assigned these values to feature matrix.
2. Transform data to the format of a feature space, i.e. transform feature matrix to feature space which is given by: $Xi\varepsilon R^d$ where X is an original feature, R is transform feature space and d is a dimension of data.

S. Agrawal (✉) • S. Sharma • S. Silakari
Department of Computer Science & Engineering, UIT, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, M.P, India
e-mail: shikha@rgtu.net; shraddhasharma2010@gmail.com; ssilakari@yahoo.com

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_105,

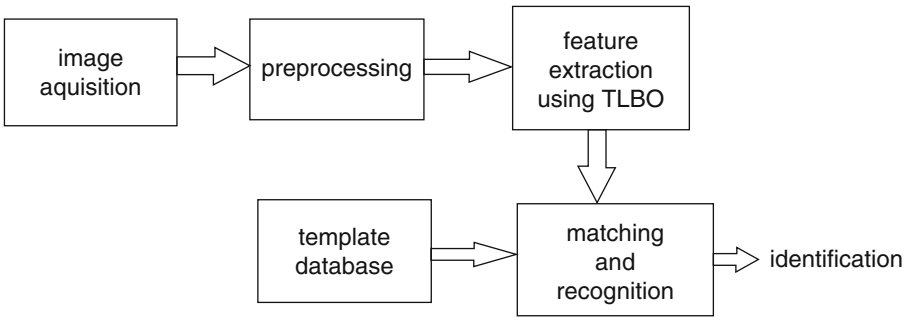


Figure 1 TLBO based architecture for iris recognition system

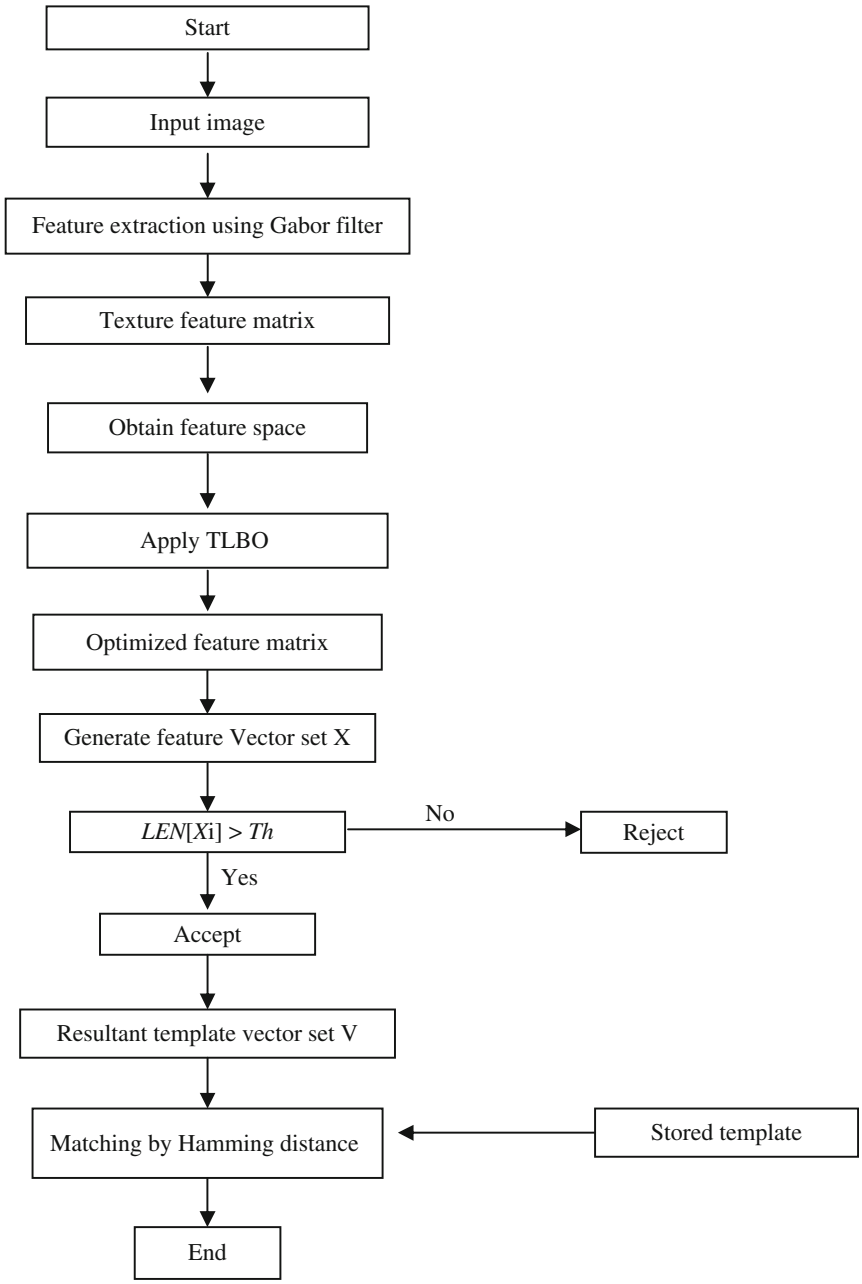


Figure 2 Proposed Model For Feature Optimization Of The Iris Using TLBO

3. Conduct scaling of the data i.e

$$\alpha = \sum_{i=1}^{m} \sum_{j=1}^{n} sim(Xi, Xj)/m*k$$

where α is scaling factor, m is total data point, k is the total number of instance and sim is the similarity function which finds the similarity between data.
4. The output of previous step is considered as input or population for TLBO algorithm. These features are analogous to number of students in TLBO.
5. The total population X is randomly distributed in total dimensions of matrix. Objective function F(x) is the maximum value of feature vector.
6. **Teacher phase**: Measure the mean columnwise in population i.e

$$M_{D^r} = [m_1, m_2 \ldots \ldots \ldots \ldots \ldots \ldots ., m_D]$$

The best solution which acts as teacher in TLBO for that iteration is the maximum value in feature matrix i.e

$$X_{teacher^r} = X_F(x) = maximum$$

The teacher is try to shift this mean from M_{D^r} to $X_{teacher^r}$ therefore the difference is calculated between $X_{teacher^r}$. And M_{D^r} which is given as:

$$difference = r(X_{teacher^r} - T_F M_{D^r})$$

Use the best value of T_F to select the optimal feature. The value of T_F is selected as 1 or 2.
The difference obtained is added to the current value to update it. It is calculated by:

$$X_{new} = X_{old} + difference$$

Now, if X_{new} gives maximum value then accept.
7. **Learner phase**: Learners increase their knowledge with mutual interation. Therefore for feature optimization select any two value from above step and update the smallest value corresponding to the largest one. This step is applied for all data calculated in above step. It is given as:

$$X_{new_i^g} = \begin{cases} x_i^g + rand \times (-X_r^g) & if \ \ f(x_i^g) < f(X_r^g) \\ x_i^g + rand \times (X_r^g - x_i^g) & otherwise \end{cases}$$

Where g is the current iteration and rand is a constant.
8. TLBO gets terminated either optimized feature matrix is obtained or maximum number of iterations are reached otherwise goto step 6.
9. Generate the feature vector from this matrix bu using Binarization.
10. Consider the vector C as $C_1, C_2, \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots .C_n$ set of optimized binary features.
11. Begin.
12. For each feature from the feature set do
{
(a) Calculate the bit length of each feature and store its value in vector X.
Let maxmimum bit length = max[X].

(b) From the vector X
{
Calculate threshold (Th) as:
Th=$floor[(maximum \ \ bit \ \ length/2)\text{-}1]$}

(c) If bit length of the feature $(X_i) >$ Th
Then accept that feature for match and store it in a vector V.
else discard the feature
}
13. We get resultant template vector V for matching step.
14. Exit.

The resultant optimized template generated by the above mentioned algorithm is then used for matching and identification. During matching process hamming distances between the optimal template and stored templates in the database are calculated and identification is done on the basis of these values. Figure 3 shows the proposed model for TLBO Based Feature Optimized Iris Recognition System.

3 Experimental Results

The performance of the proposed method for the Iris verification system is evaluated by performing some experimental task using MATLAB 7.8 on CASIA Iris Image Database, which contains different types of iris and the results have been calculated on the basis of FAR FRR, CRR, ERR and recognition rate parameters.

3.1 Iris Image Database

The proposed algorithm is tested on CASIA iris image database in which 756 iris images from 108 different eyes are present, which describes the efficiency and reliability of the proposed method.

3.2 Experimental Analysis

This paper first uses Gabor Wavelet Transform for extraction of features which are further optimized using TLBO algorithm. At last, the results are evaluated in terms of recognition rate (%) which is calculated using the false rejection rate (FRR) and false acceptance rate (FAR) parameters. The efficiency of iris recognition system has a high recognition rate.

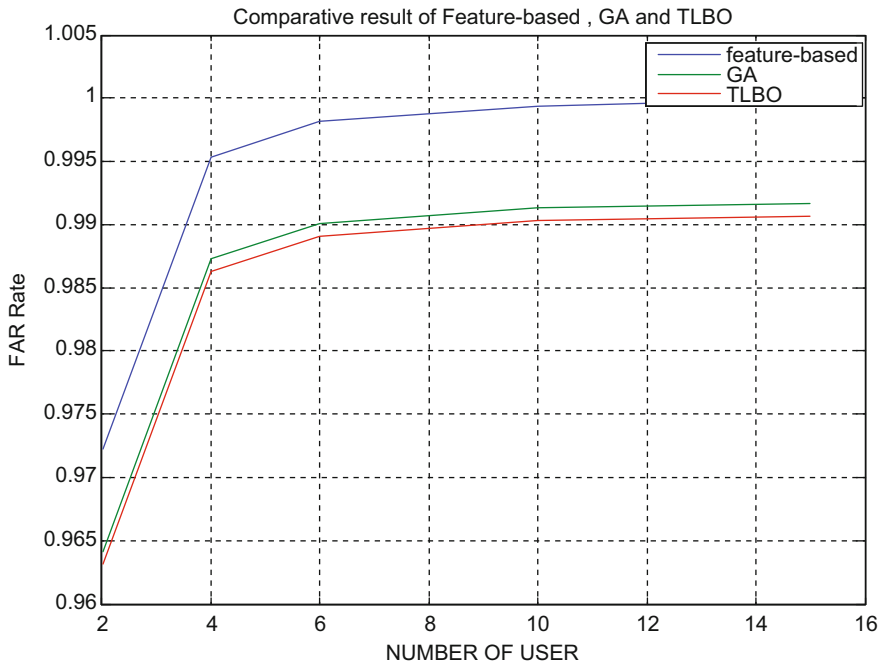


Figure 3 Comparative Result of FAR Rate

Table 1 Comparision Of Performance Metrics Using CASIA Dataset

Method	FAR	FRR	Best EER	Recognition Rate
Standard Iris Recognition System	0.998148	0.98627	0.00236914	94.4712
GAoptimized Iris Recognition System	0.990048	0.990048	0.00155914	95.4712
TLBO based Iris Recognition System	0.989048	0.980789	0.00145914	97.4712

FRR, FAR, CRR and Recognition Rate are defined as follows:

$$FAR = \frac{the\ number\ of\ false\ acceptation}{the\ total\ number\ of\ test\ sample}$$

$$FRR = \frac{the\ number\ of\ false\ rejection}{the\ total\ number\ of\ test\ sample}$$

$$CRR = \frac{Correctly\ recognized\ user\ number}{the\ total\ number\ of\ person\ enrolled}$$

$$Recognition\ Rate = \frac{1}{Recognition\ Time} * 100$$

In this experiment, 50 users are randomly selected from the database as experimental data in which each user has 7 images (total 360 images) out of which 3 images per users are considered.

3.3 Results

Our proposed method is an optimization technique to optimize elapse time with higher recognition rate. To calculate the performance of the proposed method over standard Iris Recognition System and Genetic algorithm optimized Iris Recognition System, we use FAR, FRR, CRR and Recognition Rate as the most important points for comparison. The experimental results of these methods are shown in Table 1. The results of the table shows the superiorty of the proposed method with recognition rate 97.4712 %.

Graphs of FAR, FRR and Recognition Rate with respect to number of users are shown in Figs. 3,4 and 5 respectively.

From Figs. 3 and 4, it can be clearly seen that FAR and FRR rate of TLBO based Iris Recognition System is less as compared to other two methods. This indicates that the proposed method is more secure and reliable than other two methods.

From Fig. 5, we conclude that TLBO based Iris Recognition System outperforms other two compared algorithms in terms of recognition rate as it shows highest recognition rate of 97.4712 %. When number of users increases, recognition rate of iris recognition system decreases in all the three cases. However, TLBO based Iris Recognition System's performance is still comparable to other two methods considered here, showing better scalability.

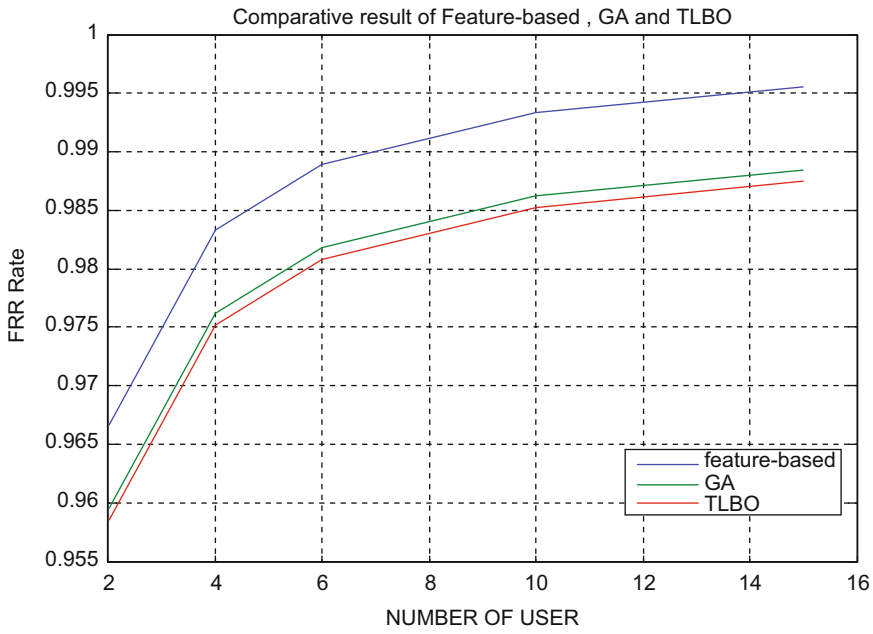


Figure 4 Comparative Result of FRR Rate

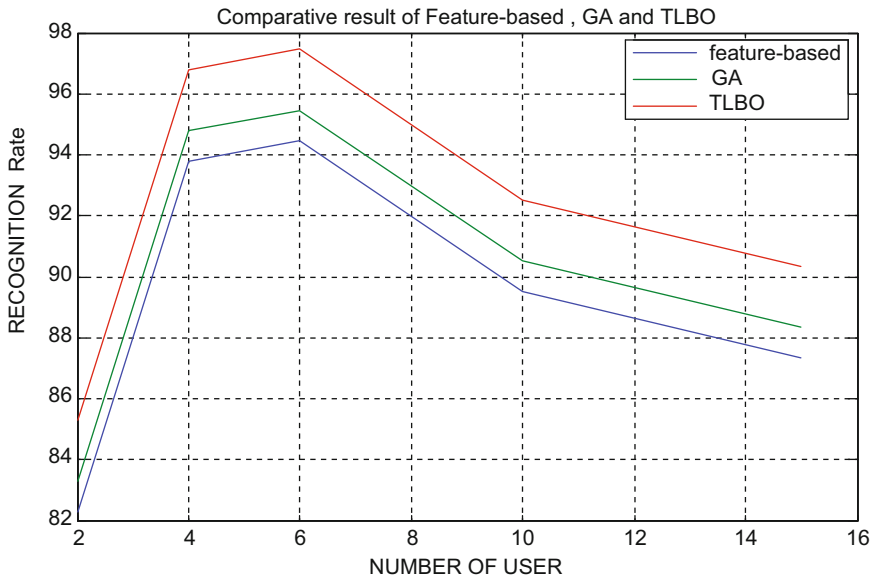


Figure 5 Comparative Result of Recognition Rate

4 Conclusion and Future Work

Iris recognition system is a novel approach of biometric authentication system which deals with security issues. Rececently, it is most secured and reliable system among other biometric systems. It consists of four steps: image acquisition, preprocessing, feature extraction and matching. Feature extraction is an important step in iris recognition system. Now-a-days, feature extraction and feature optimization is an open research area in the field of iris recognition system. Many researchers used SI based optimization techniques to optimize the features like Particle Swarm Optimization, Genetic Algorithm, Fuzzy Logic and Neural Network as these are very efficient, low cost and gives accurate results in optimization. TLBO is also one of the SI based technique for optimization. This paper proposed a TLBO based iris recognition system for feature optimization. This technique is processed on feature matrix extracted from feature extraction process. Here, gabor wavelet transform is used for texture feature extraction process. For optimization, TLBO is used and optimized feature matrix is obtained. Templates are generated with this optimized matrix by binarization and then matching is performed.The TLBO based iris recognition system gets 97.4712 % recognition rate and reduces the number of false rejection ratio.

The TLBO based iris recognition system shows superior performance in feature optimization when compared with standard iris recognition system and genetic algorithm optimized iris recognition system. In future,we will continue to examine the performance of our proposed method on other biometric system and after optimization we can also further classify the templates by using various classifiers like SVM.

References

[1] Hollingsworth K.P., Bowyer K.W., and Flynn P.J., "Using Fragile Bit Coincidence to Improve Iris Recognition," Proc. IEEE Third Int'l Conf. Biometrics: Theory, Applications, and Systems, pp. 1-6 (2009).

[2] Daugman J., "How Iris Recognition Works," IEEE Trans. Circuits and Systems for Video Technology, vol. 14, no. 1, pp. 21-30(2004).

[3] Rao R.V. and Kalyankar V.D., "Multi-objective multi-parameteroptimization of the industrial LBW process using a new optimization algorithm", Journal of Engineering Manufacture, 2012b, DOI: 10.1177/ 0954405411435865

[4] Bowyer K.W., Hollingsworth K.P., and Flynn P.J., "Image Understanding for Iris Biometrics: A Survey," Computer Vision and Image Understanding, vol. 110, no. 2, pp. 281-307(2008)

[5] Bolle R.M., Pankanti S., Connell J.H., and Ratha N., "Iris Individuality: A Partial Iris Model," Proc. Int'l Conf., Pattern Recognition, pp. 927-930 (2004)

[6] Hollingsworth K.P., Bowyer K.W., and Flynn P.J., "All Iris Code Bits Are Not Created Equal," Proc. IEEE Int'l Conf. Biometrics: Theory, Applications and Systems, pp. 1-6 (2007).

[7] Hollingsworth K.P., Bowyer K.W., and Flynn P.J., "The Best Bits in an Iris Code," IEEE Trans. Pattern Analysis and Machine Intelligence, vol 31, no. 6, pp. 964-973(2009)

[8] Barzegar N. and Moin M.S., "A New User Dependent Iris Recognition System Based on an Area Preserving Pointwise Level Set Segmentation Approach," EURASIP J. Advances in Signal Processing, pp. 1-13 (2009)

[9] Dozier G., Bell D., Barnes L., and Bryant K., "Refining Iris Templates via Weighted Bit Consistency," Proc. Midwest Artificial Intelligence and Cognitive Science Conference, pp. 1-5 (2009).

[10] Dozier G., Frederiksen K., Meeks R., Savvides M., Bryant K., Hopes D., and Munemoto T., "Minimizing the Number of Bits Needed for Iris Recognition via Bit Inconsistency and Grit," Proc. IEEE Workshop Computational Intelligence in Biometrics: Theory, Algorithms, and Applications, pp. 30-37(2009).

[11] Daugman J., "Demodulation by complex-valued wavelets for stochastic pattern recognition," Internat. J. Wavelets, Multi-Res. And Info. Processing, vol. 1, pp.1–17(2003).

[12] Adam M., Rossant F., Mikovicova B., and Amiel F., "Iris identification based on a local analysis of the iris texture". Proceedings of 6th International Symposium on Image and Signal Processing and Analysis (ISPA 2009), pp523-528 (2009).

[13] Adjedj M., Bringer J., Chabanne H., and Kindarji B., "Biometric Identification over Encrypted Data Made Feasible," Information Systems Security: Lecture Notes in Computer Science, pp86-100 (2009).

[14] Agrawal N.and Savvides M., "Biometric data hiding: A 3 factor authentication approach to verify identity with a single image using steganography, encryption and matching".IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops 2009), pp85-92 (2009).

[15] Alghamdi A.S., Ullah H., Mahmud M., and Khan M.K., "Bio-chaotic Stream Cipher-Based Iris Image Encryption". International Conference on Computational Science and Engineering (CSE '09), pp739-744 (2009)

[16] Al-Qunaieer F.S. and Ghouti L., "Color Iris Recognition Using Hypercomplex Gabor Wavelets". Symposium on Bio-inspired Learning and Intelligent Systems for Security (BLISS '09), pp18-19 (2009).

[17] Baig A., Bouridane A., Kurugollu F., and Qu G., "Fingerprint-iris fusion based identification system using a single Hamming distance". Symposium on Bio-inspired Learning and Intelligent Systems for Security (BLISS '09), pp 9-12 (2009).

[18] Baker S., Bowyer K., and Flynn P., "Empirical Evidence for Correct Iris Match Score Degradation with Increased Time-Lapse between Gallery and Probe Matches". Advances in Biometrics: Lecture Notes in Computer Science, pp 1170-1179 (2009).

[19] Baker S. E., Hentz A., Bowyer K. W., and Flynn P. J., "Degradation of Iris Recognition Performance Due to Non-Cosmetic Prescription Contact Lenses". Computer Vision and Image Understanding, pp 1030-1044 (2010).

[20] Bastys A., Kranauskas J., and Masiulis R., "Iris Matching by Local Extremum Points of Multiscale Taylor Expansion," Advances in Biometrics: Lecture Notes in Computer Science pp 1070-1079 (2009).

[21] Suresh Chandra Satapathy, Anima Naik and K Parvathi" A teaching learning based optimization based on Orthogonal design for solving global optimization problems" in Springer Open Journal (2013).

Acceleration based Particle Swarm Optimization (APSO) for RNA Secondary Structure Prediction

Jitendra Agrawal and Shikha Agrawal

1 Introduction

A ribonucleic acid (RNA) molecule consists of a chain of ribonucleotides connected together by covalent chemical bonds. Each ribonucleotide consists of one of the four bases such as adenine (A), cytosine (C), guanine (G) or uracil (U), and the exact sequence of bases along the chain, the primary structure of the molecule, determines the kind of RNA it is. The secondary structure is the outcome of hydrogen bonds between nucleotides that are situated far away in the chain. Characteristically these hydrogen bonds exist only between G and C, or A and U, or G and U (or vice versa). The nucleotides so linked are called base pairs and are labeled as GC, CG, AU, UA, or GU, or UG, the first base being the one with the smaller index in the chain.

Though RNA secondary structure forecast is fundamentally based on Nuclear Magnetic Resonance and X -ray crystallography, these methods are very complex, time-consuming and highly costly. This has given birth to the development of various mathematical and computational models which have now become highly indispensable in bioinformatics for RNA secondary structure forecast. Several investigational efforts have been carried out in the area of RNA secondary structure prediction with optimization techniques. The cardinal issue here is to find out, from among all the probable secondary structures, the one which has the minimum energy, because it is the most likely one to have a robust stable secondary structure. But from the computer perspective, it emerges into an optimization issue spread over a search space and has an objective function [11]. No wonder, a number of diverse optimization techniques and strategies have been introduced as an aftermath. With a view to eliminate this vexed dilemma, the contribution of diverse meta heuristics, intimately linked with the GA in the iterative search procedure like simulated annealing (SA), particle swarm optimization (PSO), ant colony optimization (ACO), and tabu search (TS), have also become subject matter of hot debate and discussion [21, 13]. Particle Swarm Optimization algorithm is employed to optimize the structure of an RNA molecule, by means of a sophisticated thermodynamic model [9]. Herbert et al. [22] have skillfully explained the advantages of SARNA-Predict, an RNA secondary structure prediction algorithm based on Simulated Annealing (SA). Estimation for the execution of SARNA-Predict in terms of forecast precision is established with native structures. Hao wu et al. [23] cleverly put forward a fuzzy adaptive particle swarm optimization (FPSO) blending particle swarm optimization (PSO) and fuzzy logic control (FLC) to forecast RNA secondary structure with the least energy. This technique is targeted at forecasting pseudoknots in the huge search spaces. The arithmetical outcomes and statistical investigations have proved beyond doubt that the innovative approach is competent to locate an optimal feature subset from a bulky noisy data set.

The forecast of RNA secondary structure with minimum free energy is a very critical function. In the perfect RNA structure prediction the calculation of base pairs plays a very effective role. With an eye on forecasting highly perfect RNA secondary structure several optimization techniques have been introduced in the literature. But it is unfortunate that each and every technique suffers from certain lacuna in the area of precise structure prediction.The drawbacks in the existing methods are solved by constructing a new optimization technique named as Acceleration base Particle Swarm Optimization (APSO) algorithm.

The outline structure of the paper is organized as follows:- In section 2, a brief discussion on the standard

J. Agrawal (✉)
School of Information Technology, UTD, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, M.P, India
e-mail: jitendra@rgtu.net

S. Agrawal
Department of Computer Science & Engineering, UIT, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, M.P, India
e-mail: shikha@rgtu.net

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_106,

PSO and the proposed APSO algorithm is presented. The experimental results on the multimodal benchmark functions and RNA structure prediction are given in Section 3 and conclusion of the paper is given in Section 4.

2 Proposed Acceleration based Particle Swarm Optimization (APSO)

2.1 Particle Swarm Optimization (PSO)

Particle swarm optimization (PSO) [26] is triggered by the community actions of organic creatures, like fishes and birds equipped with supreme qualities of clustering jointly to function hand in hand in recognizing advantageous locations in a particular zone, as evidenced by the fishes looking out for a food source.

PSO emulates the swarm behavior and individuals represent potential solutions in a D-dimensional search space. Particle i is often composed of four vectors:

$X_i = (x_i^1, x_i^2, \cdots, x_i^D)$ with x_i^d being its position in the dth dimension, $pbest_i = (pbest_i^1, pbest_i^2, \cdots, pbest_i^D)$ with $pbest_i^d$ being the best position in the dth dimension that particle i has found by itself, $V_i = (v_i^1, v_i^2, \cdots, v_i^D)$ with v_i^d being the velocity in the dth dimension, and $gbest = (gbest^1, gbest^2, \cdots, gbest^D)$ with $gbest^d$ being the global best position in the dth dimension that all particles have found. Particles in a swarm move through the search space by

$$V_i^d = V_i^d + c_1.r_1.\left(pbest_i^d - x_i^d\right) + c_2.r_2.\left(gbest^d - x_i^d\right) \quad (1)$$

$$x_i^d = x_i^d + \delta V_i^d \quad (2)$$

Where c_1 and c_2 are two constants often with the value of 2.0, and r_1 and r_2 are two independent random numbers uniformly generated in the range [0, 1] at each updating iteration from $d = 1$ *to* D, respectively.

The standard PSO technique has been extensively applied in various investigational efforts to achieve an optimal key for the dilemma. With a view to achieve a highly accurate optimal outcome, the defects inherent in the PSO technique such as premature convergence and loss of diversity have to be properly tackled and surmounted by initiating effective alteration or augmentation in the procedure of PSO. The cardinal defect of the PSO technique is a random value choice during the new particles generation procedure where during the course of the velocity calculation procedure the acceleration coefficients are engendered arbitrarily. The arbitrary value choice in the velocity procedure makes it essential that the created particles are in random. The unfortunate fact is that the random populations fail miserably in generating further precise outcomes. Therefore, to attain a further precise outcome and to steer clear of the PSO defects, rather than fixing the value of acceleration coefficient we introduce an Acceleration base Particle swarm optimization (APSO) technique in which acceleration coefficient values are updated on the basis of evaluation function. Next section briefly explain the proposed APSO.

2.2 Acceleration based Particle Swarm Optimization (APSO)

In our proposed technique, an Acceleration based Particle Swarm Optimization (APSO) algorithm is developed to solve premature convergence problem of the standard PSO. In APSO, the acceleration coefficients are selected based on the fitness value which will increase the accuracy of the results. The selection of acceleration coefficient values in APSO algorithm is described as follows,

$$nc_1 = c_1 \times (1 - \lambda) \quad (3)$$

$$nc_2 = c_2 \times (1 - \lambda) \quad (4)$$

In Equ. (3) and (4), the λ value is computed based on the fitness values is calculated as,

$$\lambda = \frac{\chi\left(1 + \phi(f_{\max} - f_{\min})^{\omega} - \left(f_{avg}{}^{\omega}\right)\right)}{\delta(f_{\max} - f_{\min})^{\omega} - f_{avg}^{\omega}} \quad (5)$$

$$\delta = \left(\frac{f_{\max} - f_{\min}}{f_{avg}}\right)^{\omega} \quad (6)$$

Where,

χ - Alteration probability

ω, ϕ - are the coefficient factors

$F_{\max}$, $F_{\min}$, F_{avg} - denotes the maximum, minimum and average fitness of the particles

By exploiting Equ. (3) and (4), the acceleration coefficients values, the velocity formula which is given in Equ. (1) is updated by,

$$V_i^d = V_i^d + nc_1.r_1.(a_1) + nc_2.r_2.(a_2) \quad (7)$$

In Equ. (7), the a_1 and a_2 values are computed by comparing the particles which have the minimum fitness values with the pbest particles values. If the both particles values are same the value of a1 and a2 is set as zero otherwise the values are 1.

In the APSO, the particles are generated by using the RNA structures base pairs. To generate the particles, different length RNA structures are stored in the database D. The process of APSO algorithm in RNA secondary structure prediction is described as follows,

Step 1: Initially, the particles are initiated by generate RNA sequence sets which contain the RNA base pairs (A, U, C, and G) in an encoded format. Then, the parameters of each particle, including its position and velocity are initiated.

Step 2: Every generated particles fitness value is calculated by the RNA fold algorithm. The particles which have the minimum free energy are selected as the best particles.

Step 3: Update $pbest_i$ of each particle and $gbest_i$. Based on these values, the particles velocity and positions are updated by exploiting the Eqn. (7) and (2). The particles are updated by using the both Equ. (7) and thus we get the new particles. From this new set of particles best particles are selected by comparing the particles values and these best particles are given to the fitness computation.

Step 4: Stop if the current optimization solution is good enough or some stopping criterion is satisfied. Other-wise, go to Step 2.

3 Experimental Results

The proposed Acceleration based PSO (APSO) is implemented in the working platform of MATLAB and results are compared with GA and standard PSO. The parameters which are utilized in the APSO are listed in Table 1.

3.1 Results of APSO on Benchmark Functions

The proposed APSO method is first tested on eight benchmark functions which are described as follows,

To accomplish the performance analysis process, we have performed 30 independent runs of APSO, GA and PSO methods. Result of convergence in terms of average fitness value for number of iterations, of all the three algorithms, for all the above mentioned functions is shown in Figure 1.

As can be seen from Figure. 1, our proposed APSO method has obtained global optima in less number of iterations as compared to PSO and GA methods. Among three optimization methods, GA has given poorer performance when compared to PSO and APSO methods whereas PSO performance degrades when compared to APSO. Hence, our proposed APSO method has better convergence rate for all the eight standard functions than the PSO and GA methods.

Table 1 Parameters Values in APSO

Parameters	Values
Population Size	20
r_1 and r_2	[0,1]
c_1 and c_2	0.02
nc_1 and nc_2	0.04

Name	Equation
Extended Powell	$f_0(x) = \sum_{i=1}^{n/4} (x_{4i-3} + 10x_{4i-2})^2 + 5(x_{4i-1} - x_{4i})^2 + (x_{4i-2} - 2x_{4i-1})^4 + 10(x_{4i-3} - x_{4i})^4$
Dixon	$f_1(x) = (1 - x_1)^2 + (1 - x_{10})^2 + \sum_{i=1}^{9} (x_i - x_{i+1})^2$
Sphere	$f_2(x) = \sum_{i=1}^{n} x_i^2$
Rastrigin	$f_3(x) = \sum_{i=1}^{n} \left(x_i^2 - 10\cos(2\Pi x_i) + 10\right)$
Rosenbrock	$f_5(x) = \sum_{i=1}^{n/2} 100\left(x_{2i} - x_{2i-1}^2\right)^2 + (1 - x_{2i-1})^2$
Schwefel	$f_6(x) = -\sum_{i=1}^{n} x_i \sin\left(\sqrt{\lvert x_i \rvert}\right)$
Griewank	$f_7(x) = \sum_{i=1}^{n} \frac{x_i^2}{4000} - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$
Bohachevsky	$f_8(x) = \sum_{i=1}^{n} \left(x_i^2 + 2x_{i+1}^2 - 0.3\cos(3\pi x_i) - 0.4\cos(4\pi x_{i+1}) + 0.7\right)$

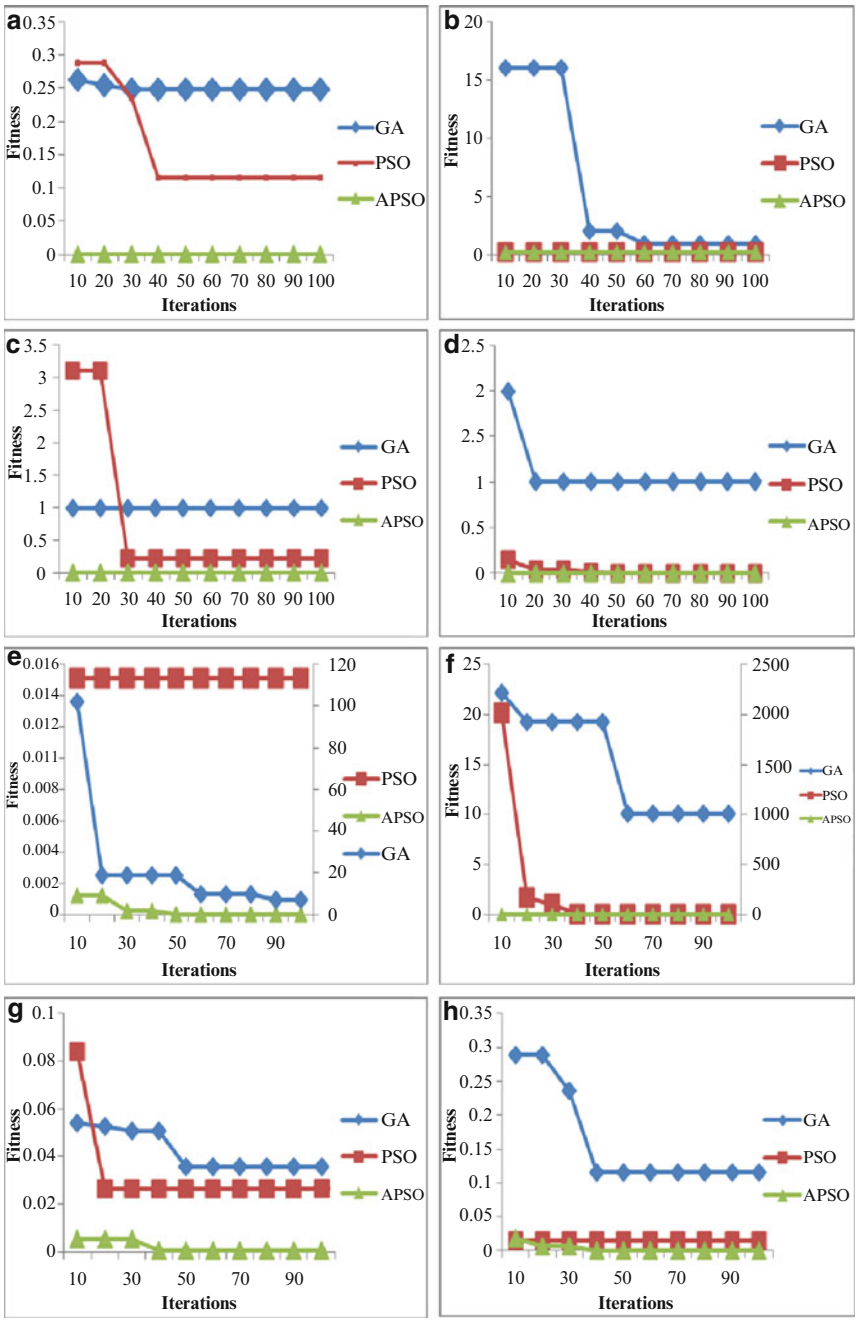


Fig. 1 Performance of APSO, PSO and GA methods

3.2 Results of APSO on RNA Secondary Structure Prediction

The proposed APSO, PSO & GA methods' performances are analyzed by conducting experiments on four different RNA sequences of varying sizes. The results of the 4 different experiments with 4 RNA sequences are given in Table 2.

Table 2 shows the performance of APSO, PSO & GA methods' performance in the RNA pair's prediction. In S. cerevisiae sequence 37 pairs are predicted in total, among which 31 pairs are correctly predicted with standard deviation 4.9 in APSO and in PSO & GA methods 29 & 28 pairs are correctly predicted respectively. In H.marismortui RNA sequence, APSO correctly predicts 201 base pairs among the 233 predicted base pairs and GA & PSO methods have correctly predicted 192 and 190 base pairs respectively. In other two RNA sequences X.laevis and D.virilis our proposed APSO correctly predicts 222 base pairs in 254 predicted base pairs and 30 base pairs in 38 predicted base pairs. When compared to the GA and PSO methods our proposed APSO method has given superior base pair prediction accuracy.

Table 2 Performance of (i) GA (ii) PSO (iii) APSO results on 4 RNA sequences

Sequences	Free Energy (ΔG kcal/mol)	Pairs Predicted	Pairs Correct	Pairs Known Structure	Percentage correct (%)
GA					
S.cerevisiae	-43.4	37 ± 0	29 ± 4.2	37	78.3
H.marismortui	-53.6	233 ± 0	192 ± 10.5	233	82.4
X.laevis	-205.9	254 ± 0	212 ± 11.6	254	83.6
D.virilis	-101.8	38 ± 0	27 ± 3.9	38	71.05
PSO					
S.cerevisiae	-44.4	37 ± 0	28 ± 4.4	37	75.6
H.marismortui	-52.5	233 ± 0	190 ± 10.3	233	81.5
X.laevis	-201	254 ± 0	215 ± 11.4	254	84.6
D.virilis	-103.5	38 ± 0	26 ± 3.8	38	68.3
APSO					
S.cerevisiae	-40.4	37 ± 0	31 ± 4.9	37	83.7
H.marismortui	-45.3	233 ± 0	201 ± 11.6	233	86.2
X.laevis	-193	254 ± 0	222 ± 12.8	254	87.4
D.virilis	-97.8	38 ± 0	30 ± 4.4	38	78.9

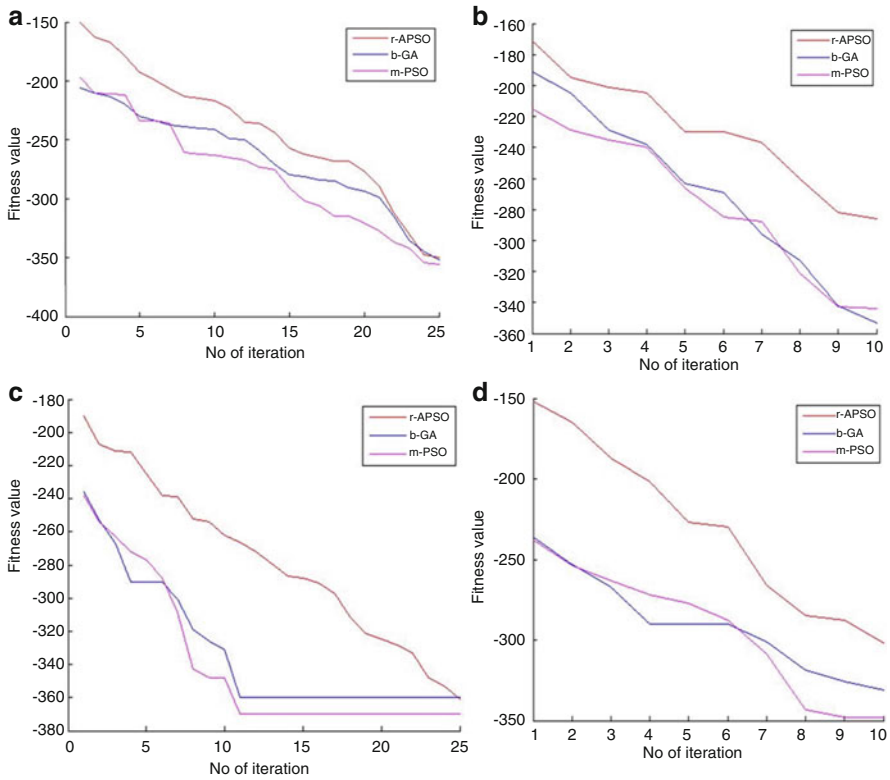


Figure 2 Performance of APSO, GA, PSO techniques in terms of free energy in RNA sequences

Performance of APSO is further evaluated on the basis of convergence characteristic on the 4 RNA sequences and is presented in Figure 2. APSO shows superior performance than GA and PSO in all the cases.

Prediction accuracy is shown in Figure 3. As can be seen from Fig. 4, our proposed APSO method has given a greater number of correctly predicted pairs than the GA and PSO methods in all 10 iterations. While comparing GA & PSO methods, GA has given superior prediction accuracy but this is lower than that of the APSO method. Hence, our proposed APSO method has given high performance in RNA structure prediction than the GA and PSO methods.

4 Conclusion

In this paper we have proposed a new Acceleration base Particle Swarm Optimization (APSO) algorithm for finding minimum energy RNA secondary structures. The proposed

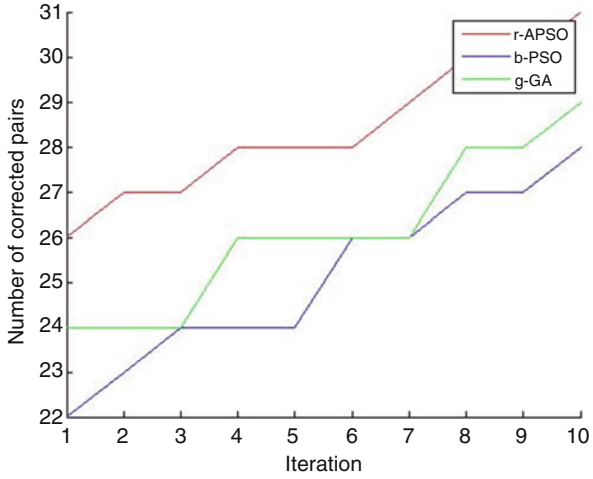


Figure 3 Comparison graph of our proposed APSO and GA, PSO methods performance in terms correctly predicted base pairs of sequence X.laevis

APSO is first tested on eight benchmark functions and then the algorithm is extended to solve RNA secondary structure prediction problems. The experimental result on benchmark functions shows that APSO acquires a better performance than the GA and standard PSO algorithms. Simulation to predict RNA secondary structure is performed on 4 different RNA sequences of varying lengths. Computational result shows that APSO algorithm not only converges well but also has high prediction accuracy as compared to GA and standard PAO algorithms. Hence, our proposed APSO algorithm is able to predict more corrected structures and also find better secondary structures in terms of free energy.

Reference

1. Hao Wu, Yan-feng Shi, Xing Jin, Gang Wang and Hao Dong, "A Fuzzy Adaptive Particle Swarm Optimization for RNA Secondary Structure Prediction", In Proceedings of International Conference on Information Science and Technology, Nanjing, pp. 1390-1393, 2011.
2. Arturo Díaz Perez Mario A. Garcia-Martinez, "FPGA Accelerator for RNA Secondary Structure Prediction", In Proceedings of 12th Euromicro Conference on Digital System Design / Architectures, Methods and Tools, Patras, pp. 667-671, 2009.
3. Eberhart R.C. and Kennedy J., "Particle swarm optimization,"IEEE International Conference on Neural Networks, Volume 4(27), pp 1942-1948, 1995.
4. Esquivel S., Coello C. and Coello A., "On the use of particle swarm optimization with multimodal functions", IEEE Congress on Evolutionary Computation (CEC), pp1130–1136, 2003.
5. Herbert H. Tsang and Kay C. Wiese, "SARNA-Predict: Accuracy Improvement of RNA Secondary Structure Prediction Using Permutation Based Simulated Annealing", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 7, No. 4, pp. 727-738, 2010.
6. Kangtai Wang and Ning Wang, "A novel RNA Genetic Algorithm for Parameter Estimation of Dynamic Systems", Journal of Chemical Engineering Research and Design, Vol. 88, No. 11, pp. 1485-1493, 2010.
7. Marais Neethling and A.P. Engelbrecht, "Determining RNA Secondary Structure using Set based Particle Swarm Optimization", In Proceedings of IEEE Congress on Evolutionary Computation, Vancouver, pp. 1670-1677, 2006.
8. Poli R., "An Analysis of the publications on the applications of Particle Swarm Optimization", Journal of Artificial Evolution & Applications, 2007.
9. S.S. Ray, M. Bachhar, and S.K. Pal, "RNA Secondary Structure Prediction in Soft Computing Framework: A Review," Proc. IEEE Third Int'l Conf. Computer Science and Information Technology, vol. 5, pp. 430-435, 2010.
10. Shubhra Sankar Ray and Sankar K. Pal, "RNA Secondary Structure Prediction Using Soft Computing", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 10, No. 1, pp. 2-17, 2013.

Performance Analysis of Zone Based Features for Online Handwritten Gurmukhi Script Recognition using Support Vector Machine

Karun Verma and R. K. Sharma

1 Introduction

Handwriting has continued to persist as a means of communication and recording information in day-to-day life even with the introduction of newer technologies. The system by which a computer can recognize characters and other symbols written by hand in natural handwriting is called handwriting recognition system. Handwriting recognition is classified into two categories, namely, offline handwriting recognition and online handwriting recognition. Online handwriting recognition involves automatic conversion of handwritten text as it is written on a special digitizer or PDA, where a sensor picks up the pen-tip movements as well as pen-up/pen-down switching. This kind of data is known as digital ink and can be considered as a dynamic representation of handwriting. The obtained signal is converted into letter codes which are usable within computer and text-processing applications.

2 Related work

Almuallim and Yamaguchi [1] proposed a technique for cursive handwritten Arabic script recognition. The words were segmented into strokes by them and these strokes were further classified using geometrical and topological features. Kurtzberg [2] proposed a method to recognize unconstrained handwritten discrete symbols based on elastic matching against a set of prototypes generated by individual writers. Since then, statistical, syntactical and structural, neural network and elastic matching are the common handwriting recognition methods [3, 4]). Dutta and Chaudhary [5] proposed a curvature feature based recognition method for Bangla alphanumeric characters, and found the technique effective with two stage feed forward neural network. Garcia et al. [6] presented a neural predictive system for online writer independent character recognition. A fixed number of predictive Neural Networks (NN) were used to model each letter. For ten different writers the system gave quite good recognition rate and improvement in results reported with the extension of the system to the durational HMM framework. Hu et al. and Takahashi et al. proposed HMM based algorithms for online hand written character recognition. Hu et al. [7] achieved a writer independent recognition rate of 94.50% on 3,823 unconstrained online handwritten word samples from 18 writers covering a 32 word vocabulary, whereas, 90.00% recognition rate was achieved by Takahashi et al. [8] for 881 Kanji characters. Connell et al. [9] proposed a recognition system for unconstrained online Devanagari characters. An accuracy of 86.50% with no rejects was achieved. Lehal and Singh [10] presented a system for recognition of machine printed offline Gurmukhi script and operated at sub-character level with a recognition rate of 96.60%. Bhattacharya et al. [11] proposed a scheme for recognition of offline Bangla characters with local chain code histogram of input character shape. Recognition accuracy of 92.14% on the test set and 94.65% on the training set was achieved. Sharma et al. [12] proposed a scheme for offline handwritten Devanagari character recognition based on quadratic classifier. They achieved an accuracy of 98.90% and 80.40% for Devanagari numerals and characters, respectively, using 5-fold cross validation technique. Bhawmik et al. [13] proposed a SVM based hierarchical classification scheme for recognition of handwritten Bangla characters. An accuracy of 88.02% was reported with hierarchical learning architecture scheme.

K. Verma (✉)
Department of Computer Science, Thapar University, Patiala, Punjab, India
e-mail: karun.verma@thapar.edu

R.K. Sharma
School of Mathematics and Computer Applications, Thapar University, Patiala, Punjab, India
e-mail: rksharma@thapar.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_107,

Rajashekraradhya and Ranjan [14] defined Zone centroid and Image centroid based Distance metric feature extraction system for offline handwritten characters of Kannada, Tamil, Telugu and Malayalam Numerals. The average recognition accuracy achieved was 97.80% for the four scripts with nearest neighbour and neural networks being used as classifiers. Pradeep et al. [15] proposed a diagonal feature based method for offline handwritten characters with neural networks as classifier. A recognition rate of 97.80% was achieved by them. Kumar et al. [16] proposed a SVM based offline handwritten recognition system using zone based features with a recognition accuracy of 94.29%. Kumar et al. [17] proposed a *k*-NN based recognition system for Gurmukhi script using zone based features and achieved an accuracy of 94.12%. Kumar et al. [18] described the PCA based approach for finding the feature classifier combinations for offline handwritten Gurmukhi character recognition. The recognition rate was improved from 94.85% to 97.71% with use of PCA. This paper is organized into six sections. Section 1 has introduced basic concepts and current section has described the literature related to the problem. Section 3 describes the features of Gurmukhi script and how online handwritten data was collected. In Section 4, preprocessing steps applied on the collected data have been explained. This section also explains how various features were calculated and organized for recognition of strokes. Section 5 describes various parameters used for support vector machine to recognize stroke classes. In section 6, we have concluded the work done with salient findings.

3 Gurmukhi script and data collection

In this work, we have addressed the problem of classifying different characters of Gurmukhi script. This script is one of the popular scripts in north India. This is also popular in Punjab province of Pakistan. Besides this, this script is very frequently used by Punjabis settled in Canada, USA, UK, Australia and other countries. Table 1 contains the characters used in Gurmukhi script. This script contains 9 vowels (laga matras; A8-I8) and 41 consonants (A0-G5), two symbols for nasal sounds *bindī* (I0) and *ṭippī* (I2), and one symbol which duplicates the sound of any consonant *addak* (I4) with writing style from left to right. Gurmukhi script characters may lie in three horizontal zones, namely, upper zone, middle zone and lower zone. The upper zone denotes the region above the head line, where some of the vowels and sub-parts of some other vowels reside, while the middle zone represents the area below the head line where the consonants and some sub-parts of vowels are present. Middle zone consists of most of the Gurmukhi characters. The lower zone represents the area below middle zone where some vowels and certain half characters lie in the foot of consonants. Some of the recognized strokes in various zones are depicted in Figure 1.

	0	1	2	3	4	5	6	7	8	9
A	ੳ	ਅ	ੲ	ਸ	ਹ	ਸ਼		ਆ	ਾ	ਕਾ
B	ਕ	ਖ	ਗ	ਘ	ਙ	ਖ਼		ਇ	ਿ	ਕਿ
C	ਚ	ਛ	ਜ	ਝ	ਞ	ਗ਼		ਈ	ੀ	ਕੀ
D	ਟ	ਠ	ਡ	ਢ	ਣ	ਜ਼		ਉ	ੁ	ਕੁ
E	ਤ	ਥ	ਦ	ਧ	ਨ	ਫ਼		ਊ	ੂ	ਕੂ
F	ਪ	ਫ	ਬ	ਭ	ਮ	ਲ਼		ਐ	ੈ	ਕੈ
G	ਯ	ਰ	ਲ	ਵ	ੜ			ਏ	ੀ	ਕੀ
H	Consonants							ਓ	ੋ	ਕੋ
I	ਂ	ਃ	ਁ	ੰ	ੱ			ਔ	ੌ	ਕੌ
J	Other Symbols							Vowels		
K	੦	੧	੨	੩	੪	੫	੬	੭	੮	੯
	Numerals									

Table 1 Chart of Gurmukhi Characters

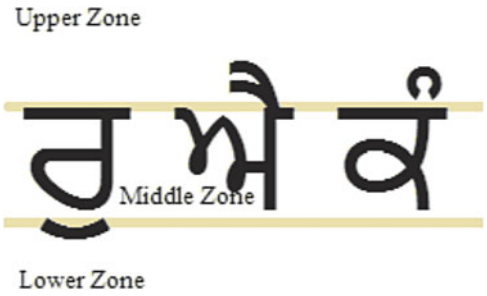


Fig. 1 Zone division of Gurmukhi characters

3.1 Data Collection

Dell Latitude XT3 Tablet PC was used to collect samples of Gurmukhi characters with strokes as the smallest unit. Stroke is a typical pen movement with any capturing device that traces points between successive PEN DOWN and PEN UP movements. In the data collection phase, samples of Gurmukhi characters from 30 different writers were taken. The data were further annotated at the stroke level assigning unique StrokeIDs to each stroke. A total of 101 unique strokes were detected from this set of data. Each stoke was assigned a class according to the zone in which the stroke appears. Details of each class is depicted in Table 2. There are 12 stroke classes in upper zone, 7 stroke classes in lower zone, and 82 stroke classes in middle zone. In the current study, one hundred samples per class of the middle zone strokes (A total of 8200 strokes) have been used for the process of classification.

4 Preprocessing and Feature Extraction

During preprocessing, a series of transformations on the stroke trace captured between successive PEN DOWN and PEN UP are done. These transformations helped in removal of noise or distortions that may arise due to hardware or software imperfections. At the start, normalization and centering of the stroke is done, where the stroke vector is fitted into a window of 500500 pixels. After normalization, the image is binarized by defining the value at each pixel as: $pix_{ij} = \frac{0,\ if\ Background\ color}{1,\ if\ Foreground\ color}$, $0 \le i,\ j < 500$, This 500×500 pixel vector has been used for further defining the features.

4.1 Zone Base Feature Extraction

Each preprocessed stroke is divided into 5050 sized 100 zones as shown in Figure 2(a). Each zone Z_{pq} in range $[Z_{00} - Z_{99}]$ is defined as a set of pixel values as,

$$Z_{pq} = \bigcup_{\substack{50*p \le i < 50*(p+1) \\ 50*q \le j < 50*(q+1)}} pix_{ij} \ \forall\ 0 \le p,q \le 9 \quad (1)$$

Table 2 Details of classes

Class of Stroke	No. of Strokes	Examples
Lower Zone	7	
Middle Zone	82	
Upper Zone	12	

in order to further find the features. Pixel information in each zone looks like Figure 2(b).

Normalized features. For finding Normalized features, a zonal sub-feature f_{pq} is calculated for each zone Z_{pq} as the sum of pixel values in the zone Z_{pq} as given below.

$$f_{pq} = \sum_{\substack{50*p \le i < 50*(p+1) \\ 50*q \le j < 50*(q+1)}} pix_{ij} \ \forall\ 0 \le p,q \le 9 \quad (2)$$

From this sub-feature a normalized feature value ${F_{pq}}^N$ which lies between [0,1] is calculated as:

$$F^N_{pq} = \frac{f_{pq}}{\sum_{0 \le i,\ j < 500} pix_{ij}} \ \forall\ 0 \le p,q \le 9 \quad (3)$$

This gives a feature set of 100 features per stroke. The feature set has been referred as D_N in further parts of the paper.

Diagonal features. In each zone, sum of pixels is calculated by moving along its diagonal. There are a total of 99 diagonals formed in the 5050 pixel zone. Each diagonal contributed to a single sub-feature ${d_{pq}}^v$ as given below.

$$\begin{aligned} d^v_{pq} = \sum_{\substack{i_{min} \le i < i_{max} \\ j_{min} \le j < j_{max}}} & pix_{ij} \ \forall\ v = (i+j) - (i_{min} + j_{min}) \\ & i_{min} = 50*p,\ i_{max} = 50*(p+1) \\ & j_{min} = 50*q,\ j_{max} = 50*(q+1) \\ & \forall\ 0 \le p,\ q \le 9 \end{aligned} \quad (4)$$

The average of these 99 sub-features is now calculated to form the feature for the zone to get D_{pq} for each zone.

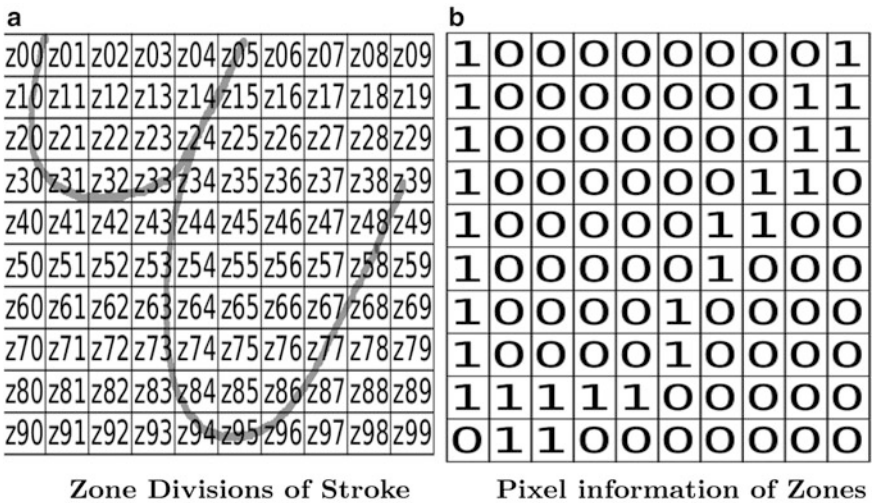


Fig. 2 Zone based Features

$$D_{pq} = \frac{1}{99} \sum_{0 \le v < 99} d_{pq}^{v} \tag{5}$$

Each zone contributed one such feature, giving a total of 100 features per stroke. This feature set will be referred as D_D in further parts of the paper.

Directional features. The directional feature for each zone was extracted using the starting point and ending point in the zone. The starting point in the zone was found by traversing the zone from left to right and top to bottom, the first vector $(v_i^s,\ v_j^s)$ whose value is 1 and ending point is the first vector $(v_i^e,\ v_j^e)$ whose value is found 1 while traversing the zone right to left, bottom to top. Algorithm 1 has been used to calculate directional feature D_{pq}

The angle between starting point and the ending point have been calculated. The zones where no pixel was present, the feature was considered as zero. This feature set will be referred as D_{Di} in further parts of the paper.

Parabola based curve fitting features. Each preprocessed stroke was divided into 5050 sized 100 zones. A parabola $j_{pq} = ai_{pq}^2 + bi_{pq} + c$ is fitted on a series of pixels in each zone using least square method. The parameters a, b and c uniquely define the parabolic curve in the zone, giving 3 features for each zone and hence 300 features per stroke. This feature set will be referred as D_{Pr} in further parts of the paper.

Power Curve based features. Each stroke is here divided into 5050 sized 100 zones. A power curve $j_{pq} = ai_{pq}^b$ is fitted on a series of ON pixels in each zone using least square method. The parameters a and b uniquely define the curve in the zone, giving 2 features for each zone and hence 200 features per stroke. The feature set has been referred as D_{Pw} in further parts of the paper.

5 Classification

As mentioned earlier, the strokes in middle zone have only been considered for recognition process in this work. Here, a total number of 8200 strokes have been taken for training. This set contains 100 samples for each of 82 strokes in middle zone. The feature set consists of the features as extracted in the previous section. SVM with kernels, namely, linear, polynomial, sigmoid, and RBF have been used for classification. Other parameters that have empirically been experimented in this work are learning rate (γ) and tolerance limit of termination (ε). Recognition rates for different experiments have been obtained using k-fold cross validation for three values of

1. **for** $i = 50 * p$ to $50 * (p + 1) - 1$ **do**
2. **for** $j = 50 * q$ to $50 * (q + 1) - 1$ **do**
3. **if** $p_{ij} == 1$ **then**
4. $v_{pq}^{si} = i$;
5. $v_{pq}^{sj} = j$;
6. $break$;
7. **end if**
8. **end for**
9. **end for**
10. **for** $i = 50 * (p + 1) - 1$ to $50 * p$ **do**
11. **for** $j = 50 * (q + 1) - 1$ to $50 * q$ **do**
12. **if** $p_{ij} == 1$ **then**
13. $v_{pq}^{ei} = i$;
14. $v_{pq}^{ej} = j$;
15. $break$;
16. **end if**
17. **end for**
18. **end for**
19. $D_{pq} = \tan^{-1}\left(\frac{v_{pq}^{ei} - v_{pq}^{si}}{v_{pq}^{ej} - v_{pq}^{sj}}\right)$

Algorithm 1 Directional feature calculation

Table 3 Parameters for SVM

Parameter	Option/Values Considered
Folds (k)	4, 5, 6
Kernel	Linear(0), Polynomial(1), RBF(2) and Sigmoid(3)
Learning rate (γ)	0.01 (0.01) 0.5
Tolerance limit of termination (ε)	0.1 (0.1) 0.5

k as mentioned in Table 3. Table 3 also contains the values of other parameters considered in this work.

The experiment results for all the training sets are shown in Figures 3-7. In Figure 3, for feature set D_N, a recognition accuracy of 88.93% has been achieved for the data set considered in this work when 5-fold cross validation with polynomial kernel is used. Here, this can be noted that polynomial kernel gives better recognition rate when compared with other kernels for the three values of folds in k-fold cross validation, whereas sigmoid kernel gives the lowest recognition rate for the same set of parameters. The experimental results for D_D feature set shown in Figure 4 explicate that a recognition rate of 92.09% could be achieved with this feature set when polynomial kernel and 5-fold cross validation is used. Here also, polynomial kernel outperforms other kernels for the folds examined in this work. The experimental results for D_{Di} feature set is shown in Figure 5. A maximum recognition rate of 56.38% was achieved for this feature set with linear kernel at 5 folds. It has been seen that linear and polynomial kernels are achieving similar recognition rates at different folds.

The results for feature sets D_{Pr} and D_{Pw} are presented in Figures 6-7. The recognition accuracies achieved using these two features is 73.60% and 20.89%, respectively.

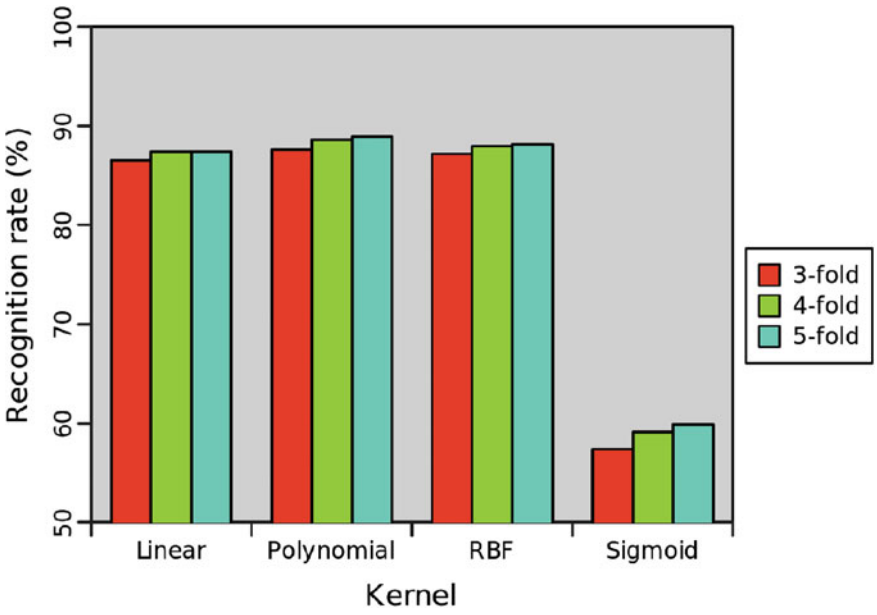


Fig. 3 Classification results of D_N for middle zone stroke set

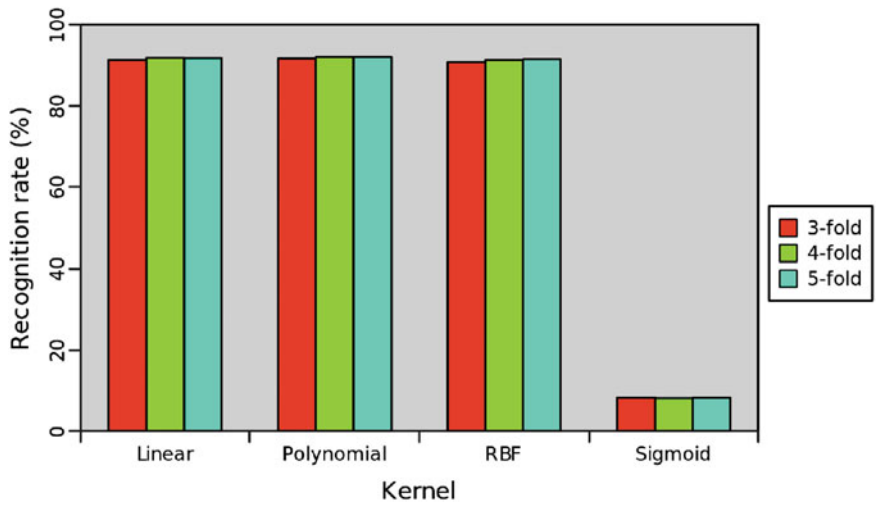


Fig. 4 Classification results of D_D for middle zone stroke set

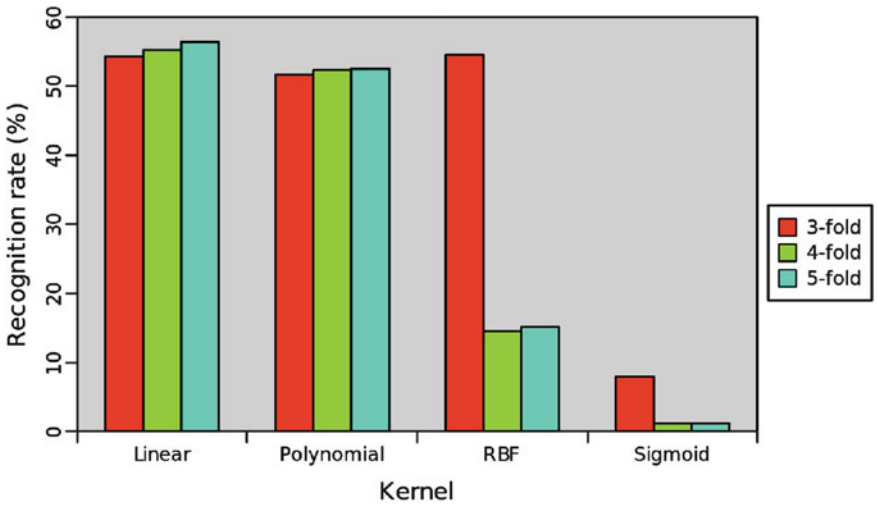


Fig. 5 Classification results of D_{Di} set

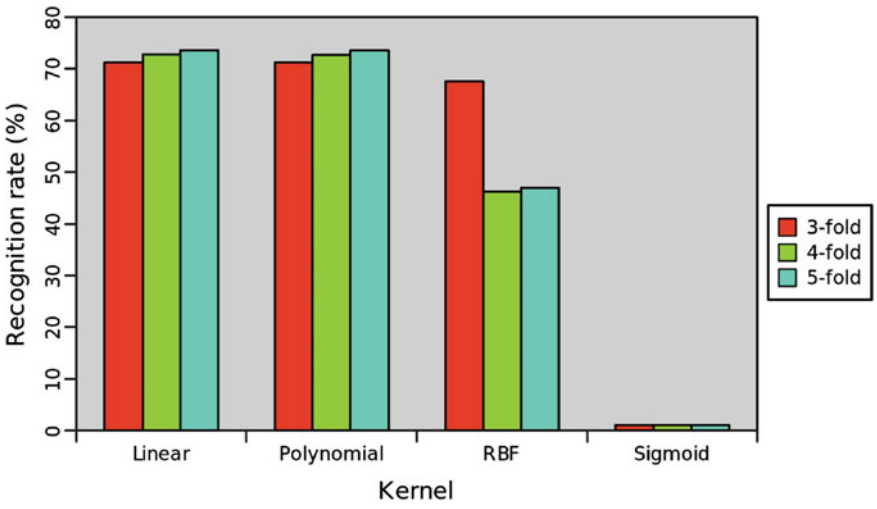


Fig. 6 Classification results of D_{Pr} set

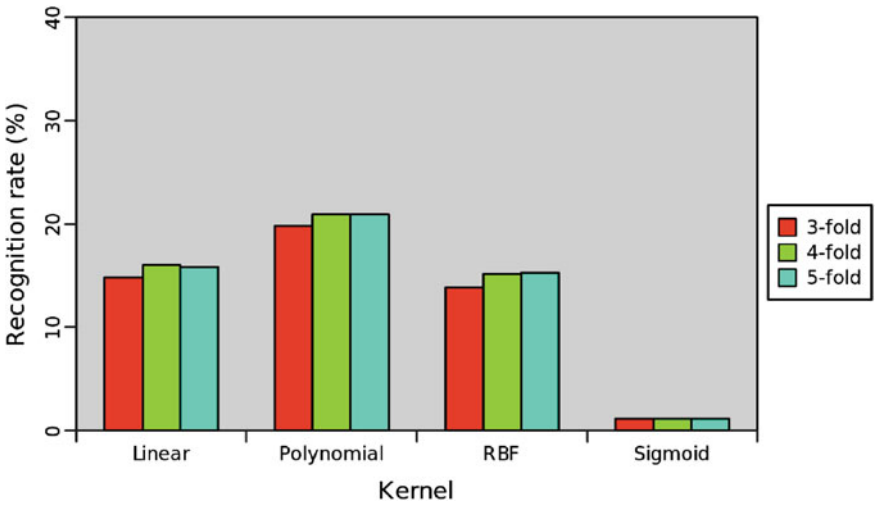


Fig. 7 Classification results of D_{Pw} set

6 Results and Discussion

The work presented here in this paper analyses the performance of various zone based features for classification of strokes in Gurmukhi script. Five different zone based features were elaborated. Four different kernels, namely, linear, polynomial, radial basis function and sigmoid for SVM training were considered. Other parameters such as *k*-folds, learning rate and tolerance limit were experimented for classification of strokes. D_D feature set yielded a 92.09% recognition rate out of the five features considered in this work. Linear and polynomial kernels achieved approximately the same recognition accuracy for all the five feature sets. As a further work in this direction, one can experiment for achieving higher recognition rates for the above feature sets by including other combinations of parameters. This study can further be extended to other strokes of Gurmukhi script. Experiments can also be done on a larger dataset.

Acknowledgements We take this opportunity to extend our special thanks to Technology Development for Indian Languages (TDIL), DeitY, MoCIT, Government of India for sponsoring the data collection used in this work.

References

1. Almuallim, H., Yamaguchi, S.: A method of recognition of arabic cursive handwriting. Pattern Analysis and Machine Intelligence, IEEE Transactions on (5) (1987) 715–722
2. Kurtzberg, J.M.: Feature analysis for symbol recognition by elastic matching. IBM journal of research and development **31**(1) (1987) 91–95
3. Bellegarda, E.J., Bellegarda, J.R., Nahamoo, D., Nathan, K.S.: A continuous parameter hidden markov model approach to automatic handwriting recognition. (July 14 1993) EP Patent 0,550,865.
4. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. Pattern Analysis and Machine Intelligence, IEEE Transactions on **22**(1) (2000) 4–37
5. Dutta, A., Chaudhury, S.: Bengali alpha-numeric character recognition using curvature features. Pattern Recognition **26**(12) (1993) 1757–1770

6. Garcia-Salicetti, S., Doizzi, B., Gallinari, P., Mellouk, A., Fanchon, D.: A hidden markov model extension of a neural predictive system for on-line character recognition. In: Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on. Volume 1., IEEE (1995) 50–53
7. Hu, J., Brown, M.K., Turin, W.: Hmm based online handwriting recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **18**(10) (1996) 1039–1045
8. Takahashi, K., Yasuda, H., Matsumoto, T.: A fast hmm algorithm for on-line handwritten character recognition. In: Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on. Volume 1., IEEE (1997) 369–375
9. Connell, S.D., Sinha, R., Jain, A.K.: Recognition of unconstrained online devanagari characters. In: Pattern Recognition, 2000. Proceedings. 15th International Conference on. Volume 2., IEEE (2000) 368–371
10. Lehal, G., Singh, C.: A gurmukhi script recognition system. In: Pattern Recognition, 2000. Proceedings. 15th International Conference on. Volume 2., IEEE (2000) 557–560
11. Bhattacharya, U., Parui, S., Shaw, B., Bhattacharya, K., et al.: Neural combination of ann and hmm for handwritten devanagari numeral recognition. In: Tenth International Workshop on Frontiers in Handwriting Recognition. (2006)
12. Sharma, N., Pal, U., Kimura, F., Pal, S.: Recognition of off-line handwritten devnagari characters using quadratic classifier. In: Computer Vision, Graphics and Image Processing. Springer (2006) 805–816
13. Biadsy, F., El-Sana, J., Habash, N., et al.: Online arabic handwriting recognition using hidden markov models. In: Tenth International Workshop on Frontiers in Handwriting Recognition. (2006)
14. Rajashekararadhya, S., Ranjan, P.V.: Efficient zone based feature extration algorithm for handwritten numeral recognition of four popular south indian scripts. Journal of Theoretical & Applied Information Technology **4**(12) (2008)
15. Pradeep, J., Srinivasan, E., Himavathi, S.: Diagonal based feature extraction for handwritten alphabets recognition system using neural network. Computer Science and Informatin Technology, International Journal o **3**(1) (2011) 27–38
16. Kumar, M., Sharma, R.K., Jindal, M.K.: Svm based offline handwritten gurmukhi character recognition. SCAKD Proceedings **758** (2011) 51–62
17. Kumar, M., Jindal, M.K., Sharma, R.K.: k-nearest neighbor based offline handwritten gurmukhi character recognition. In: Image Information Processing (ICIIP), 2011 International Conference on, IEEE (2011) 1–4
18. Kumar, M., Sharma, R.K., Jindal, M.K.: Offline handwritten gurmukhi character recognition: Study of different feature-classifier combinations. In: Proceeding of the Workshop on Document Analysis and Recognition. DAR '12, New York, NY, USA, ACM (2012) 94–99

Words Are Analogous To Lymphocytes: A Multi-Word-Agent Autonomous Learning Model

Jinfeng Yang, Xishuang Dong, and Yi Guan

1 Introduction

Words are the basic structural units of language, and they interact with each other and follow certain rules to form sentences. During interacting, some words depend on others or are depended upon by others. As an example, Figure 1 shows that a Chinese sentence "上海浦东开发与法制建设同步(Development is synchronized with legal construction in Pudong of Shanghai)", which is excerpted from Penn Chinese Treebank 5.1(CTB) [21], is annotated to a dependency tree. In the dependency tree, each dependency relation is also called head-dependent pair represented by an arrow pointing from the head to the dependent. Inter-dependent relation between words is a kind of combinative relation. Relations between different words may exhibit different strengths. The combination strengths measure the degrees of affinities of combinative relation, i.e., higher strengths between words mean they more prefer combine together. The combination strengths are determined by the features of the context of the combinative relations and are the rules dominating the order by which words compose a sentence. Reversely, based on combination strengths between two words in a sentence, the sentence structure, syntax structure or semantic structure, may be created in a bottom-up paradigm [8]. With this understanding, learning and regulating the combination strengths between words is of key importance for sentence structure analysis.

An important analogy between language and the immune system was first made in Jerne's Nobel lecture [13]. The immune system(IS) is considered as a continuous learning system for its ability of adaptation to foreign pathogens [9]. When pathogens, also called antigens (Ags), invade the human body, bone cells (B cells), an important class of lymphocytes in the IS, recognize Ags by their receptors and then undergo a sequence of state changes resulting in higher affinities between the B cells and Ags. Inspired by Jerne's lecture, we make analogous comparison between B cells and words: B cells can only recognize some kind of antigens with specific receptors on the surface of them; and words can only combine with certain words with their properties, which including word-token, word class (i.e. POS), meaning, valency, etc [11]. The analogies between words and B cells guide our research to learn from the immune system. Actually, the human immune system has attracted more attention as an evolution model and has inspired the research of artificial immune systems (AIS) [4].

This research presents a multi-word-agent autonomous learning model (MWAALM) based on an AIS using clonal selection [2] and an idiotypic immune network [12] to regulate strengths of combinative relations between words. First, words are viewed as B cells and antigens and modeled as B cell word agents (BWAs) and antigen word agents (AgWAs). The word combination strengths are represented by the affinities between B cells and are regulated by applying clonal selection and idiotypic immune network theory. Considering the sentence dependency tree bank is available, we narrow the combinative relation to syntax dependency relation. Since the dependency tree of a sentence can be created based on combination strengths, this model is evaluated by a graph-based dependency parsing method using a maximum spanning tree (MST) algorithm [8]. Secondly, cellular automation(CA) [20] is used to construct the model, which is composed of autonomous agents and the environment in which agents reside. This paper presents a completely new perspective on language and proposes an autonomous learning model to learn word combination strengths. The most significant advantage of the proposed model is the ability of continuous learning and the concise implementation method. This model is verified by sentence dependency parsing. Experimental results on CTB indicate that our model can learn word combination strengths effectively and continuously.

J. Yang (✉) • X. Dong • Y. Guan
Web Intelligence Lab, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
e-mail: yangjinfeng2010@gmail.com; dongxishuang@gmail.com; guanyi@hit.edu.cn

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_108,

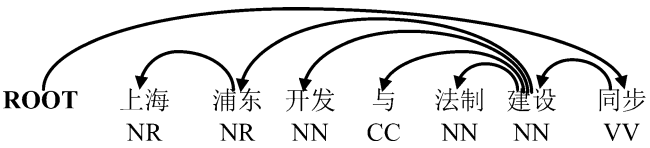


Fig. 1 The dependency structure for a Chinese sentence "上海浦东开发与法制建设同步(Development is synchronized with legal construction in Pudong of Shanghai)".

The remainder of the paper is organized as follows. In Section 2, related work is summarized. An autonomous learning model based on adaptive immune theories is proposed in detail in Section 3. In Section 4, experimental results of the model are presented and analyzed. Finally, the conclusion of the work is given in the last section.

2 Related Work

To the best of our knowledge, this research is the first attempt to investigate the optimization problem of combination strengths between words, and is also the first attempt to try agent modeling method in natural language processing study. Related works including immune-based learning and agent-based modeling are overviewed in the remainder of this section.

Immune-based learning has gained most attentions in the application area of AIS. Immune-based learning mainly involves clonal selection algorithms and immune network algorithms. Clonal selection describes the basic feature of adaptive immune response: only those B cells that recognize antigens proliferate and the offspring may undergo somatic hypermutation resulting in higher affinities with antigens [2]. Clonal selection theory has inspired an unsupervised learning model, such as CLONALG [6], and a supervised AIS classifier, such as AIRS [19]. The idiotypic immune network theory was first proposed by Jerne [12] and formalized into a model by Farmer [9]. In this theory, B cells can recognize or be recognized by other B cells until the idiotipc level reaches the maximum, which leads to the creation of a network among B cells [18]. By employing the metaphor of the immune network theory, unsupervised learning models [5] and supervised AIS classifiers [7] were proposed. For more applications about AIS, please refer to [4, 10].

Agent-based modeling employs large numbers of autonomous agents that interact with each other in an artificial environment. The agents' behaviors are described by rules that determine how they learn, interact with each other and adapt. The agents and environment are generally implemented with a cellular automaton (CA). One of the most referenced and peer reviewed IS simulators called ImmSim was based on CA with probabilistic rules [3]. At each time step, cellular entities in the same CA site can interact with each other stochastically and diffuse through the lattice. C-ImmSim is a version of ImmSim developed in the C programming language, with a focus on improved efficiency and simulation size and complexity [1]. C-ImmSim is the most advanced IS simulator based on the original version; so it is simplified to construct the proposed model.

Hart and Dasgupta proposed that one of future efforts would be to develop distinctive immune inspired algorithms without any logical and technique overlap for any existing techniques [10, 4]. This research exploits the consistency between immune system and language comprehensively and develops an autonomous learning model based on clonal selection and immune network theory. This research is expected to make positive contributions as pointed out by Hart and Dasgupta.

3 Multi-Word-Agent Autonomous Learning Model

The proposed model, MWAALM, aims to learn and regulate the combination strengths between words. Constructed by CA, the model consists of two components: a group of autonomous word agents(BWAs and AgWAs) and the artificial immune environment. In order to train and evaluate this AIS based model, a Chinese dependency Treebank is divided into a training set and a test set. The learning process of this AIS model includes three stages. In the initialization stage, the immune environment and B cell word agents are initialized. The immune environment is initialized as an $M \times M$ grid. BWAs are built from the training set and then are distributed into the grid uniformly. In the learning stage, AgWAs are constructed from a sentence from the dependency Treebank one by one and are injected into the immune environment. With the principles of clonal selection and idiotypic immune network, BWAs and AgWAs interact with each other resulting in higher strengths between BWAs. In the last evaluation stage, the sentences in the test set are structured as dependency trees based on the model and evaluated by computing unlabeled attachment scores(UAS) [14], i.e. the percentage of words that have the correct heads.

3.1 Representation of BWA and AgWA

Representation of BWA. The receptors of B cells have a Y-shaped structure. At the tips of the Y, there exist paratopes and idiotopes. The paratopes are responsible for recognizing the unique set of antigenic determinants of Ags, also called epitopes, and the idiotopes can function as antigens. So, the paratopes of one B cell can also recognize the idiotopes of

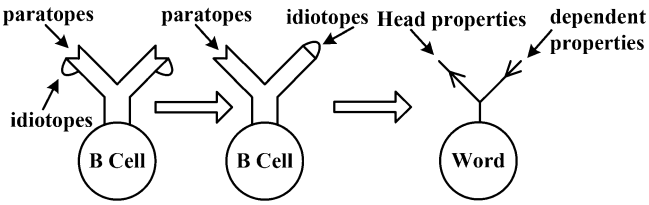


Fig. 2 The design of B cell receptors and word properties

Table 1 Feature templates of dependency pairs

Word(W)	POS(P)	Word and POS
W_h	P_h, P_d, P_h_P_d	W_h_W_d_P_d
W_d	P_h_P_{h+1}_P_{d-1}_P_d	W_h_P_h_W_d
W_h_W_d	P_{h-1}_P_h_P_{d-1}_P_d	W_h_P_h_P_d
	P_h_P_{h+1}_P_d_P_{d+1}	P_h_W_d_P_d
	P_{h-1}_P_h_P_d_P_{d+1}	W_h_P_h_W_d_P_d

another B cell. The two tips of Y are identical. For simplicity in our model, the right tips of the Y are idiotopes and the left tips are paratopes. Accordingly, in the proposed model, properties of words are grouped into dependent properties and head properties. Figure 2 shows the simplified design of B cell receptors and the word agent with head properties and dependent properties. Dependency features extracted from head-dependent pairs of training set, according to the feature templates shown as Table 1, are used as properties of words. For a word w, $\{hf_1^w, hf_2^w, \cdots, hf_i^w, \cdots, hf_{N_{hf}}^w\}$ is the head feature set of w extracted from all head-dependent pairs in which w is the head word, ω_i^w is the weight of hf_i^w. Respectively, $\{df_1^w, df_2^w, \cdots, df_j^w, \cdots, df_{N_{df}}^w\}$ is the dependent feature set of w which comprises features extracted from all head-dependent pairs in which w is the dependent word.

In Table 1, given a head-dependent pair, W_h donates the head word, W_d donates the dependent word, P_h donates the POS of the head word, P_d donates the POS of the dependent word, + 1 donates the right adjacent word, − 1 donates the left adjacent word. For example, 法制 (legal)建设 ← (construction) is a head-dependent pair in the dependency tree shown in Figure 1, and their corresponding POS both are NN tagged below them. Then features of the head-dependent pair include 法制, 建设, 法制_建设, *NN*, *NN*, *NN*_*NN*, etc.

Provided with word properties extracted from dependency relations, the paratopes P^w and idiotopes I^w of a BWA w are formulated as Equation (1) and Equation (2).

$$P^w = \{(hf_1^w, \omega_1^w), (hf_2^w, \omega_2^w), \cdots, (hf_{N_P}^w, \omega_{N_P}^w)\} \quad (1)$$

$$I^w = \{df_1^w, df_2^w, \cdots, df_{N_I}^w\} \quad (2)$$

Representation of AgWA. In each round of learning, one sentence dependency tree is picked from the training set. The dependent word of each head-dependent pair of the dependency tree is used to construct an AgWA and features of the head-dependent pair are used as epitopes of the AgWA. The epitopes E^w of an AgWA w is formulated as Equation (3).

$$E^w = \{df_1^w, df_2^w, \cdots, df_{N_E}^w\} \quad (3)$$

Affinity Measurement. Affinity between two word agents, the same as combination strength between the two words, is calculated based on the similarity between paratopes and epitopes or idiotopes, accumulating weights of the matched properties. Affinity between a BWA w_B and an AgWA w_{Ag} is measured by equation (4).

$$f_{affinity}(w_B, w_{Ag}) = \sum_{i=1}^{N_{hf}^{w_B}} \sum_{j=1}^{N_{df}^{w_{Ag}}} \delta\left(hf_i^{w_B}, df_j^{w_{Ag}}\right) \omega_i^{w_B} \quad (4)$$

$$\delta(x, y) = \begin{cases} 1, \text{if} x = y \\ 0, \text{otherwise} \end{cases} \quad (5)$$

3.2 Cloning and Hypermutation

Both BWAs and AgWAs can move randomly to adjacent sites or stay where they reside. A BWA w can be activated by other neighbor AgWA, which reside in the same site of the grid as w, according to affinity between them. Once w is activated by another agent, it reproduces a group of clones $\{w_1', w_2', \cdots, w_i', \cdots, w_k'\}$ where K is the number of clones. Each clone w' of the BWA w suffers hypermutation individually. The process of hypermutation is that the weight $\omega_i^{w'}$ of each paratope of the agent's receptor is assigned a random increment $\Delta_i^{w'}$ with a certain probability $p_{mutation}$. $\Delta_i^{w'}$ is inversely proportional to the weight and fitness of the agent and also inversely proportional to the affinity between the agent and the recognized AgWA. The mutation is performed according to equation (6):

$$\begin{aligned} \omega'^{w'}_i &= \omega_i^{w'} + \Delta_i^{w'}, \\ \Delta_i^{w'} &= \alpha \times (1/\beta) \times N(0,1), \\ \alpha &= e^{-\omega_i^{w'}} \times e^{-f_{affinity}} \times e^{-f_{fitness}(w')} \end{aligned} \quad (6)$$

where $\omega'^{w'}_i$ is the mutated weightiness, $N(0, 1)$ is a Gaussian random variable of zero mean and standard deviation $\sigma = 1$, β is a parameter that controls the decay of the inverse exponential function, $f_{affinity}$ is the affinity determined by the equation (4), and $f_{fitness}(w')$ is the fitness of each clone

determined by a fitness function. These clones will be evaluated by a fitness function and then the best fit one will be reserved and replace its parent. In the model, two initialization modes are considered to initialize the value of ω_i^w, formulated as a parameter m_{init}. The first mode ($m_{init} = 1$) is to initialize the value of ω_i^w as $(1/\beta) \times N(0, 1)$, and the second ($m_{init} = 0$) is to initialize the value of ω_i^w as zero. The comparisons between the two modes will be investigated in the experiment stage.

3.3 Fitness Function

When a clone w' of the BWA w finishes its hypermutation, the word combination strength between w' and the AgWA may be regulated. The fitness of w' determines whether the word combination strengths are tuned better or worse. The fitness function is designed as a UAS function, formulated as equation (7), for the predicted dependency tree of the training sentence from which AgWAs are built. The dependency tree prediction is implemented by using the MST algorithm [8]. Let S be the training sentence, T be the annotated dependency tree, T' be the predicted dependency tree, then the fitness of w' is measured by the equation (8). Then the clone w'^* which has a maximum fitness value may be reserved and replaces its parent and others are eliminated.

$$UAS = \frac{\#words\ with\ correctly\ assigned\ heads\ in\ the\ predicted\ tree}{words\ in\ training\ set\ in\ the\ annotated\ tree} \tag{7}$$

$$f_{fitness}(w') = f_{UAS}(T, T') \tag{8}$$

$$w'^* = \arg\max_i (f_{fitness}(w_i')) \tag{9}$$

The fitness function of this model is a global measurement for the performance of word strength regulation, which guides the model to evolve towards the desired state in which combination strengths between words are well tuned.

4 Experimental Results

4.1 Data sets and Experimental Design

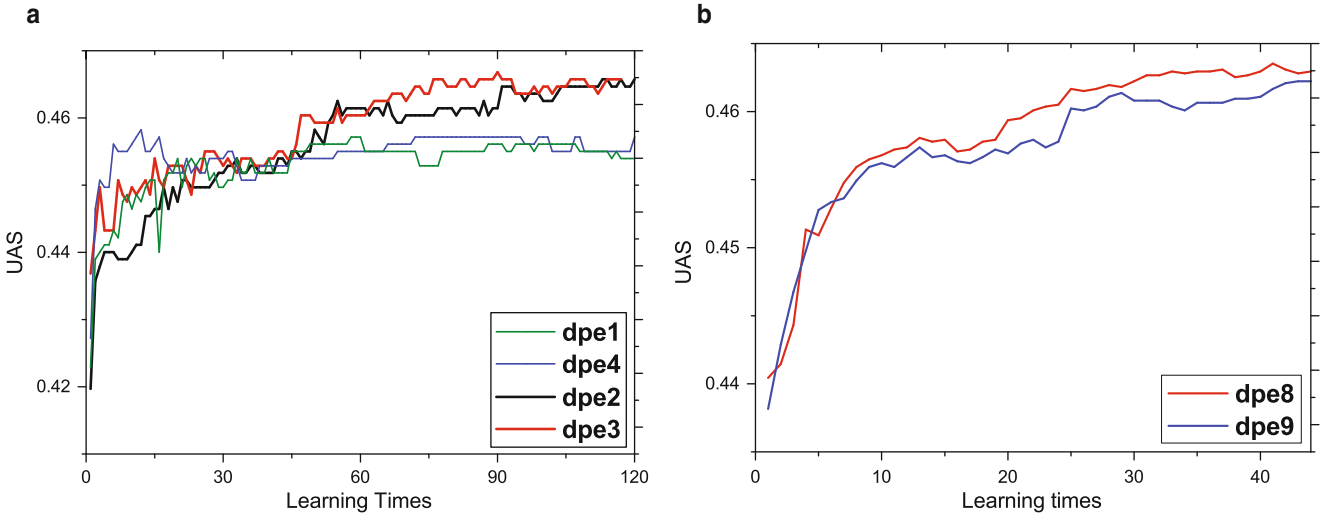
The primary purpose of the experiments is to investigate the effectiveness of regulation of word combination strengths by the proposed model. This model is verified on a dependency parsing task and a dependency Treebank converted from the CTB is employed as experimental data [17]. The dependency Treebank is divided into training data and test data. All words of sentences in the training set were used to initialize BWAs and dependency relations between words were used to initialize the artificial immune network. The performance of dependency parsing of the model is evaluated by the UAS, and also is compared with the MSTParser [15], for which is a graph-based parser and also uses the MST algorithm [8].

Since each word in the sentences of the training set is constructed as a word agent, there are so many word agents in the model that the learning process is time-consuming. A set of preliminary experiments are conducted on a small data set, including 100 training sentences and 50 test sentences, to investigate the performance of the model and the two parameters. Another set of experiments are conducted on a large data set, including 1000 training sentences and 300 test sentences, to verify the observations in the preliminary experiments.

4.2 Results and analysis

The model is evaluated by computing UAS of predicted dependency trees on a test data set when the model finishes a round of learning, then another round of learning follows. With the continuous injection of AgWAs, combination strengths between BWAs can be regulated continuously. Therefore, curves with x-axis values denoting learning times and y-axis values denoting UASs are expected to climb higher and higher with the learning times increments and eventually converges towards a certain level. Experimental results both on a small data set and a large data set are presented below.

Results of on different scale data sets. Four result curves with different parameter values are shown in Figure 3(a). The first group includes curve dpe2 and curve dpe3 ($m_{init} = 1$ in dpe2, $m_{init} = 0$ in dpe3, and other parameters are the same). The second group includes curve dpe1 and curve dpe4 ($m_{init} = 1$ in dpe1, $m_{init} = 0$ in dpe4, and other parameters are the same). These four experiments give evidence that the proposed model can continuously learn and regulate relation strengths between words effectively. Moreover, different weights initialization methods are considered to study whether the weight initialization method can influence the performance of the model. In figure 3(a), the only difference between dpe1 and dpe4 is the initialization method of weights, but the two curves converge the same level. The same effect can be observed in the curve dpe2 and dpe3. Experimental results in Figure 3(a) show that different weights initialization



Results on a small-scale data set. Results on a large-scale data set.

Fig. 3 Results of dependency parsing on different scale data sets

Table 2 Comparisons of UASs between MWAALM and MSTParser.

different scale data	MWAALM	MSTParser
small-scale data(100)	0.4657	0.5846
large-scale data(1000)	0.4629	0.6460

methods will not influence the performance of the model. This advantage indicates that m_{init} can be omitted, and more importantly that a distributed version of the model can be developed.

Preliminary results on the small-scale data set are encouraging. A group of experiments on a large-scale data set are carried out to verify the preliminary conclusions; results are shown in Figure 3(b).For the two curves, $m_{init} = 1$ in curve dpe8 and $m_{init} = 0$ in curve dpe9. The resulting curves also show that the model performs the ability of continuous learning on large-scale data. With less learning times, the performance on the large-scale data set can reach the almost same level as on the small-scale data set.

Comparisons between MWAALM and MSTParser. Comparative experiments were conducted on the same data sets with the same feature templates as shown in table 1. Experimental results of MWAALM and those of MSTParser are displayed in table 2. Both on the small-scale data set and the large-scale data set, MSTParser outperforms MWAALM. As a completely new methods to this classical NLP task, MWAALM's performance is acceptable. The gap of performance does not deny the effectiveness of the proposed model, but indicate that the learning algorithm of the model, mutation mechanisms and the fitness function, should be designed more elaborately.

5 Conclusions and Future Work

This research presents a multi-word-agent autonomous learning model to regulate combination strengths between words based on the adaptive immune theory. The model is evaluated on the dependency Treebank built from the CTB and proven effective. This research provides a completely new perspective on language and words and introduces biological inspirations from the immune system into the proposed model. With a concise and multi-agent modeling method, this AIS-based model achieves the ability of continuous learning and performs effectively for sentence dependency parsing, which is a classical research task of natural language processing (NLP). In the area of researches on NLP, applications of statistical machine learning methods are more prevalent. However, most statistical machine learning methods fail to adapt to new circumstances and lack the characteristic of continuous learning; this disadvantage greatly hampers both researches and applications of NLP. The performance of this model may provide certain inspirations to the researches on NLP, as well as machine learning.

Two aspects of future work will be mainly focused. In this research, words are viewed as lymphocytes and represented as BWAs. This new lymphocyte-style representation is actually a two-vector word representation and has the potential of expressing combinative relation, which is an inherent limitation of existing word representation, such as distributed word representation [16]. So, one of the future work is to investigate lymphocyte-style representation in some classical NLP tasks and make comparisons with existing word representations. As compared with MSTParser, this model seems to be somewhat immature.

The second future work is to improve the learning algorithm of the model, primarily including mutation mechanisms and the fitness function.

References

1. Bernaschi, M., Castiglione, F.: Design and implementation of an immune system simulator. Computers in biology and medicine 31(5), 303–331 (Sep 2001)
2. Burnet, S.: The Clonal Selection Theory of Acquired Immunity, vol. 105. Nashville: Vanderbilt University Press (Jun 1959)
3. Celada, F., Seiden, P.: A computer model of cellular interactions in the immune system. Immunology today 13(2), 56–62 (Feb 1992)
4. Dasgupta, D., Yu, S., Nino, F.: Recent Advances in Artificial Immune Systems: Models and Applications. Applied Soft Computing 11(2), 1574–1587 (Mar 2011)
5. De Castro, L., Von Zuben, F.: aiNet: An Artificial Immune Network for Data Analysis. In: Abbass, H.A., Sarker, R.A., Newton, C.S. (eds.) Data Mining A Heuristic Approach, chap. XII, pp. 231–259. Idea Group Publishing (2001)
6. De Castro, L., Von Zuben, F.: Learning and optimization using the clonal selection principle. IEEE Transactions on Evolutionary Computation 6(3), 239–251 (Jun 2002)
7. Deng, Z., Tan, G., He, P., et al.: A decision hyper plane heuristic based artificial immune network classification algorithm. Journal of Central South University 20(7), 1852–1860 (Jul 2013)
8. Eisner, J.: Three new probabilistic models for dependency parsing: An exploration. Proceedings of the 16th conference on Computational linguisticsVolume 1 96(August), 340–345 (1996)
9. Farmer, J., Packard, N., Perelson, A.: The immune system, adaptation, and machine learning. Physica D: Nonlinear Phenomena 22 (1–3), 187–204 (1986)
10. Hart, E., Timmis, J.: Application areas of AIS: The past, the present and the future. Applied Soft Computing 8(1), 191–201 (Jan 2008)
11. Hudson, R.: An Introduction to Word Grammar. Cambridge University Press (2010)
12. Jerne, N.: Towards a network theory of the immune system. Annales immunologie 125C(1–2), 373–89 (Jan 1974)
13. Jerne, N.: The generative grammer of the immune system. Nobel Lecture 229, 1009–1057 (1985)
14. Kubler, S., McDonald, R., Nivre, J.: Dependency Parsing. Synthesis Lectures on Human Language Technologies 2(1), 1–127 (Jan 2009)
15. McDonald, R., Crammer, K., Pereira, F.: Online large-margin training of dependency parsers. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05. pp. 91–98. Association for Computational Linguistics, Morristown, NJ, USA (2005)
16. Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed Representations of Words and Phrases and their Compositionality. ArXiv Preprint (Oct 2013), http://arxiv.org/abs/1310.4546
17. Nivre, J., Hall, J., Nilsson, J.: MaltParser: A data-driven parser-generator for dependency parsing. In: Proceedings of LREC. vol. 6, pp. 2216–2219. Citeseer, Citeseer (2006)
18. Perelson, A.: Immune network theory. Immunological reviews 110, 5–36 (Aug 1989)
19. Watkins, A., Timmis, J., Boggess, L.: Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm. Genetic Programming and Evolvable Machines 5(3), 291–317 (Sep 2004)
20. Wolfram, S.: Cellural Automaton as Models of Complexity. Nature 311, 419–424 (1984)
21. Xue, N., Xia, F., Chiou, F., et al.: The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. Natural Language Engineering 11(2), 207–238 (Jun 2005)

Agile Rough Set Based Rule Induction to Sustainable Service and Energy Provision

Chun-Che Huang, Tzu-Liang (Bill) Tseng, Yu-Sheng Liu, Jun-Wei Chu, and Po-An Chen

1 Introduction

The concept of sustainable service systems is emerging and becomes more and more critical recently [1]. Sustainable service and energy refers to the service offering the tasks embedded with economic, social and environmental protection elements that are so far still non-existence especially to the existing traditional, conventional and competing offers in the market [2]. Providing sustainable service and energy has been becoming a trend due to environmental concerns. It is also because the service concept of sustainable development, not only can generic demand, and guide green consumption, but also can create of the green benefits [3]. In recent years, despite expansive literature could be found, contributions to its implementation are currently limited not yet emerging.

One of the academic challenge is: Obtaining pertinent, consistent and up-to-date information across a large region is a complex and time consuming process since data are dynamically changed from time to time. Due to such situation, rule induction to support decision making is required to be agile and effective. In addition, in the complex service sector, those data (e.g., from questionnaires) may be complicated, qualitative and in large scale. Moreover, numerous attributes which have the impact on service performance are involved. A sustainable service must take many attributes into account and evaluate how maximum benefits could be provided through the use of different indicators. Numerous indicators of sustainability have been studied for many years and have been identified as desirable instruments and/or measuring rods to assess and monitor progress towards sustainable development [4]. Decision making may redundantly utilize these indicators. Consequently, a rule induction approach to reduce the number of attributes which is derived from indicators and amount of data is required.

In data mining, rule induction approaches are numerous [5]. The main difference between the rough set based induction and other methods is that these approaches are not suitable in selection of "sustainable attribute" because they are with population based approaches which may require several statistical assumptions and view the solution approach as a black box, and they have limitation to handle qualitative type of data [6]. One of the promising solution approaches to the aforementioned challenges is Rough Set (RS) based approach can deal with qualitative information and provide an individual object based approach [6] which the relationship between each object and the rule can be recognized. However, traditional RS approaches have four disadvantages: (1) Most of previous studies on rough sets focused on finding certain rules and possible rules that a decision attribute is in one level only. However, hierarchical attributes are usually predefined in real-world applications [7]. In practice, the concept of each attribute could be hierarchical in nature. It is required to take into account the overall hierarchy of attributes, both condition or decision attributes. (2) Previous RS approaches used two stages to generate reducts and induct decision rules, respectively. Large computing space is required to store the reducts from the first stage, and solution searching is complex. Moreover, comparison of the reducts is limited to the same decision attribute and to the same number of attributes selected in the reducts. When the number of attributes is more, the strength index, which is introduced to identify meaningful reducts, is larger. This conflicts with the definition of a reduct, a minimal subset of attributes that provides the same descriptive ability as the entire set of attributes.

C.-C. Huang (✉) • Y.-S. Liu • J.-W. Chu • P.-A. Chen
Department of Information Management, National Chi Nan University, Taiwan No. 1, University Road, Pu-Li 545, Taiwan (R.O.C.)
e-mail: cchuang@ncnu.edu.tw; S96213016@ncnu.edu.tw; ppig0117@gmail.com; annieab.chen@gmail.com

T.-L. (Bill) Tseng
Department of Industrial, Manufacturing and Systems Engineering, The University of Texas at El Paso, 500 West University Avenue, El Paso TX 79968, USA
e-mail: btseng@utep.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_109,

In this paper, an exntended RS based rule induction is proposed to induct decision rules and handle the four disadvantages.

Next, the hierarchical rough set problem is defined. To resolve this problem, the extended solution approach is proposed in Section 3 and Section 4 concludes this paper.

2 Hierarchical rough set problem

In this study, the hierarchical rough set problem is defined as:

Given: (1) An information system $I = (U, A)$, where U is a finite set of objects and A is a finite set of attributes. The elements of A are called condition attributes. (2) A hierarchical transportation decision table $I = (U, A \cup \{d\})$, where $d \notin A$ is a distinguished hierarchical outcome attribute at a different level. (3) A concept hierarchy Pk refers to a set of domains $Ox,\ldots,Oz$. Pk_{sk}:$\{Ox \times \ldots \times Oz\} \rightarrow Pk_{sk-1} \rightarrow Pk_1$, where: Pk_{sk} denotes the set of concepts at the the sk^{th} level, Pk_{sk-1} denotes the concepts at one level higher than those at Pk_{sk}, Pk_1 represents the top level denoted as "ANY".

Objective: *Minimize* the subset of U, in which the specific outcome attribute of decision rules (r_i) has the highest Strong Index (SI) and the maximal level number.

Subject to the following constraints:

(1) $A = C \cup D$,
(2) $B \subset C$,
(3) $POSB(D) = \{x \in U: [x]B \subset D\}$,
(4) For any $a \in B$ of C, $K(B, D) = K(C, D)$, and $K(B, D) \neq K(B-\{a\}, D)$,

where, C is the condition domain and D is the outcome domain.

POSB(D) is the positive region and includes all objects in U which can be classified into classes of D, in the knowledge B.

$K(B, D) = \frac{card(POSB(D))}{card(POSC(D))}$ is the degree of dependency between B and D

Attribute $a \in A$, set of its values

3 The Solution Approach

In this study, the concept hierarchy is considered in both condition and decision attribute based on [8]. Suppose a concept hierarchy PI^D refers to a set of decision attribute domains $D_i,\ldots, D_k$. $PI^D{}_{ll}$:$\{D_i \times \ldots \times D_k\} \rightarrow PI^D{}_{ll-1} \rightarrow P^D{}_1$, where $PI^D{}_{ll}$ denotes the set of concepts at the ll^{th} level, $PI^D{}_{ll-1}$ denotes the concepts at one level higher than those at $PI^D{}_{ll}$, and $PI^D{}_1$ represents the top level denoted as "ANY". D_x refers to the value of attribute = x at level 1. $D_{x.y}$ refers to the value of attribute = x at level 1 and y at level 2. $D_{x.y.z}$ refers to the value of attribute = x at level 1, y at level 2, and z at level 3, …, etc. I refers to decision attribute hierarchical index. ll refers to the number of levels in the PI^D.

$D_{x.y.z} \in D_{x.y} \in D_x$, implying more levels will have more specific information. And, in the concept tree, D_x is father node of $D_{x.y}$. $D_{x.y}$ is a son node of D_x. $D_{x.y}$ and $D_{x.n}$ are the brother node for each other. In the concept tree, the most general concept can be a universal concept, whereas the most specific concepts correspond to the specific values of attributes in the data set. Each node in the tree of the concept hierarchy represents a concept. For example, $D_{X.1} \cup D_{X.2} = PX_2 \in D_X$, $D_{Y.1} \cup D_{Y.2} = PY_2 \in D_Y$, $D_X \cup D_Y = P_1 \in D_{ANY}$.

In this study, the procedure starts from the bottom nodes to further search and terminate such search based on three principles: (1) L_l is more informative but less coverage cardinal number than L_{l-1}. (2) When the data sets of condition attribute are conflicted, one should move up in the hierarchy of decision attribute. Terminate exploring until that data set of decision attributes are not conflicted. (3) The nodes whose attributes have small number cardinality less than Minimal Cardinal Number Index (MCNI) and have the same father node should bemerged, and then they should be moved up to their father nodes to generate reducts.

Searching the desired reducts (rules) which are comprised of the maximum of Strength Index (SI) is a way to rule extraction, where $SI(f) = \left(\sum_{j=1}^{m} vjWjxnf\right) / \left(\sum_{j=1}^{m} vj\right)$. In the implementation of algorithm, after the pre-process steps, select a merged reduct or an individual reduct with the highest SI form the objects under the assumption that the number of object cardinality is greater than N_{min}, which refers to the minimum number of objects for the reducts qualified to become as a rule. If the reduct with highest SI is not greater than N_{min}, then the reduct is excluded in the decision rule set. Repeat the aforementioned steps (1-14) until all objects are used then terminate the Extended Rules Induction Algorithm (ERIA) procedure.

Notation:

O: Object
SN: Support Number
t: reduct index
$O_{(SN)}$: Support Number of object
$A_{.cond}$: the condition attribute (feature)
$A_{.out}$: the outcome attribute
OC: the object have the same conditional attributes, but different decision attributes.
$OC_{(SN)}$: supper number of the object have same conditional attributes, but different decision attributes.
m: the total number of condition attributes.
n: the total number of outcome attributes
On: Minimal Support Number Index (MSNI)

sf: same parent node but not same child node

sfC: same parent node but not same child node of condition attributes

$O_{(sfC)}$: the object have same parent node but not same child node of condition attributes

$\ddot{O}$: smeller differences $O_{(sfC)}$

$\ddot{O}_{(A.cond)}$: the number of same attributes of $\ddot{O}$

L: level index

O_L: level index of object

Input: *MSNI* value.
Output: The set of decision rules and alternative rules.
Step 0 Initialization
  Select the block of the data that needs to be analyzed.
  Entering the *On* of the *MSNI* value.
Step 1 Find the complicit object *OC*.
  For $O = 1$ to t
Step 1.1 If any objects each other that have same $A_{.cond}$, but different $A_{.out}$
  Set *OC* then go to Step 1.2
  Else if
  go to Step 2
  End if
  End for
Step 1.2 Combine the *OC*, Change the $OC_{(A.out)}$.
  For each *OC* set
Step 1.2.1 If $m > 1$
  Delete the most law data of the $OC_{(A.out)}$.
  Check each set $OC_{(A.out)}$ are same.
  If each set $OC_{(A.out)}$ same
    Combine each set $OC_{(CN)}$, then go to Step 2.
  Else if
    go to Step 1.2
  End if
Step 1.2.2 If $m \leq 1$
  Delete the *OC*
  End if
  End if
Step 2 Set the value of *On* to avoid too detail information.
Step 2.1 Determine the Entering *On* value,
  For $O = 1$ to t
    If $O_{(CN)} < On$
      If have $O_{(sfC)}$
        Go to Step 2.1.1
Step 2.1.1 Do combine $O_{(sfC)}$
  (According the same $A_{.out}$, Sequence the $O_{(sfC)}$ and set the classification)
  In each classification, find $\ddot{O}$ of $A_{.cond}$
  For $\ddot{O}_{(A.cond)} = n-1$ to 1
    Delete the most law *L* data of the $O_{(sfC)}$
    Combine the *CN*
    If reach the *On*
      Recheck if any objects have *OC*
If have *OC*
        Go to Step 1
      Else
        go to Step2.1.2
      Else
        go to Step 2.1
      check if have $O_{(sfC)}$
      End if
    Else
      delete the object
      then go to Step 3
    End if
  Else go to Step3
  End if
  End if
Step 3 Run *RST*, and generate the total rule.
Step 4 Check all rules there is no objects have less information level than the other.
  If $O_L \geq O_{L-1}$
    Delete O_{L-1}
  End if
Step 5 Stop and output the decision rule.

Illustrative example: The primary goal of developing green sources of energy is to generate electronic power while minimizing both waste and pollution, to thereby reduce the impact of energy production on the environment, and also to maintain the stability of electronic power supply with desired electronic distribution, control, and storage. To achieve the goal, obtaining pertinent, consistent and up-to-date information across a large administration is a complex and time consuming process. Due to such situation, rule induction to support decision making is required to be agile and effective because data are dynamically updated at the implementation stage. In a particular region, the green energy hierarchy concept of decision table is provided in Figure 1, where *A1-A4* refer to condition attributes: Sustainable indicators, social, economic, environment, respectively (from the sustainable service perspectives [9]). *O* refers to decision attribute, control parameters of Green Energy. The condition attributes on level 1 refer locations and those on level 2 refers to types of energy, e.g. hydro, fire, wind, geothermal, solar, nuclear.

With the proposed solution approach, 18 decision rules are induct, and only 8 rules are induct by the tradition RS approach (without hierarchical concept). The coverage comparison of similar 8 rules are presented in Table 1, where

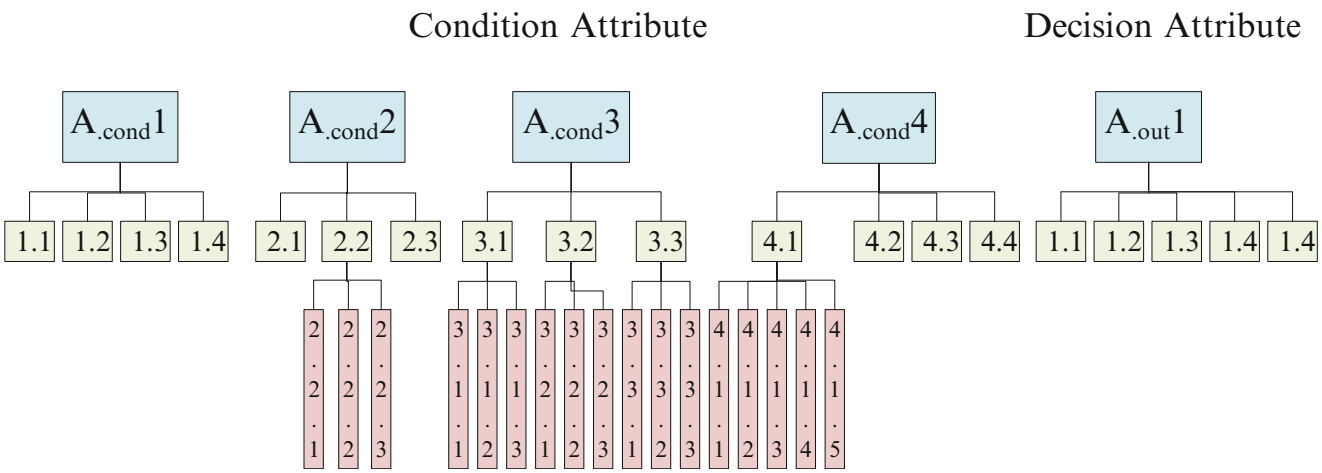


Fig. 1 The concept hierarchy.

Table 1 The comparison of accuracy and coverage between the proposed and tradition RS approaches

Rule No.	1	2	3	4	5	6	7	8
ERIA	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
Tradition RS approach	75 %	33.3 %	25 %	75 %	75 %	100 %	33.3 %	33.3 %

shows the proposed solution approach can provides more detailed rules and higher coverage than tradition RS approach. The rules aims at supporting managers to determine resource re-allocation with low cost such that the level of the desired control parameters.

4 Conclusion

In this paper, the proposed solution approach extract significant attributes and inducts decision rules for sustainable service and energy. Comparing to the traditional RS based approach, the proposed rough set approach can induct decision rules that decision tables are with concept hierarchy. The proposed RS approach resolves the problem of hierarchical data and provides a great promise application of RS in different types of service and energy systems. With the proposed approach, the decision makers are expected to complete jobs more efficiently due to all decision rules are derived to follow the green policy.

The further study requires a case study to test the validity. That is, to use different data sets to prove the correctness of the model. In the other words, the generated rules should be capable of predicting the correct outcomes from different data sets. In addition, in a decision table, the importance and separation degree between two levels are crucial. It aids to rule induction by leaning importance of each hierarchical level if the detailed information are in need.

References

1. Roy, R. Sustainable product-service systems. Futures, 32(3–4), 289–299 (2000).sectors/tourism/documents/communications/commission-communication-2007/index_en.htm
2. Dyllick, T., & Hockerts, K. Beyond the business case for corporate sustainability. Business Strategy and the Environment, 11(2), 130–141. (2002).
3. Pan, H., & Ren, J. Discuss about Sustainable Development Service Concept of Tourism Industry. 2011 IEEE International Conference on Management and Service Science (MASS), Wuhan, 12–14 Aug. 2011, 1–4. (2011).
4. Selman, P. Three decades of environmental planning: What have we really learned: London: Routledge. (1999).
5. Han, J., & Kamber, M. Data mining: concepts and techniques (Vol. 5). San Francisco, CA, itd: Morgan Kaufmann. (2001).
6. Kusiak, A. A feature transformation methods in data mining. IEEE Transaction on Electronics Packaging Manufacture, 24(3), 214–221. (2001).
7. Hong, T.-P., Liou, Y.-L., & Wang, S.-L. Fuzzy rough sets with hierarchical quantitative attributes. Expert Systems with Applications, 36(3, Part 2), 6790–6799 (2009).
8. Pan, Z., Tang, J., & Fung, R. Y. K. Synchronization of Inventory and Transportation Under Flexible Vehicle Constraint: A Heuristics Approach Using Sliding Windows and Hierarchical Tree Structure. European Journal of Operational Research, *192*(3), 824–836 (2009).
9. Shaharudin, M. R., & Zailani, S. Sustainable services in Closed Loop Supply Chains (CLSCs). 2011 IEEE Colloquium on Humanities, Science and Engineering (CHUSER) (2011).

Intelligent Web Application Systems Testing through Value Based Test Case Prioritization

Abdul Rauf and Adel Ibrahim AlSalem

1 Introduction

Websites entail some exclusively novel challenges in this era of software quality focus. Inside hours, users of a web based application increases by thousands numbers comparing to a conventional, non-web application. A unique core characteristic of web applications is dynamic nature of these applications. This dynamic nature of web applications make it possible to respond to users based on the users' inputs and to allow users to affect the business logics on the servers. These applications are normally based on enabling technologies and gets evolve from web sites or web systems, and this combination of heterogeneous nature of web pages makes web applications quite complex. While ensuring the quality of a web application, the major challenge that arises is of this complexity and heterogeneous nature of web applications. Although web testing is an effectual process to make sure the excellence of the web applications, but these unique characteristics of web applications make it impossible to use the traditional software testing techniques and methods to be applied. Distinctiveness and complex nature of web applications is the major hindrance in direct application of the traditional testing theories and methods. What's more, web applications have the vibrant, interactive and indecisive characters e.g. in a form page, if users input different contents, the output pages, which are dynamically generated, are usually different. Therefore, during the process for the test case generation, we must consider all the possible output and the successive order of the executed actions, so as to determine the testing steps, the input data, the expected output and the relationships among themselves. On the other hand, propinquity of the web is increasing the rate of prospects regarding quality and fast delivery of the application, but the technical challenges of a website and variations available kinds of the browser are making it very difficult to have complete testing and quality control for web applications, and in some ways, more restrained, than "conventional" client/server or application testing. This kind of issues gives an opportunity as well as posing a challenge to have an automated efficient testing.

The testing of web based applications has much in common with the testing of desktop systems like testing of functionality, configuration, and compatibility. There are various non-equivalence issues between traditional software testing and web application testing. There are various non-equivalence issues between traditional software testing and web application testing. Some of the issues are:

- Web applications maintenance faster rate than other software systems.
- Web applications have a huge & versatile user population
- Unexpected state change
- Web application multitier architecture feature
- Web applications were hosted on clients where there are different operating system and browsers
- Navigation mechanism implemented by hyper textual link demands a lot of verification that no unreachable, no misleading link is included in the application
- Asynchronous behavior
- Numerous technologies, languages, and unknown components

Automated testing techniques contain many advantages and benefits. In manual testing it takes a lot of time in testing but in automated tools can run test suite faster by reducing the human intervention times. Automated test tools are being considered more reliable as they test the same operations repeated again and again (Regression testing). Another obvious advantage of automated testing is reusability of test suite. Comprehensiveness is another benefit of automated testing as it can contain a group of tests that covers each and every feature in the application. Automated testing helps to improve the quality of the test effort as well as to minimize

A. Rauf (✉) • A.I. AlSalem
College of Computer and Information Sciences, Al-Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia
e-mail: rauf.malik@ccis.imamu.edu.sa; alsalem@ccis.imamu.edu.sa

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_110,

the test schedule. Finally, it reduces cost. There have been many tools and techniques employed for verification and testing of these hydrogenous kinds of web applications. Recent research has shown that website testing can be made automated and easy to some extent by using algorithms and techniques of artificial intelligence field.

The application of metaheuristic search techniques belonging to artificial intelligence (AI) domain into software engineering (SE) is known as Search Based Software Engineering (SBSE) [1]. A growing interest can be seen today to bring research directions of both disciplines (AI and SE) closer and such efforts are now building new research areas. With the application of artificial intelligence technique in software engineering and testing, there will be emerging the zone of a study which brings near the cross ertilization of the ideas from these two domains [2, 5]. It is resource consuming and infeasible to adequately test the web applications. More over it is very difficult and expensive to automate web applications testing. By employing SBSE, we can make possible the feasible automation the testing process for web applications and minimize the consumption of resources.

This paper provides the initial results of a test automation framework specifically using the concepts of the value. The framework using the artificial intelligence techniques with the concept of value based testing. Value based testing revolves around the concept of testing the feature/elements based on the value of those parts. This value is further used to prioritize the test cases and making testing feasibly economically depending on the requirements and scenario.

The paper is organized as follows. Section II contains a brief literature review. Implementation and relevant initial results are presented in (Section IIII). Finally, Section IV ends the paper with several conclusions drawn from the design and the work with the proposed system.

2 Related Work

Testing and modeling a web application is more complex than the traditional application [11]. Many types of web application methodologies have been used. Most of these focus different aspects of validation. B Kam and T R. Dean listed six potential web application problems [4]. Each researcher has focused on different areas of web applications like static testing, testing based on dynamic pages, links and frames, architecture, model and scenarios etc [3].

In the field of search based software engineering, there is huge contribution of Mark Harman who first coined this term in 2001 [5]. In 2007 [6], he gave an overview of current state of widely used search based optimization techniques and discussed the open areas, challenges and gaps for future work in SBSE. Later on in 2009 [7] he with companions conducted a complete systematic literature review on SBSE. In this SLR, he conferred the importance of search based techniques in every field of software engineering collectively and discussed trends and relationships between the techniques and the applications to which they have been applied. Leo Rela [8] also performed a study on evolutionary algorithms used to solve different search and optimization problems in the area of software engineering.

59 % work of SBSE done on software testing [10]. Hence, there is a lot work in Software testing. McMinne [9] has surveyed the application of Meta heuristics search techniques to software test data generation. Ali et al, [10] conducted systematic literature review (SLR) on Search based test case generation (SB-TCG) in which he discussed the application of SB-TCG through empirical evidence present in literature. W. Afzal performed systematic mapping and SLR [11] on search based software testing for non functional system properties and pin-point challenges in the application of search based techniques for testing the non-functional properties of the system.

Optimization of test cases is very important in software testing and there is a lot of work in literature related to it, as Zehng Li [4] et al discussed and compared different search based algorithms for test case prioritization. And L. Camila [13] performed comparison between NSGA-II, MoCell, SPEA2 and Random algorithm. Here comparison is done on the bases of requirements coverage and execution time. On the base of Requirements coverage; Random testing is worst whereas, SPEA2 seems to have a better performance for smaller search spaces, while NSGA-II and MOCell have better performance for larger search spaces and on time bases; Random algorithm had the best lowest execution time. The second best execution time was obtained by MOCell, followed by NSGA-II. SPEA2 was slowest algorithm.

Now, identify the process that used: select test tool, define the area within software (scope), plan and write the scripts, develop test suits and execute test. After that I mention some of automated testing tools that I read in papers like (1) ”IBM Rational Functional” is functional testing and regression testing tool that uses java language to code and uses excel as database to store data in web application to be tested (2) ”Selenium” is record and playback testing, it implemented like Firefox extension and allows to record, edit, and debug tests also it converts the data recorded into XML and generates report for comparison of data (3) ”QuickTest” Professional provided by HP/Mercury Interactive also it uses visual basic scripting language to build its flows. It integrates easily with other Mercury testing solutions. It can be used for both types of Testing (4) ”Sahi” it is record and playback scripting for web applications. It developed in Java and JavaScript (5) “AsT” enable the users to automate regression tests and create a test case in Excel sheet. It consist of different methods like it creates easy manual

of tests, execute of test scripts and the comparison of test results with expected results. (6) "JUnit" and "MuJava" the definition of "JUnit: it is open source testing framework for Java it was written by "Erich Gamma and Kent Beck". It is used for white box testing. About this two tools in the experimental of paper [12] for generating high quality test cases was checked, they used the tool in two groups of students some of them use java tool and the others do it manually the final result is the students using tool has better results than using manually.

3 Implementation and Experimental Results

Most of the available tools are working on the individual functionalities needed by software testing teams i.e. Hyperlinks management, test case generation, search engine optimization or test results reporting etc.. Very few of the available tools are providing an optimized bundled number of functionalities that testing teams need mostly like test case generation, test suite optimization and running of test suite depending on different constraints i.e. time and cost for a certain project. This framework is being developed with them aim to answer the following research questions and then develop a framework to implement these results.

- Determine the possibility of a single technique for modeling all diverse nature of web application?
- Evaluation of different model based techniques for test case generation of web applications?
- Evaluation of different AI based algorithms like Genetic Algorithm, Particle Swarm Optimization and Ant Colony optimization for generating test suite for web application.
- What are different factors that are affecting the performance of these algorithms?
- Using concepts of value based software engineering, determine the appropriate size of the test suite and level of testing considering different constraints.
- Determine and apply the best AI based algorithms for test case optimization?

We have implemented the proposed system by using the MATLAB environment. Practical Swarm Optimization (PSO) has been used for initial verification of the produced results. A website has been developed to do the experiment with.

To use the concept of value based prioritization, we estimated the value of the test cases in two different ways. Initial calculation was made based on the customer value calculated by using the Customer Value based Partitioning Decision (CVPD) proposed by Neunghoe Kim et.al [9]. First phase of tests were run based on this formula, while in the next iteration, we embedded the results of first iteration in CVPD and revised the formula. The values that were taken from first iteration to embed in CVPD were the test case efficiency, test coverage and fault detection capability. TRate of fault detection has been determined by APFD metric has been used. This metric is developed by Elbaum et al. [8] that measures the average rate of fault detection per percentage of test suite execution. The APFD is calculated by taking the weighted average of the number of faults detected during the run of the test suite.

Formula:

$$\text{APFD} = 1 - \frac{(TF1 + TF2 + \ldots + TFm)}{nm} + \frac{1}{2xn} \quad (1)$$

Where,

T - > The test suite under evaluation

m - > the number of faults contained in the program under test P

n - > The total number of test cases

TFi - > The position of the first test in T that exposes fault i.[8].

This metric assumes that test cost and fault severity are uniform. Although this technique can yield the results itself, but to increase the confidence we have used the results of this technique embedded with other factors and test cases has been prioritized by PSO. Following diagram represents the block diagram of value based test case prioritization using PSO.

Following graphs shows the number of faults detected by using only CVPD values given to PSO Algorithm, CVPD and AFPD values given to PSO as seed. In the end there is a comparison of results provided in tabular form.

Results have shown two kind of improvement with respect to defect detection. First there is a marginal improvement in traditional testing results when evolutionary testing (Use of PSO) has been introduced. Also an improvement in results with the inclusion can be seen with the inclusion of AFPD factor in the calculation of value. Results in all cases have shown a rapid increase in efficiency and then there is a marginally consistent performance. One reason behind this can be that most of the faults have been detected earlier by the test cases, and the newer test cases have very less probability to find new defects/faults.

Conclusions and Future Work

We have described an automatic method for prioritization of test cases based on the concept of value using evolutionary testing. This is in progress research and the effort aims to answer all the questions raised in section 03, while this paper has addressed only a part of the ongoing research effort. Modeling of websites and comparison with other evolutionary techniques has not been made part of this effort, and will be published later. This research part has only focused on determination of effect of value based testing in comparison to the traditional testing.

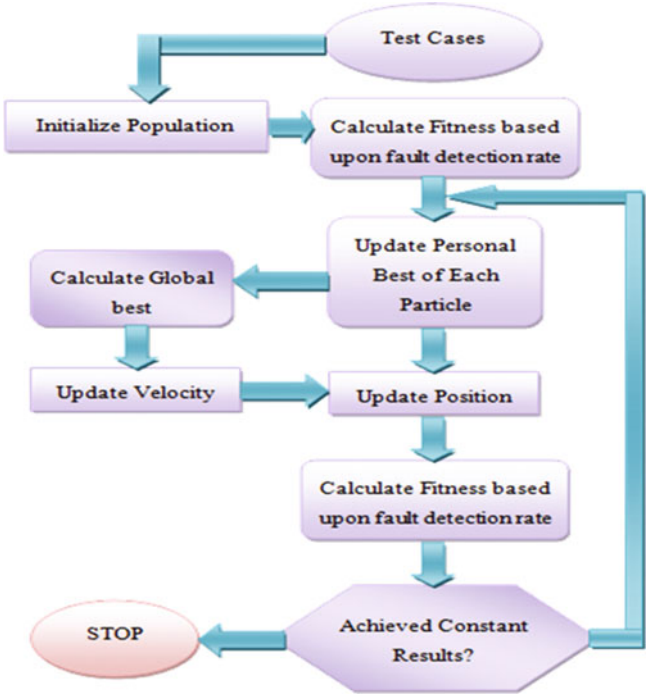


Fig. 1 Block Diagram of PSO for VBTCP

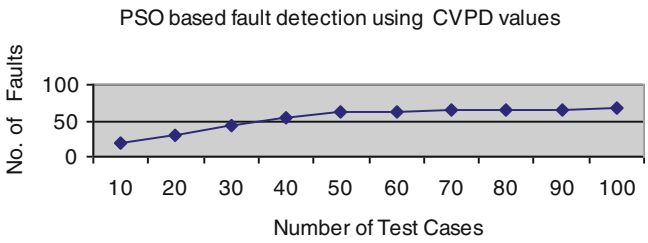


Fig. 2 Results using CVPD Values

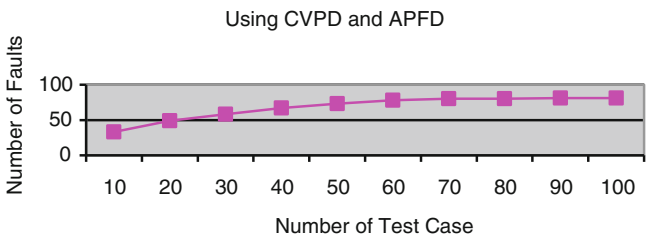


Fig. 3 Results using CVPD and APFD Values

Table 1 Comparison of Results

Number of Test Cases	# of Faults (without Using PSO	# of Faults (CVPD Values Given to PSO)	# of Faults (CVPD and AFPD Values Given to PSO)
10	12	18	33
20	17	29	49
30	23	43	58
40	27	55	67
50	29	61	73
60	34	62	78
70	37	66	80
80	38	66	80
90	41	66	81
100	42	67	81

References

1. B. Boehm and L. Huang, "*Value-Based Software Engineering: A Case Study*" *IEEE Computer*, vol. 36, pp. 33-41, March 2003.
2. S. Elbaum, A. Malishevsky and G. Rothermel, "*Test Case Prioritization: A Family of Empirical Studies*" IEEE Transactions on Software Engineering, vol. 28, pp. 159–182, February, 2002.
3. B.Beizer, "*Software Testing Techniques*" International Thomson Computer Press, 1990.
4. S.Berner, R.Weber and R.K.Keller, "*Observations and Lessons Learned From Automated Testing*" in Proceedings of the 27th International Conference on Software Engineering (ICSE '05), pp. 571–579, St. Louis, Mo, USA, May 2005
5. A. Rauf, S. Anwar, N. Kazim Khan, A. A. Shahid, "*Evolutionary based Automated Coverage Analysis for GUI Testing*", Communications in Computer and Information Science (Springer) ISSN: 1865-–0929
6. Favaro, Favaro, K. R., Favaro, P. F "*Value-based Reuse Investment, Annals of Software Engineering*", (1998)
7. B. Boehm, B. W "*Value-Based Software Engineering. Software Engineering Notes*", 28(2):2003
8. Z. Li, M. Harman, and R.M.Hierons "Search Algorithms for Regression Test Case Prioritization",IEEE Transaction on Software Engineering, VOL. 33, NO. 4, APRIL 2007
9. Neunghoe Kim; Taek Lee; Donghyun Lee; Keun Lee; Hoh Peter In, "Customer Value-based HW/SW Partitioning Decision in Embedded Systems," Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD '08. Ninth ACIS International Conference on, vol., no., pp.257,262, 6–8 Aug. 2008
10. Harman, Mark; Jones, Bryan F. (2001-12-15). "Search-based software engineering". Information and Software Technology 43 (14): 833–839. doi:10.1016/S0950-5849(01)00189-6. ISSN 0950–5849. Retrieved 2013-10-31
11. Y. Fazlalizadeh, A. Khalilian, M. Abdollahi Azgomi and S. Parsa "Prioritizing Test Cases for Resource Constraint Environments Using Historical Test Case Performance Data" IEEE2009
12. Abdul Rauf, Sajid Anwar, Arfan Jaffar, Arshad Ali Shahid, "Automated GUI Test Coverage Analysis using GA", 7th International Conference on Information Technology New Generations (ITNG 2010) Las Vegas, Nevada, USA, April 12-14, 2010
13. K. R.Soffa"Time Aware Test Suite Prioritization"ISSTA'06, July 17–20, 2006, Portland, Maine, USA.
14. E. Wong, J. Horgan, M. Syring, W. Zage, and D. Zage, "Applying design metrics to predict fault-proneness: a case study on a large-scale software system," Software Practice and Experience, vol. 30, pp. 1587–1608, 2000.
15. Standish.Group, "CHAOS." http://www.standishgroup.com/chaos.htm.
16. G.Mogyorodi, "Requirements-Based Testing: An Overview," 39th International Conference and Exhibition on Technology of Object-Oriented Languages and Systems, Santa Barbara, California, pp. 286–295, August 2001.

Iterative Hybrid Identification of Spatial Bilinear Models in the Presence of Uncertainty

James E. Trollope and Keith J. Burnham

1 Introduction

In many engineering problems where control system design is necessitated there is a need for a mathematical model of the plant or sub-system to be controlled. Often, in practice, the systems are nonlinear and it is normal to assume, where possible, that local linearity holds. In some cases, however, it is not possible to assume linearity and it becomes necessary to adapt a nonlinear approach.

Whilst there is an abundance of literature on the topic of linear systems there are rather few articles on nonlinear systems, with each nonlinear system requiring a unique approach. There are, however, some useful categorisations of nonlinear phenomena, and the class of bilinear systems is one such category with a wide variety of practical engineering applications, see (Burnham 1991) and (Mohler 1973).

To establish whether or not a system may be approximated by a bilinear model a number of tests or experiments may be conducted. Some of these tests are equally applicable to other forms of nonlinear systems, so the approaches are not restrictive. These tests include multiple step responses, single and multiple sinewave responses as well as the more general frequency response testing, including single and multi sinewave signals.

Models obtained may be either in continuous-time or in discrete-time and may also involve spatial variables as well time, e.g. diffusion and mechanical compression. The motivation for the paper stems from the latter, where it is known that as the frontal crash structure of a vehicle will buckle under axial compression, its stiffness and damping (assumed to be proportional) will change as a nonlinear (bilinear) function of the deformation. Depending on the geometry of the structure, the greater the compressible deformation and the greater or lesser the stiffness and damping become. Due to the fact that the deformation takes place over an extremely short period of time e.g. of the order 400 microseconds, it is convenient to consider the stiffness, here proportional to damping, to be dependent on deformation only. Consequently attention is restricted here to spatially dependent quasi-static bilinear models of the form

$$f = f_p + \gamma\delta \pm \eta\delta f \tag{1}$$

where f denotes the axial force, δ denotes the axial deformation and f_p denotes a buckling threshold force, which defines the point of buckling and γ and η are the coefficients of the linear and bilinear terms, respectively. A typical force versus displacement characteristic is given in Fig. 1. Note that only the portion between the dashed vertical lines is considered.

2 Concept of active buckling control

Under ideal active buckling control the characteristic shown in Fig. 1 will change dependent on the mass compatibility of two colliding vehicles, thereby requiring tuning of the quantities f_p, γ and η in the longitudinal members of each vehicle. This is achieved to ensure that the energy absorption between the vehicles is appropriately managed in a controlled manner; with use being made of modal control (Trollope and Burnham 2013a) in order to uncouple the structural dynamics of the two colliding vehicles. The area under the bilinear curve given by (1) relates to the collision energy absorbed, so that an accurate representation is required in order to share the energy between the colliding vehicles accordingly.

In previous work of the authors, the collision energy compatibility mitigation problem has been described, see (Trollope and Burnham 2013b, 2013c) for details.

Consider for example, a collision between a moving and a stationary vehicle of dissimilar mass, whereby the larger

J.E. Trollope • K.J. Burnham (✉)
Control Theory and Applications Centre, Coventry University, Coventry CV1 5FB, UK
e-mail: james.trollope@coventry.ac.uk; k.burnham@coventry.ac.uk

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_111,

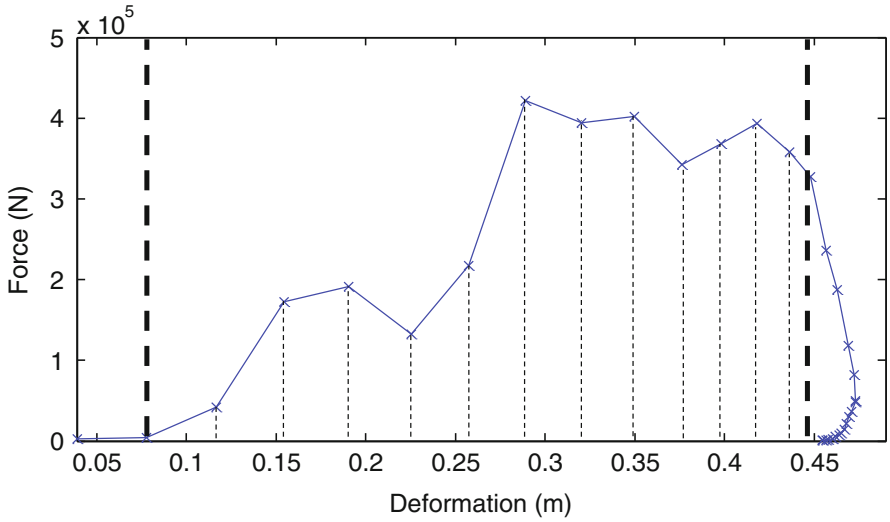


Fig. 1 Typical force versus deformation characteristic of a vehicle frontal collision

moving vehicle, initially travelling at a velocity given by $V_a = 12m/s$, collides with a smaller stationary vehicle given by $V_b = 0m/s$. Denote the mass of the vehicles as m_a and m_b given by 1000kg and 500kg, respectively. Denote the final velocity i.e. after the collision as V_f. It is well known that the conservation of momentum can be expressed as

$$m_a V_a + m_b V_b = m_{(a+b)} V_f \tag{2}$$

where, $m_{(a+b)} = m_a + m_b$. It can be deduced that the final velocity of the combined mass of the two vehicles is 8m/s.

From the principle of conservation of energy, the kinetic energy before and after the collision must be equal, consequently the following relationship occurs

$$\frac{1}{2}m_a V_a^2 + \frac{1}{2}m_b V_b^2 = \frac{1}{2}m_a V_f^2 + \frac{1}{2}m_b V_f^2 + \Delta E \tag{3}$$

where ΔE is the collision energy dissipated within the vehicle structures. It can be deduced that ΔE for this particular collision is 24kJ. It is known (Schmidt et al. 1998) that the ratio of absorption of energy from a collision is proportional to the change in the vehicle velocities where

$$\Delta V_a = \left|V_f - V_a\right| \; and \; \Delta V_b = \left|V_f - V_b\right| \tag{4}$$

It can be deduced that the ratio of $V_a : V_b$ is the same as $m_b : m_a$, so that the smaller vehicle becomes the more vulnerable of the two, and will absorb the larger proportion of the collision energy (Elmarakbi and Jean 2004).

The area under the curve of the force verses deformation directly relates to the work done or the so-called delta energy ΔE, so that

$$force \; x \; deformation = work \; done = delta \; energy$$

and an accurate representation of the nonlinear function is required. The trapezoidal rule is used to approximately calculate the region under the curve of the function f as a sum of trapezoids to obtain the area. It follows that for a typical force versus deformation curve the following is utilised, see Fig. 1

$$\int_a^b f(x)dx \approx (b-a)\frac{f(a)+f(b)}{2} \tag{5}$$

To best determine the integral, the intervals of integration $[a, b]$ are split into smaller subintervals relating to the captured data and the trapezoidal rule is applied to each of them. Unfortunately, there is a lack of data points available, hence prompting the need for an approximate bilinear function which may be readily integrated.

3 Estimation of a spatial bilinear model

One of the problems facing the above task is the lack of available data from crash tests conducted under controlled scenarios. Consequently the paper considers a controlled test performed under simulated conditions. The detail of the test is omitted here for succinctness of the paper, as the main emphasis is placed on the methodology of the spatial modelling approach rather than the experimental detail. The approach developed here is considered to be applicable to other bilinear system models and is not limited to spatially dependent systems. The approach is focused on establishing the coefficients of the model of the form (1), but can be extended to consider the case of higher spatial derivatives (Meyer et al. 2002).

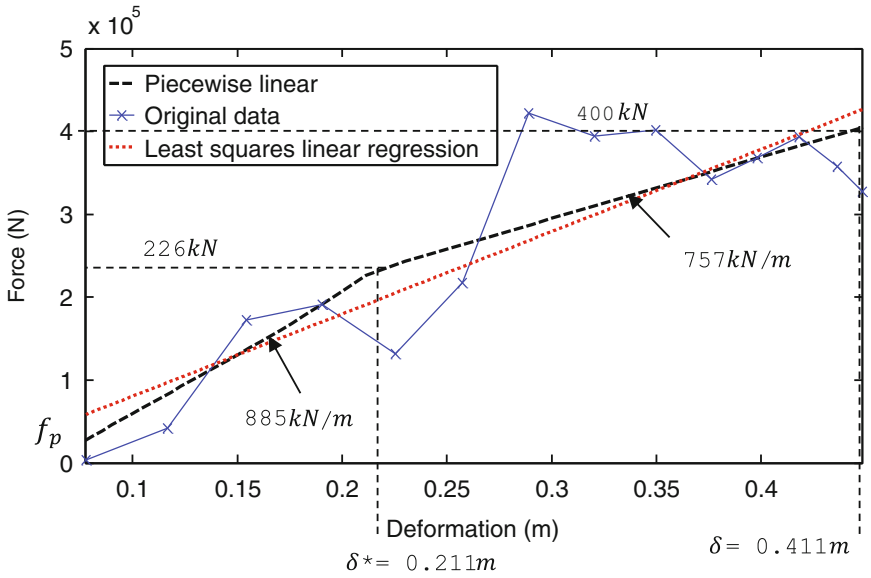


Fig. 2 Data and initial linear/piecewise linear models for simulated experiment

3.1 Piecewise linear benchmark model approach

An initial linear model is obtained from the data by applying linear regression. Subsequently a piecewise linear model (two segments) is obtained with the 'knee' corresponding to δ^*, see Fig. 2, relating to 0.211m of deformation. Of particular note is the value of f_p and the maximum force at the extent of deformation, i.e. 400kN at a deformation of 0.441m.

The following piecewise linear model is thus obtained using a least squares regression procedure

$$f_a = f_p + 885x\delta \qquad \delta < \delta^* \tag{6}$$

$$f_a = f_p + 885x\delta^* + 757x(\delta - \delta^*) \qquad \delta \geq \delta^* \tag{7}$$

where $0 < \delta \leq 0.441$ and $\delta^* = 0.211m$.

It is considered that each crumple zone of a given longitudinal member of the vehicles frontal crash structure will behave in a similar manner, in the case of a full frontal collision, with an initial force f_p to be overcome before buckling compression commences. In practice, due to the multiple repetitive nature of the buckling action of a frontal crash structure, attention is restricted here for a single crumple zone only.

3.2 Conceptual bilinear spatial approach

Based on the form of the model in (1) it is observed that the nonlinearity (bilinearity) is continuous with a decreasing gain as the deformation increases, being representative of a negative bilinearity, as will be introduced, see (1).

Prompted by the need for a convenient and accurate representation of the collision energy a nonlinear (bilinear) spatial model is derived in an iterative manner, with the piecewise linear model providing a useful starting point. With reference to Fig. 2, it is clear that the first piecewise linear segment has a gradient that is greater than the second segment. Recall again equation (1) with a negative bilinear term

$$f = f_p + \Upsilon\delta - \eta\delta f \tag{8}$$

Exploiting the spatial nature of the modelling task, (8) is rearranged to give,

$$f + \eta\delta f = f_p + \Upsilon\delta \tag{9}$$

so that,

$$(1 + \eta\delta)f = f_p + \Upsilon\delta \tag{10}$$

hence,

$$f = \left(f_p + \Upsilon\delta\right)/(1 + \eta\delta) \tag{11}$$

Since a negative bilinearity is observed, it is clear that an initial choice of η must be such that $\eta > s_1$, where $s_1 = 885kN/m$, so that division by the linear factor $(1 + \eta\delta)$ leads to a matching of the maximum force at the full extent of deformation. This initial choice of function is evaluated, so that $\theta = \left[f_p \Upsilon \eta\right]^T$ becomes an initial vector within a least squares estimation algorithm to provide an update to be fed back into the above procedure, defined by equations (9) to (11).

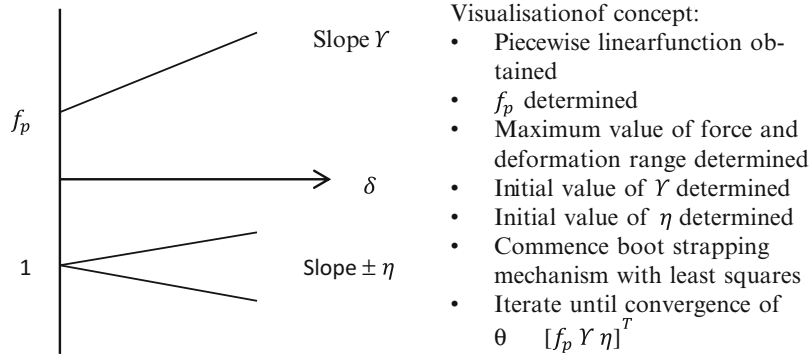


Fig. 3 Pictorial representation of the construction of a spatial bilinear function

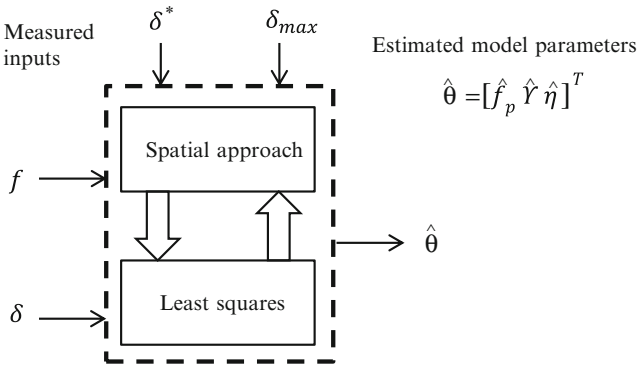


Fig. 4 Bootstrapping configuration of hybrid identification procedure

From equation (11) it follows that the force f may be represented as the ratio of two linear functions of deformation, yielding an overall bilinear function, which needs to be tuned to match a given set of force versus deformation data.

A conceptual visualisation of the approach is illustrated and described in Fig. 3.

The iterative spatial bilinear modelling approach is essentially a hybrid identification procedure whereby the novel approach described above feeds into a least squares algorithm, which recursively/iteratively generates the parameter vector, see Fig. 4.

4 Illustration of approach yielding an initial bilinear model

The novel hybrid approach is demonstrated and compared to the linear and piecewise linear results of Fig. 2. Following the steps outlined above with the piecewise linear function being the starting point, an initial bilinear model is selected as follows, e.g. $\gamma = 1200$ and η evaluated accordingly to match the maximum force at the full extent of deformation. This generates a bilinear model which is subsequently refined using the first step of a least squares procedure and the result is given in Fig. 5.

5 Overall Spatial Bilinear Modelling Procedure

The development of the overall algorithm, prompted by the presence of uncertainty and the absence of a sufficient quantity of data, leads naturally to a combined approach which exploits a least squares procedure in the form of an extended Kalman filter configured for spatial parameter model estimation. The result obtained following the initialisation is shown in Fig. 5. Further iterations of the combined approach yields a convenient and accurate model representing the energy under the force versus deformation curve for use within a control scheme.

The extended Kalman filter algorithm adopted here, see (Young 1974) takes the following prediction/correction form:

Prediction:

$$\widehat{\theta}_{\delta|\delta-\Delta\delta} = \widehat{\theta}_{\delta-\Delta\delta|\delta-\Delta\delta} + \Delta\widehat{\theta}(f + \Delta f) \tag{12}$$

$$\Phi_{\delta|\delta-\Delta\delta} = \Phi_{\delta-\Delta\delta|\delta-\Delta\delta} + W_\delta \tag{13}$$

Correction:

$$f\widehat{\theta}_{\delta|\delta} = \widehat{\theta}_{\delta|\delta-\Delta\delta} + \theta_\delta\left[\delta_\delta - x_\delta^T\widehat{\theta}_{\delta|\delta-\Delta s}\right] \tag{14}$$

$$_\delta = \Phi_{\delta|\delta-\Delta\delta}x_\delta\left[r_\nu + x_\delta^T\Phi_{\delta|\delta-\Delta s}x_\delta\right]^{-1} \tag{15}$$

$$\Phi_{\delta|\delta} = \left[I - _\delta x_\delta^T\right]\Phi_{\delta|\delta-\Delta\delta} \tag{16}$$

where $\widehat{\theta} = \left[\widehat{f}_p\widehat{\Upsilon}\widehat{\eta}\right]$ and $x^T = [1\delta\delta f]$, with subscripts δ *and* $\delta - \Delta\delta$ denoting the spatial nature of the evolving estimation and Φ, W, θ *and* I denoting the covariance matrix, process noise covariance matrix, gain vector and identity matrix, respectively.

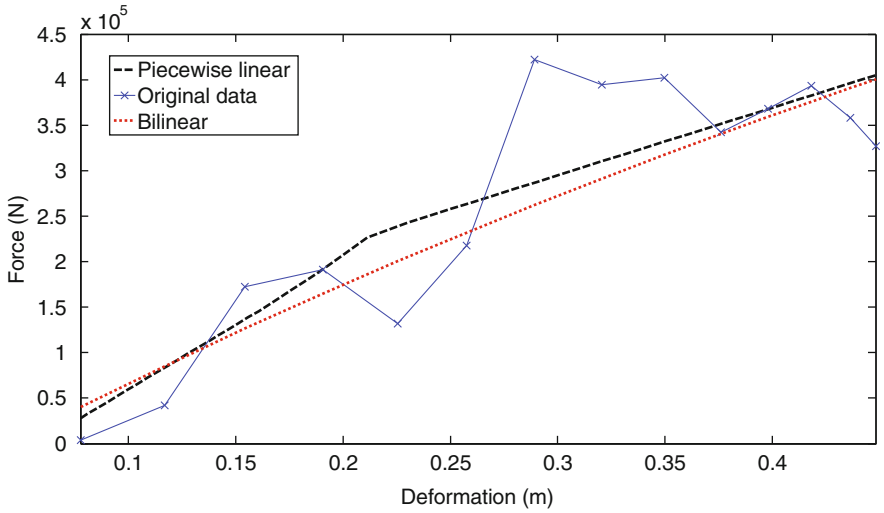


Fig. 5 Data and initial piecewise linear/bilinear models for simulated experiment

6 Concluding remarks

Prompted by the need for modelling the nonlinear energy absorption phenomenon when a vehicle experiences a frontal impact collision, a novel bilinear spatial modelling procedure involving the ratio of two linear functions of deformation is proposed. When implemented iteratively in conjunction with an extended Kalman filter approach the resulting hybrid procedure provides an effective representation of the collision energy in the presence of uncertainty and with a paucity of measured data.

References

1. Burnham, K. J. (1991), Self-tuning Control for Bilinear Systems, PhD thesis, Coventry Polytechnic, UK.
2. Elmarakbi, A.M. and Jean, W. Z. (2004), Numerical Analysis and Optimisation of Vehicle Compatibility Using a Smart Vehicle Structure, SAE Technical Paper 2003-01-2802.
3. Meyer, J., Burnham, K. J., Haas, O. C. L., Mills, J. A. and Parvin, E. M. (2002), Application of a least-squares parameter estimation approach for 2-D spatial modelling of compensators for intensity-modulated radiotherapy, Transactions of the Institute of Measurement and Control, Vol. 24 (5), 369-386.
4. Mohler, R. R. (1973), Bilinear Control Processes. New York: Academic Press.
5. Schmidt, B., Haight, W. R., Szabo, T. and Welcher, J. (1998), System-based Energy and Momentum Analysis of Collisions, SAE Technical Paper 980026.
6. Trollope, J. E. and Burnham, K. J. (2013a), Collision Energy Mitigation through Active Control of Future Lightweight Vehicle Architectures. Proc. CD-ROM, Vol. 2, 10th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2013), pp 477-484, Reykjavik, Iceland, July 29-31.
7. Trollope, J. E. and Burnham, K. J. (2013b), Active Buckling Control, (patent pending), GB1320489.6.
8. Trollope, J. E. and Burnham, K. J. (2013c), Active Buckling Control for Future Lightweight Vehicle Body Structures. Measurement and Control. 46 (10), 315-320.
9. Young, P. C. (1974), Recursive approaches to time-series analysis. Bull. of Inst. Mathematics and its Applications, 10, 209-224.

Special Session: Intelligent Video Surveillance Systems

A Fast Non-searching Algorithm for the High-Speed Target Detection

Jibin Zheng, Tao Su, Wentao Zhu, and Qing Huo Liu

1 Introduction

The high-speed target detection plays an important role in the modern radar system [1, 2, 3, 4, 5, 6]. The conventional detection method is the moving target detection (MTD) algorithm [7], which requires the range profile alignment and can be efficiently implemented via the fast Fourier transform (FFT). However, for the high-speed target, the linear range migration may easily happen during the coherent processing interval. Thus, due to the effect of the linear range migration, the range profile alignment is disturbed and the MTD algorithm cannot work any more [7, 8].

For the high-speed target detection, several detection algorithms have been developed. The Hough transform [4, 5, 6], the Radon transform [7], the quantifying based method [8, 9] and the Radon-Fourier transform (RFT) [10, 11] are based on motion parameters searching method to compensate the effect of the linear range migration. The large computational cost and the complicated radar system are always necessary for the searching method, which limit its application in the high-speed target detection. A fast implementation method of the RFT has been proposed based on the chirp-z transform [12], which is searching free for the slow moving target. However, due to the velocity ambiguity, the searching procedure is still necessary for the chirp-z transform based RFT when the RFT is utilized to deal with the high-speed target. The keystone transform can blindly compensate for the linear range migration without *a priori* knowledge of accurate motion parameters. It has been widely applied in the detection and the motion parameter estimation of the air moving target [13] and the ground moving target [14, 15]. A fast implementation of the keystone transform is proposed based on the chirp-z transform [15], while the velocity ambiguity is also a limitation.

In this paper, aiming at problems of the MTD algorithm and searching algorithms above, a fast non-searching algorithm is proposed for the high-speed target detection. This algorithm employs a novel symmetric autocorrelation function and the inverse fast Fourier transform (IFFT). Compared to the MTD algorithm and searching algorithms above, the proposed fast algorithm can complete the detection of the high-speed target with the lower computational cost and the less complicated radar system. It is worthwhile noting that, in this paper, we propose to utilize the radial velocity to determine the high-speed target and this may provide a novel idea for the echo processing of the high-speed target. The analysis of the computational cost and several simulation examples on the synthetic model are shown to validate the effectiveness of the fast non-searching algorithm.

2 Fast Non-searching Algorithm for the High-Speed target Detection

Suppose the radar transmits the LFM signal, which takes the form

$$x(\hat{t}) = \mathrm{rect}\left(\frac{\hat{t}}{T_p}\right)\exp\left[j2\pi\left(f_c\hat{t} + \frac{1}{2}\gamma\hat{t}^2\right)\right] \quad (1)$$

where $\mathrm{rect}[x] = \begin{cases} 1, |x| \le 1/2 \\ 0, |x| > 1/2 \end{cases}$. $\hat{t}$, T_p, f_c and γ denote the fast time, the pulse width, the carrier frequency, and the frequency modulation rate, respectively.

For N targets, the received baseband signal can be represented as

J. Zheng (✉) • T. Su • W. Zhu
National Lab of Radar Signal Processing, Xidian University, Xi'an, China
e-mail: jz119@duke.edu; sutao@xidian.edu.cn; wtzhufr@163.com

Q.H. Liu
Department of Electrical and Computer Engineering, Duke University, Durham, USA
e-mail: qhliu@ee.duke.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_112,

$$s(\hat{t},t_m)=\sum_{i=1}^{N}A_i\mathrm{rect}\left[\frac{\hat{t}-2R_i(t_m)/c}{T_p}\right]\exp\left[-j2\pi f_c\frac{2R_i(t_m)}{c}\right]\cdot\exp\left[j\pi\gamma\left(\hat{t}-\frac{2R_i(t_m)}{c}\right)^2\right]+n(,\hat{t},t_m) \tag{2}$$

where t_m and A_i are the slow time and the amplitude of the echo, respectively. $R_i(t_m)$ denotes the time-varying range of the ith target during the coherent processing interval. $n(\hat{t},t_m)$ is the additive complex white Gaussian noise with a variance of δ^2.

Utilizing the matched filter $H(\hat{t})=\mathrm{rect}(\hat{t}/T_p)\exp(j\pi\gamma\hat{t}^2)$ to complete the pulse compression, we can obtain the baseband echo in the spatial frequency domain.

$$C_s(f,t_m)=\sum_{i=1}^{N}B_i\mathrm{rect}\left[\frac{f}{\gamma T_p}\right]\exp\left[-j2\pi(f+f_c)\frac{2R_i(t_m)}{c}\right]+n(f,t_m) \tag{3}$$

where f is the frequency domain with respect to $\hat{t}$. B_i denotes the amplitude after the pulse compression.

The target model is illustrated in Fig. 1. The Y axis is the radar line-of-sight (RLOS). Suppose the velocity and the radial initial range of the *ith* target are v_i and R_{0i}, respectively. The velocity v_i can be decomposed into the radial velocity v_{0i} and the perpendicular velocity v_{1i}, while v_{1i} does not cause the radial motion. Therefore, the radial distance $R_i(t_m)$ between the *ith* target and the radar at the instant time t_m can be expressed as

$$R_i(t_m)=R_{0i}+v_{0i}t_m \tag{4}$$

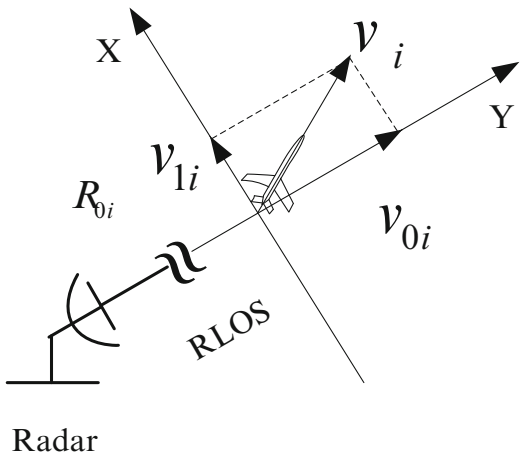


Fig.1 Target model

Substituting (4) into (3), we rewrite the baseband echo in the spatial frequency domain.

$$C_s(f,t_m)=\sum_{i=1}^{N}B_i\mathrm{rect}\left[\frac{f}{\gamma T_p}\right]\exp\left[-j2\pi(f+f_c)\frac{2R_{0i}}{c}\right]\cdot\exp\left[-j2\pi(f+f_c)\frac{2v_{0i}t_m}{c}\right]+n(f,t_m) \tag{5}$$

Due to the high velocity and the low radar pulse repetition frequency, the undersampling will occur and the Doppler frequency can be expressed as

$$f_{di}=\frac{2v_{0i}}{\lambda}=f_{d0i}+N_{ki}\mathrm{PRF} \tag{6}$$

where λ is the wave length. f_{d0i} and N_{ki} denotes the ambiguous Doppler frequency and the velocity ambiguity, respectively. PRF is the pulse repetition frequency.

Based on (6), (5) can be rewritten as

$$C_s(f,t_m)=\sum_{i=1}^{N}B_i\mathrm{rect}\left[\frac{f}{\gamma T_p}\right]\exp\left[-j2\pi f_c\frac{2R_{0i}}{c}\right]\exp\left[-j2\pi f\frac{2R_{0i}}{c}\right]\cdot\exp(-j2\pi f_{d0i}t_m)\exp\left(-j2\pi f\frac{2v_{0i}t_m}{c}\right)+n(f,t_m) \tag{7}$$

In (7), the second exponential term indicates the initial range, the third exponential term indicates the ambiguous Doppler frequency, and the fourth exponential term indicates the range migration induced by the high velocity. It is easily seen from (7) that, if v_{0i} is obtained first, the target detection and the estimation of R_{0i} can be easily completed by utilizing the IFFT and the FFT. That is, the target can be also determined by v_{0i}. Therefore, we define a novel symmetric autocorrelation function

$$R(f,t_m)=C_s(f,t_m+\tau)C_s^*(f,t_m-\tau)=R_{self}(f,t_m)+R_{cn}(f,t_m) \tag{8}$$

where τ denotes a constant lag-time, which is chosen according to [16] in this paper. $R_{self}(f, t_m, \tau)$ and $R_{cn}(f, t_m, \tau)$ represent the self-term and the noisy cross-term, respectively.

Substituting (7) into (8), we obtain

$$R(f,t_m)=\sum_{i=1}^{N}B_i^2\mathrm{rect}\left[\frac{f}{\gamma T_p}\right]\exp(-j4\pi f_{d0i}\tau)\exp\left(-j4\pi f\tau\frac{2v_{0i}}{c}\right)+R_{cn}(f,t_m) \tag{9}$$

Two variables, f and t_m, appear in (9), while t_m appears only in the cross-term $R_{cn}(f, t_m)$. That is, the energy of the

self-term can be accumulated along t_m axis with the discrete Fourier transform (DFT) at the frequency zero, which can be realized by the add operation.

$$R_D(f) = \sum_{i=1}^{N} B_i^2 \mathrm{rect}\left[\frac{f}{\gamma T_p}\right] \exp(-j4\pi f_{d0i}\tau)\exp\left(-j4\pi f \tau \frac{2v_{0i}}{c}\right)\delta_{t_m} + R_{D,cn}(f) \tag{10}$$

where $R_{D,cn}(f)$ and δ_{t_m} denote the noisy cross-term and the accumulated energy after the add operation, respectively.

Performing IFFT to f axis, we can obtain

$$R_k(\hat{t}) = \sum_{i=1}^{N} B_i^2 G_i \mathrm{sinc}\left[B\left(\hat{t} - \frac{4v_{0i}}{c}\tau\right)\right]\delta_{t_m}\exp(-j4\pi f_{d0i}\tau) + R_{k,cn}(\hat{t}) \tag{11}$$

where G_i is the gain after the IFFT operation. $R_{k,cn}(\hat{t})$ denotes the noisy cross-term.

In (11), we estimate the target velocity and detect the target [9, 12]. Thereafter, dechirping $C_s(f, t_m)$ with exp $[j2\pi f(2v_{0i}t_m)/c]$, we obtain the initial range with the FFT and the IFFT.

$$\left(R'_{0i} = \frac{\hat{t}\,c}{2}\right) = \underset{(,\hat{t},f_{t_m})}{\mathrm{argmax}} \left|\mathrm{IFFT}_f\left\{\mathrm{FFT}_{t_m}\left[C_s(f,t_m)\exp\left(j2\pi f\frac{2v_{0i}t_m}{c}\right)\right]\right\}\right| \tag{12}$$

Above is the proposed algorithm for the detection of high-speed target. It is obvious the proposed algorithm, which can be easily implemented by using the complex multiplications, the add operations and the IFFT, is searching free. In Section 3, through several numerical examples and the analysis of the computational cost, the effectiveness of the proposed fast non-searching algorithm will be demonstrated.

3 Numerical Examples

In Section 2, a fast non-searching algorithm is proposed for the detection of the high-speed target. In this section, two numerical examples will be shown under the signal to noise ratio (SNR) equals to -5 dB. **Example 1** is utilized to verify the effectiveness of the proposed algorithm with the high-speed mono-target. **Example 2** is utilized to verify the effectiveness of the proposed algorithm with high-speed multi-targets. All simulations in this paper are completed on a personal computer with an Intel Core 2 Duo (2.83 GHz) processor and 2 GB memory. The radar parameters and the motion parameters are listed in Table I.

TABLE I. RADAR PARAMETERS AND MOTION PARAMETERS

Carrier frequency	10 GHz	Pulse width		2.56 us
Bandwidth	100 MHz	Sample frequency		100 MHz
Pulse repetition frequency	256 Hz	Effective echo pulses		256
Targets	**A1**	**A2**	**A3**	**A4**
Initial distance (km)	62	61.97	62	62.03
Initial velocity (m/s)	90	96	102	108

In these two numerical examples, we focus on (1) the proposed fast algorithm can complete the high-speed target detection under the low SNR; (2) the computational cost of the proposed fast algorithm to complete the target detection is lower than that of the MTD method.

Example 1: Fast non-searching algorithm with the high-speed mono-target (A1)

Fig. 2 (a) shows the signal after the pulse compression. It is obvious the range profile alignment is disturbed due to the linear range migration. Thus, the MTD method cannot accumulate the target energy in Fig. 2 (b). Through the proposed fast algorithm, the simulation result is given in Fig. 2 (c). In Fig. 2 (c), the target energy is accumulated into a sole peak. The target detection and the estimation of the radial velocity can be completed with the result of Fig. 1 (c). According to (12), the range migration can be eliminated with the dechirp method and the result after the compensation is shown in Fig. 2 (d), where the initial range estimation can be completed. For the proposed fast algorithm, implementation procedures include the symmetric autocorrelation function (the number of complex multiplication is $MN/2$, where M is the number of the fast time and N is the number of the slow time), the add operation and the IFFT (the number of complex multiplication is $(M/2)\log_2 M$). Therefore, the number of complex multiplications shown in TABLE II is $MN/2 + (M/2)\log_2 M$ for the proposed algorithm. For the conventional MTD method, implementation procedures include the FFT along the slow time and the IFFT along the spatial frequency. Therefore, the number of complex multiplications is $(MN/2)(\log_2 M + \log_2 N)$. It is obvious that the proposed fast algorithm is more efficient than the conventional MTD method.

Example 2: Fast non-searching algorithm with high-speed multi-targets (A1- A4)

Fig. 3 shows the simulation results of the proposed algorithm and the MTD method. It is obvious that the MTD method cannot work due to the linear range migration. Fig. 3 (c) gives the result of the proposed algorithm. The self-term is accumulated, while the cross-term cannot be

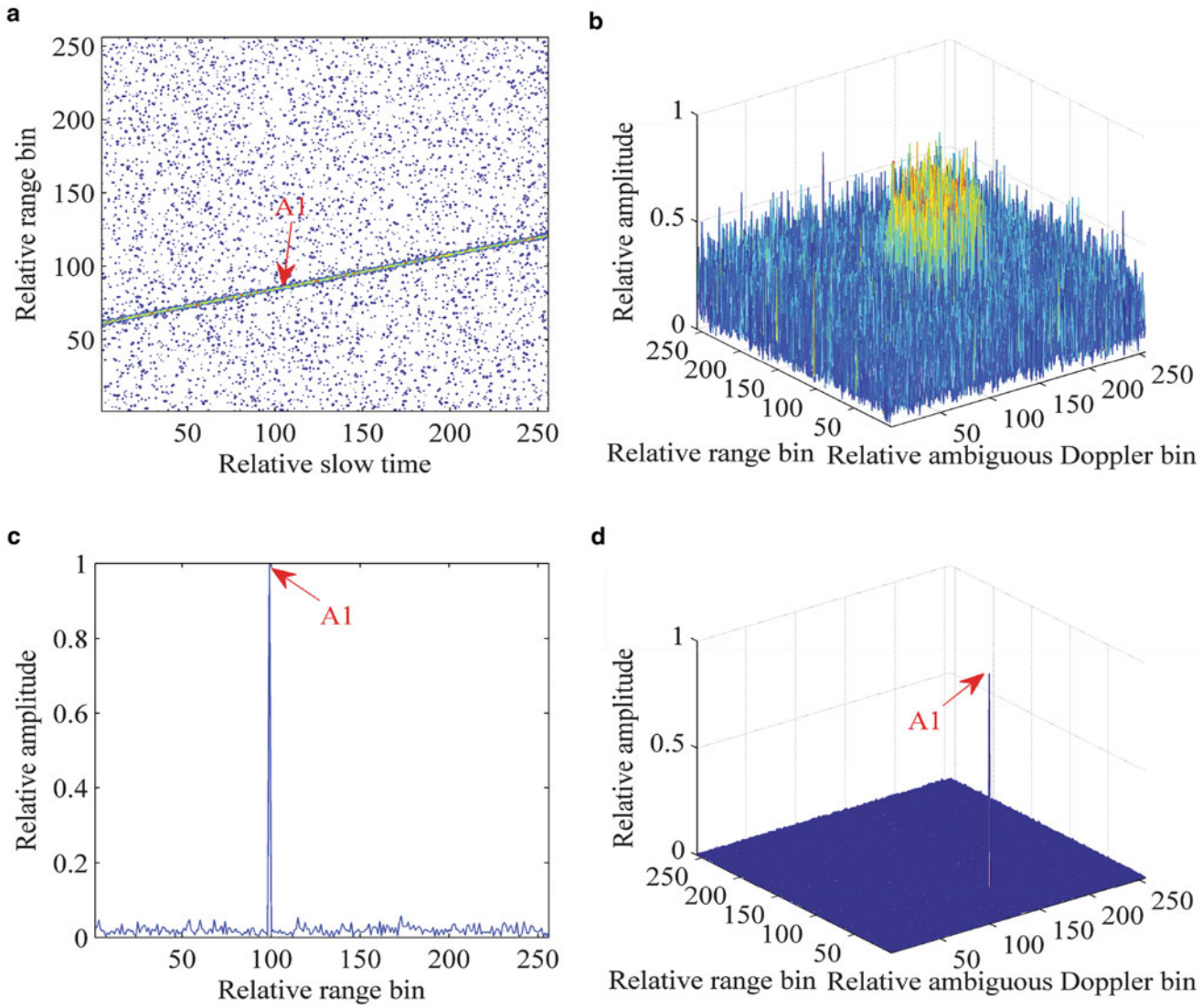


Fig. 2 Simulation results of Example 1. (a) Result after the pulse compression. (b) Result of the MTD method. (c) Result of the proposed algorithm. (d) Result after the compensation of the range migration.

TABLE II. COMPUTATIONAL COST

Computation cost Methods	Number of complex multiplications	Simulation time (s)
MTD method	$(MN/2)$ $(\log_2 M + \log_2 N)$	0.0089
Fast non-searching algorithm	$MN/2 + (M/2)\log_2 M$	0.0049

accumulated. The radial velocity estimation and the target detection can be completed in Fig. 3 (c). With the estimated radial velocity, we compensate the range migration in Fig. 3 (d) and the range can also be estimated.

4 Conclusion

In this paper, a fast non-searching algorithm is proposed for the high-speed target detection. In this algorithm, we employ the novel symmetric autocorrelation function and the inverse fast Fourier transform. Thus, the searching procedure is eliminated in the proposed fast algorithm. Compared to the MTD method and searching algorithms, the proposed algorithm can complete the target detection with the lower computational cost and the less complicated radar system. Furthermore, we utilize the radial velocity to determine the

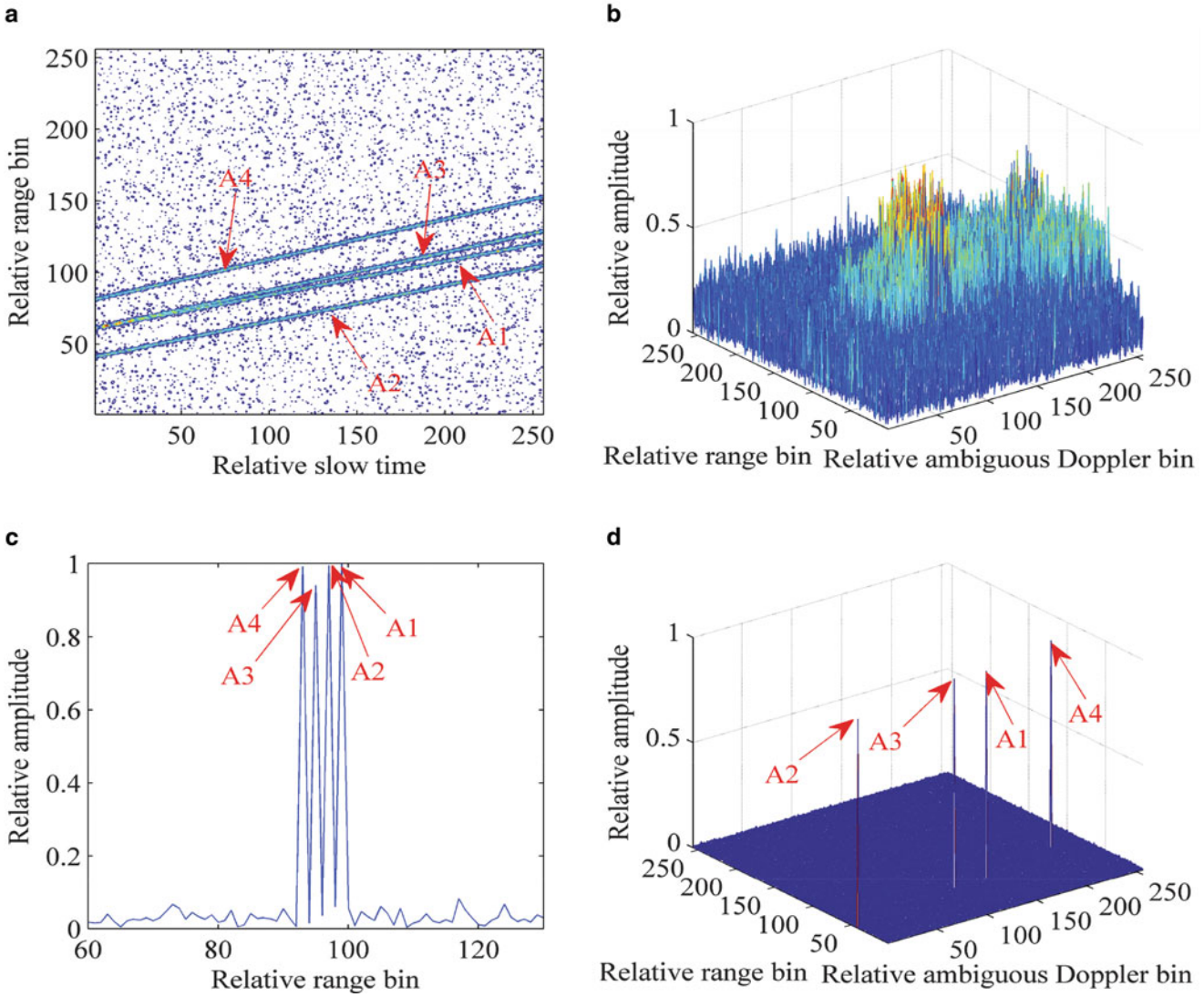


Fig. 3 Simulation results of Example 2. Result after the pulse compression. (b) Result of the MTD method. (c) Result of the proposed algorithm. (d) Result after the compensation of the range migration.

target, which may provide a novel ideal for the signal processing of the high-speed target. In this paper, we utilize several numerical examples and the analysis of the computational cost to demonstrate the effectiveness of the proposed algorithm.

Acknowledgement This work was supported in part by the National Natural Science Foundation of China under Grants 61001204, the Science and technology Foundation of Shaanxi Province (2012JM8015), and the Xi'an Polytechnic University Dr Support Foundation (BS1119).

References

1. Xu, J., Xia, X., Peng, S., Yu, J., Peng, Y., Qian, L.: Radar maneuvering target motion estimation based on generalized Radon-Fourier transform. IEEE Trans. Signal Process., vol. 60, no. 12, pp. 6190-6201 (2012)
2. Sun G., Xing M., Xia X., Wu Y., Bao Z.: Robust ground moving-target imaging using deramp-keystone processing. IEEE Trans. Geosci. Remote Sens., vol.51, no. 2, pp. 966-982 (2013)
3. Zheng, J., Su T., Zhu W., Liu, Q. H.: ISAR Imaging of Targets With Complex Motions Based on the Keystone Time-Chirp Rate Distribution. IEEE Geosci. Remote Sens. Lett., vol.11, no. 7, pp. 1275-1279 (2014)
4. Carlson B. D., Evance E. D., Wilson S. L.: Search radar detection and track with the hough transform part I: system concept. IEEE Trans. Aerosp. Electron. Syst., vol. 30, no. 1, pp. 102–108 (1994)
5. Carlson B. D., Evance E. D., Wilson S. L.: Search radar detection and track with the hough transform part I: detection statistic. IEEE Trans. Aerosp. Electron. Syst., vol. 30, no. 1, pp. 109–115 (1994)
6. Carlson B. D., Evance E. D., Wilson S. L.: Search radar detection and track with the hough transform part I: detection performance with binary integration. IEEE Trans. Aerosp. Electron. Syst., vol. 30, no. 1, pp. 116–125(1994)

7. M. Xing, J. Su, G. Wang, and Z. Bao." New parameter estimation and detection algorithm for high speed small target," *IEEE Trans. Aerosp. Electron. Syst.,* vol. 47, no.1, pp. 214–224, 2011.
8. R. Tao, N. Zhang, and Y. Wang, " Analysing and compensating the effects of range and Doppler frequency migrations in linear frequency modulation pulse compression radar," *IET Radar, Sonar and Navigation,* vol.5, iss. 1, pp. 12-22, 2011.
9. Zheng J., Su T., Liu Q. H., Zhang L., Zhu W.: Fast Parameter Estimation Algorithm for Cubic Phase Signal Based on Quantifying Effects of Doppler Frequency Shift. Progress In Electromagnetics Research, vol.142, pp. 57-74 (2013)
10. Xu J., Yu J., Peng Y., Xia X.: Radon-Fourier transform for radar detection, I: generalized Doppler filter bank. IEEE Trans. Aerosp. Electron. Syst., vol. 47, no. 2, pp. 1186-1200 (2011)
11. Xu J., Yu J., Peng Y., Xia X.: Radon-Fourier transform for radar detection, II: blind speed sidelobe suppression. IEEE Trans. Aerosp. Electron. Syst., vol. 47, no. 4, pp. 2473–2489 (2011)
12. Yu J., Xu J., Peng Y., Xia X.: Radon-Fourier transform for radar detection, III: optimality and fast implementations. IEEE Trans. Aerosp. Electron. Syst., vol. 48, no. 2, pp. 991–1004 (2012)
13. Zhang S., Zeng T., Long T., Yuan H.: Dim target detection based on keystone transform. In: Proc. IEEE Int. Radar Conf., pp. 889-894 (2005)
14. Perry R. P., Dipietro R. C., Fante R. L.: SAR imaging of moving targets," IEEE Trans. Aerosp. Electron. Syst., vol. 35, no. 1, pp. 188-200 (1999)
15. Zhu D., Li Y., Zhu Z.: A keystone transform without interpolation for SAR ground moving-target imaging. IEEE Geosci. Remote Sens. Lett., vol.4, no. 1, pp. 18-22 (2007)
16. Djurovic I., Simeunovic M., Djukanovic S., Wang P.: A hybrid CPF-HAF estimation of polynomial-phase signals: detailed statistical analysis. IEEE Trans. Signal Process., vol.60, no. 10, pp. 5010–5023 (2012)
17. Lv X., Bi G., Wang C., Xing M.: Lv's distribution: principle, implementation, properties, and performance. IEEE Trans. Signal Process., vol.59, no. 8, pp. 3576-3591 (2011)

A Comparative Study of Video Splitting Techniques

Abdul Khader Jilani Saudagar and Habeeb Vulla Mohammed

1 Introduction

With the swift escalation of the number of television channels, media, internet and online information services, information becomes progressively obtainable and accessible. The computerization enhances preservation of records and makes the access to documents easier. On the other hand, when the quantity of documents become important the digitalization is not enough to ensure an efficient access. Indeed, we need to have system to extract text appearing in video, which often reflects a prospects semantic content and properly index for efficient and fast access. Many research works are proposed for text extraction from motion video for effective indexing and searching. But very few researchers has made on Arabic videotext extraction. So in this process the first step involves the splitting of the video into frames and if this is done in apt manner then the percentage of accuracy in text extraction will be more.

2 Methods

Numerous works [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16] have been proposed in the past for the text extraction process from a video but little attention was given to the technique used in splitting the video into frames which is the first step in text extraction. A new video feature extraction technique based on the Generalized Hough Transform (GHT) [17] is proposed, which calculate the differences between two frames in terms of rotation, scale and displacement. An overview of existing techniques for video segmentation that operate on both uncompressed and compressed video stream are discussed by I. Koprinska and S. Carrato [18]. A hierarchical approach for segmentation of videos into shots and scenes using visual content is proposed by Andrew Thompson [19].

The well-known methods for video splitting are FFmpeg [20] and VLC [21]. FFmpeg is a complete, cross-platform solution to record, convert and stream audio and video files. It contains libraries for encoding or decoding media. FFmpeg is a command line program for transcoding multimedia files where there is no need to play the file in-order to get the frames and the user can set the time-stamp according to the requirement. VLC media player is a portable, free and open-source, cross-platform media player. The default distribution of VLC includes a large number of free decoding and encoding libraries, avoiding the need for finding/calibrating proprietary plugins. When compared to FFmpeg, VLC is also a command line program, user has the facility to fix the time interval but VLC takes more time, as the generation of frames will be completed only after playing of the video reaching to an end.

The proposed approach is simple for the splitting the video into frames and has the following advantages: it can be run by the command line as well as through an interface, no need to play the video file and user can decide the time-stamp. The output frames from the proposed approach, FFmpeg and VLC are compared in-terms of image difference and PSNR[22, 23, 24].

In this work we selected the same frame number from the output frames generated by the three approaches with a time-stamp of 5 seconds and calculated the values of frame differences and PSNR between proposed approach vs FFmpeg, proposed approach vs VLC and FFmpeg and VLC and plotted the values as shown in Fig. 5 and Fig. 6. The following equations are used in-order to calculate the frame difference.

$$rgb1 = image1.gerRGB(width, height)$$

$$rgb1 = image2.gerRGB(width, height)$$

A.K.J. Saudagar (✉) • H.V. Mohammed
College of Computer & Information Sciences, Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia
e-mail: saudagar_jilani@ccis.imamu.edu.sa; habeebvulla@ccis.imamu.edu.sa

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_113,

$$r1 = (rgb1) \gg 16)\&0 \times ff$$

$$g1 = (rgb1) \gg 8)\&0 \times ff$$

$$b1 = (rgb1)\&0 \times ff$$

$$r2 = (rgb2) \gg 16)\&0 \times ff$$

$$g2 = (rgb2) \gg 8)\&0 \times ff$$

$$b2 = (rgb2)\&0 \times ff$$

$$Difference = \sum_{i=0,j=0}^{n=width,m=height} (Math.abs(r1 - r2) + Math.ads(g1 - g2) + Math.ads (b1 - b2)) \quad (1)$$

$$\begin{aligned} &where\ r1, r2\ red\ values\ of\ image\ 1\ and\ image\ \ 2; \\ &where\ g1, g2\ red\ values\ of\ image\ 1\ and\ image\ \ 2; \\ &where\ b1, b2\ red\ values\ of\ image\ 1\ and\ image\ \ 2; \\ &Percentage\ Difference\ = Difference/n/n255.0 \\ &where\ n\ = width * height*3 \end{aligned} \quad (2)$$

The following equations are used in-order to calculate the PSNR.

$$PSNR = 20 \quad \log_{10}\left(\frac{MAX_f}{\sqrt{MSE}}\right) \quad (3)$$

Where MSE (Mean Square Error) is

$$MSE = \frac{1}{mn}\sum_{0}^{(m-1)}\sum_{0}^{n-1} f(i,j) - g(i,j)^2 \quad (4)$$

f represents the matrix data of our first image.

g represents the matrix data of our second image.

m represents the numbers of rows of pixels of the images and i represents the index of that row.

n represents the number of columns of pixels of the image and j represents the index of that column.

MAX_f is the maximum signal value.

The output frames for the proposed approach, FFmpeg and VLC are shown in Fig. 2, Fig. 3 and Fig. 4.

3 Implementation of Proposed Approach

Splitting the video into frames is carried out by using Java Programming where the sample video file is split into frames using the java library with milliseconds of time span between them. The time between the frames can be provided by the user to split the video into image frames.

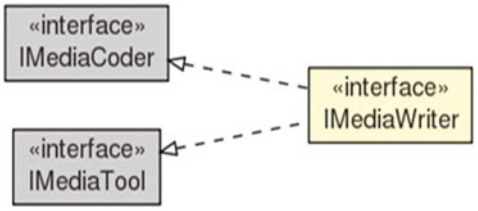


Fig. 1. Class diagram

The implementation process begins with reading the input video file using File Class of java, then read this file in to IMediaReader Class (This class contains IContainer Inner class). This IMediaReader Class is used to read the video and decode the video media. The reader opens up a media container, reads packets from it, decodes the data, and then dispatches information about the data to any registered IMediaListener objects(This is an interface used for media objects).IMdeiaListener extends the MediaListenerAdapter, which is an adapter (provides empty methods) implementing the IMediaListenerinterface. This interface is notified about events generated during the video processing. We just see the video events so we implement IMediaListener.OnVPicture Method. Inside that, we use the provided IVPictureEvent object to find what stream (video only) we are dealing with.

In order to capture frames in specific times, we have to use timestamps. First, we take the very first frame with specific value that means no time stamp is set for a given object. When the minimum elapsed time has passes, we capture the frame by invoking IVPictureEvent .getImage method, which returns the underlying BufferedImage. Here we concentrate on elapsed time into real time. After that, we dump image data to a PNG format using the ImaageIO.write method. Finally, we update the last write time.

In addition, we use java listeners to handle the events.

The programme consists of one class Image Listener this consists of two methods getPicture and putPicture Methods.

The main class imports the java packages used in this programme.

We specify the time span between the frames and a file object to take the video file. Then IMediaReader class will read the video file. Once the video file encoded by the IMediaReader class we use the indexing of the images by giving them sequence of numbers. Image naming can be done by using time stamp.

Using the putImage method, we write the generated image thumbnails to disk in particular location provided with .png extension.

frame # 2 frame # 18 frame # 21

frame # 38 frame # 39

Fig. 2. Output frames by proposed approach

frame # 2 frame # 18 frame # 21

frame # 38 frame # 39

Fig. 3. Output frames by FFmpeg

Fig. 4. Output frames by VLC

Fig. 5. Frame difference in percentage

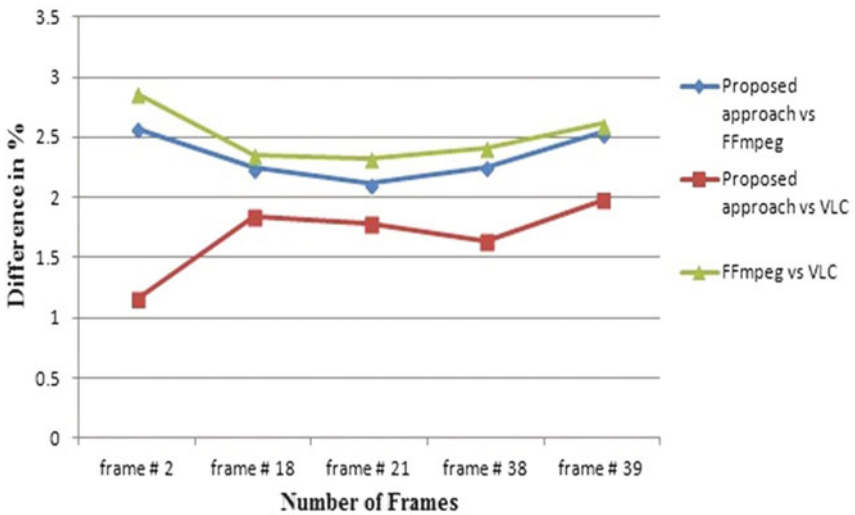
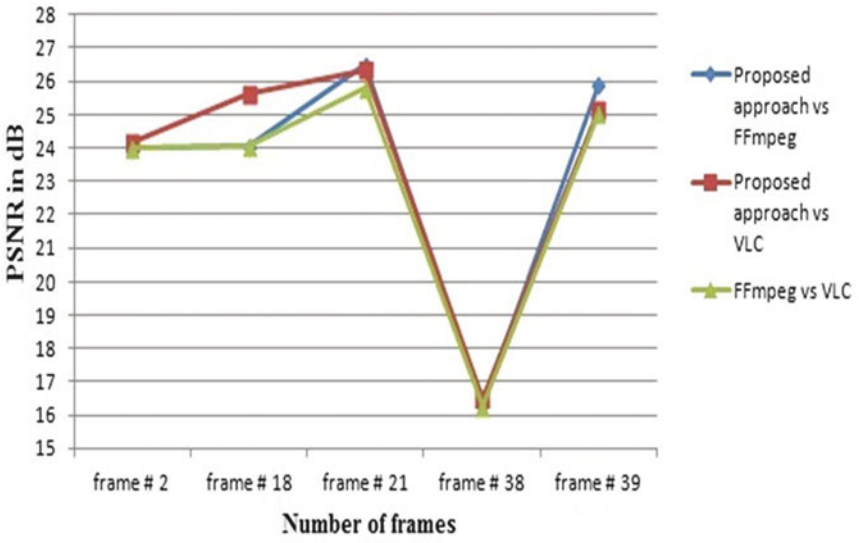


Fig. 6. PSNR in dB




```
Algorithm: videoThumbNails
procedure getVideoPicture(IVideoPicture
Event)
If the videostream id is not yet set go head
   If videoStreamIndex ← negative 1
   videoStreamIndex← videoStreamId
   else
   return;
 if videoStreamId uninitialized
      pictureWrite<-get the timestamp
-micro seconds between the frames;
    if get the timestamp- pictureWri-
te>=micro seconds between the frames
        outFile←Procedure picturetoFile
(Image event)
        sec←(double)get the timestamp;
        pictureWrite← + second between
the frames;
Procedure picturetoFile(BufferedImage)
          Outfile← outfileprefix+Sys-
tem.currentTimeMillis()+".png";
          ImageIO.write(image, "png",
new File(outfile));
```

4 Results

The proposed approach is tested on 16 video samples, for this paper we consider a video sample of length 6 minutes and 22 seconds from Aljazeera television containing scrolling Arabic text from left to right direction and split the video with a time-stamp of 5 seconds using the proposed approach, FFmpeg and VLC. We should take into consideration scrolling rate of the Arabic text in a video and based on that we can decide the time-stamp so that the scrolling text never disregard from the frames. The following are the generated output frames which are of 480×360 pixel dimension and they are in .png format.

5 Conclusion

It is observed from figure 5 that the frame difference values are minimum when the proposed approach is used with FFmpeg and VLC as against to FFmpeg and VLC. The PSNR value is also slightly higher as shown in figure 6 when the proposed approach is used with FFmpeg and VLC as compared to FFmpeg and VLC. This shows that the generated frames by this approach have less error when compared to frames generated by other techniques and can be efficiently used for further video processing such as segmentation, text extraction, localization etc.

6 Future Work

The future work will focus on the Arabic text extraction which includes text detection and text recognition process.

Acknowledgements This research is supported by King Abdulaziz City for Science and Technology, Saudi Arabia, vide grant no. AT-32-87.

References

1. Agnihotri, L., Dimitrova, N.: Text Detection for Video Analysis. In: IEEE Workshop on Content-Based Access of Image and Video Libraries, pp. 109–113 (1999)

2. Gargi, U., Crandall, D., Antani, S., Gandhi, T., Keener, R., Kasturi, R.: A System for Automatic Text Detection in Video. In: 5th International Conference on Document Analysis and Recognition, pp. 29–32 (1999)
3. Li, H., Doermann, D.: Automatic Text Detection and Tracking in Digital Video. IEEE T Image Process. 9(1), 147–156 (2000)
4. Chen, D., Odobez, J.M., Bourlard, H.: Text Detection and Recognition in Images and Video Frames. Pattern Recogn. 37(3), 595–608 (2004)
5. Huang, W., Shivakumara, P., Tan, C.L.: Detecting Moving Text in Video using Temporal Information. In: 19th International Conference on Pattern Recognition, pp. 1–4 (2008)
6. Yen, S. H., Chang, H. W., Wang, C. J., Wang, C.W.: Robust News Video Text Detection based on Edges and Line-Deletion. WSEAS Transactions on Signal Processing. 6(4), 186–195 (2010)
7. Wredenhagen, G.F.: Moving Text Detection in Video. Patent US20100254605 A1. (2010)
8. Moradi, M., Mozaffari, S.: Hybrid Approach for Farsi/Arabic Text Detection and Localisation in Video Frames. IET Image Processing.7(2), (2013)
9. Antani, S., Crandall, D., Kasturi, R.: Robust Extraction of Text in Video. In: 15th International Conference on Pattern Recognition, pp. 831–834 (2000)
10. Palma, D., Ascenso, J., Pereira, F.: Automatic Text Extraction in Digital Video based on Motion Analysis. Image Analysis and Recognition, LNCS, vol. 3211, pp. 588–596. Springer, Heidelberg (2004).
11. Lyu, M.R., Song, J., Cai, M.: A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction. IEEE T Circ Syst Vid. 15(2), 243–255 (2005)
12. Karray, H., Salah, M., Alimi, A. M.: TEVI Text Extraction for Video Indexing. http://pdf.aminer.org/000/331/442/a_robust_algorithm_for_text_extraction_in_color_video.pdf (2006)
13. Moradi, M., Mozaffari, S., Orouji, A.A.: Farsi/Arabic Text Extraction from Video Images by Corner Detection. In: 6th Iranian Conference on Machine Vision and Image Processing, pp. 1–6 (2010)
14. Moradi, M., Mozaffari, S., Orouji, A.A.: Farsi/Arabic Text Extraction From Video Images. In: 19th Iranian Conference on Electrical Engineering, pp. 1–6 (2011)
15. Halima, M. B., Karray, H., Alimi, A. M.: A Comprehensive Method for Arabic Video Text Detection, Localization, Extraction and Recognition. In: Advances in Multimedia Information Processing, LNCS, vol. 6298, pp 648–659 (2010)
16. Halima, M. B., Karray, H., Alimi, A. M., Vila, A. F.: NF-SAVO Neuro-Fuzzy System for Arabic Video OCR. International Journal of Advanced Computer Science and Applications. 3(10), 128–136 (2012)
17. Sáez,E., González,J.M., Palomares,J,M., Benavides,J.I., Guil, N.: New Edge-based Feature Extraction Algorithm for Video Segmentation. In: SPIE Proceedings on Image and Video Communications and Processing, 5022, pp. 861–872 (2003)
18. Koprinska, I., Carrato,S.: Temporal Video Segmentation: A Survey. Signal Process-Image.16, 477–500 (2001)
19. Thompson, A.: Hierarchical Segmentation of Videos into Shots and Scenes using Visual Content. Thesis, School of Information Technology and Engineering Faculty of Engineering University of Ottawa, Canada (2010)
20. FFmpeg, http://www.ffmpeg.org
21. VideoLAN, http://www.videolan.org/vlc/index.html
22. Stanislav, P., Jurgen, H., Lei, Z.: Image Noise Level Estimation by Principal Component Analysis. IEEE T Image Process. (2012)
23. Ce, L., William T. F., Richard, S., Sing, B. K.: Noise Estimation from a Single Image. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, pp. 901–908 (2006).
24. Karibasappa, K.G., Shivarajkumar, H., Karibasappa, K.: Neural Network Based Noise Identification in Digital Images. ACEEE Int. J. on Network Security.2(3), 28–31 (2011)

Trajectory Based Unusual Human Movement Identification for Video Surveillance System

Himanshu Rai, Maheshkumar H. Kolekar, Neelabh Keshav, and J.K. Mukherjee

1 Introduction

Intelligent video surveillance, an area of computer vision is becoming important mainly because it is effective way of safeguarding life and property [1], [2], [3], [4].Very few attempts have been made to recognize and classify abnormal or abnormal trajectories. In [5], an attempt has been made to identify two distinct trajectories spiral and closed, based on the distribution of points in space. The limitation is the small number of trajectories that can be identified. In [6], the authors have made an attempt for wandering trajectory recognition by method of angles. In this method they connect points in subsequent frames to obtain line segments and then on the basis of the angles between the line segments they classify the trajectories into three distinct categories. The problem in this method is that it requires very robust tracking and if the tracking is not robust enough then stray points may result into a very different pattern of angles that might result into wrong identification. Probabilistic graphical models for video processing such as Markov random field, Hidden Markov models [7], Bayesian Belief networks [8] have proved to be very powerful for video object tracking [9].

Our method for trajectory recognition is based on obtaining the Motion History Image (MHI) [10] of the centroid of the moving person and classification based on hu-moments. This method produces accurate results even when tracking is not very robust. This is because a few stray points won't affect the MHI much and the hu-moments will be fairly the same. In addition to this, our system can detect whether a person is entering a prohibited area where he/she is not supposed to enter. The last stage of the system calculates the speed in units of meter per second unlike other works that calculate speeds in terms of pixel distances. We followed an exhaustive testing procedure using 180 video clips taken in real life conditions. In order to reliably detect only the moving objects, optical flow field based methods such as flux tensor [11] are widely used. For tracking of the person we have also used optical flow method which requires less number of parameters to be provided by the programmer. The proposed system is robust and dynamic scheme that has been proposed so far produces very high success rate.

2 Proposed Method

In order to perform surveillance, we used a network of cameras at sensitive and strategic locations which provided real time video data. This data was analyzed by our algorithm to detect abnormal activity based on trajectory of the subjects. The flowchart in Fig. 1 indicates the procedure that we followed. Trajectory extraction was performed on the video clip using optical flow method and it was ascertained whether the target is entering prohibited area. Entry into such an area was immediately classified as abnormal activity. Multiple restricted areas can be defined by the administrator of the system. If the subject's trajectory does not enter the restricted area then we subject it to trajectory classification. Here we determined if the path that the subject followed had "normal trajectory" or "abnormal trajectory". "Normal" and "abnormal" trajectories considered in our experimentation are shown in Fig. 2. A abnormal trajectory is immediately identified as "abnormal activity", shown by the subject. So any trajectory which is spiraling or is helical (f, g, i) or is a closed circular path (h) is treated as abnormal. Now the normal trajectory is usually linear or has at most one loop (a-d), which might be the case if the moving subject drops his belonging and comes back to collect it forming a loop. But if there are more than one loops involved in the normal trajectory then it is an abnormal activity and we can

H. Rai (✉) • M.H. Kolekar • N. Keshav
Department of Electrical Engineering, IIT Patna, Patna, India
e-mail: himanshu.ee10@iitp.ac.in; mahesh@iitp.ac.in; neelabh.mtee13@iitp.ac.in

J.K. Mukherjee
EISD, Bhabha Atomic Research Center, Mumbai, India
e-mail: jkmukh@yahoo.co.uk

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_114,

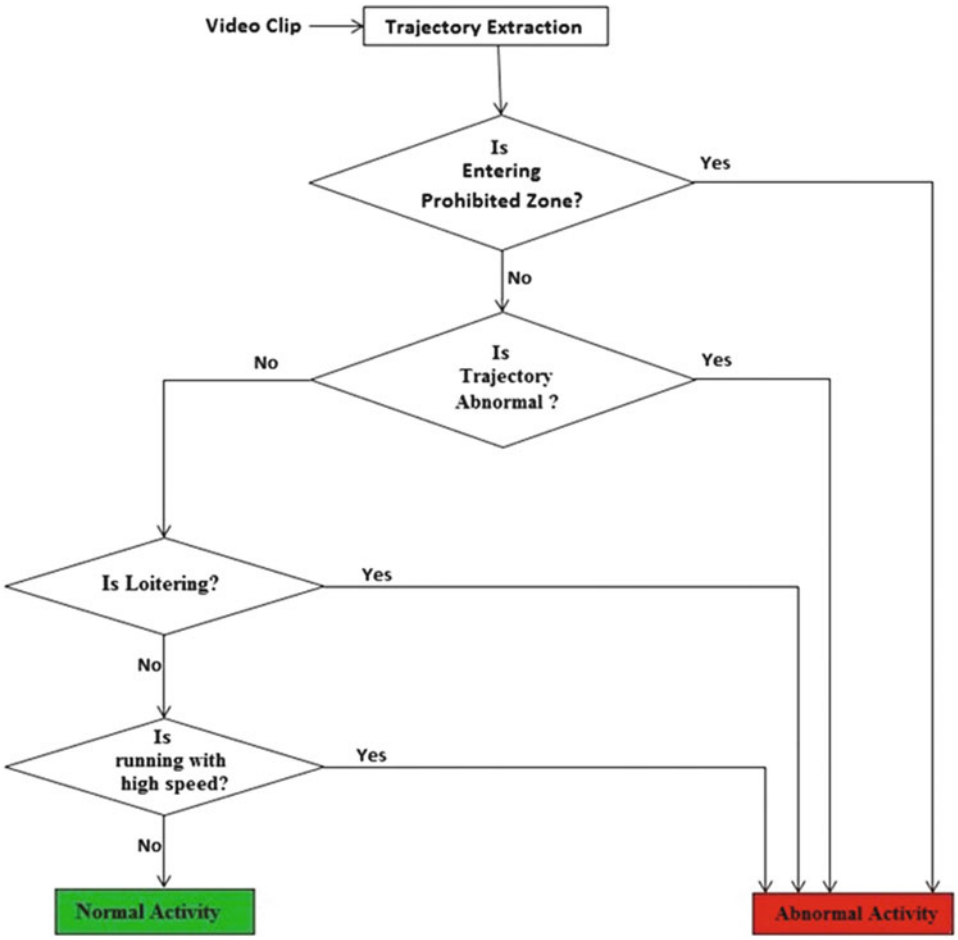


Fig. 1 Process of trajectory based abnormal Human Activity Detection

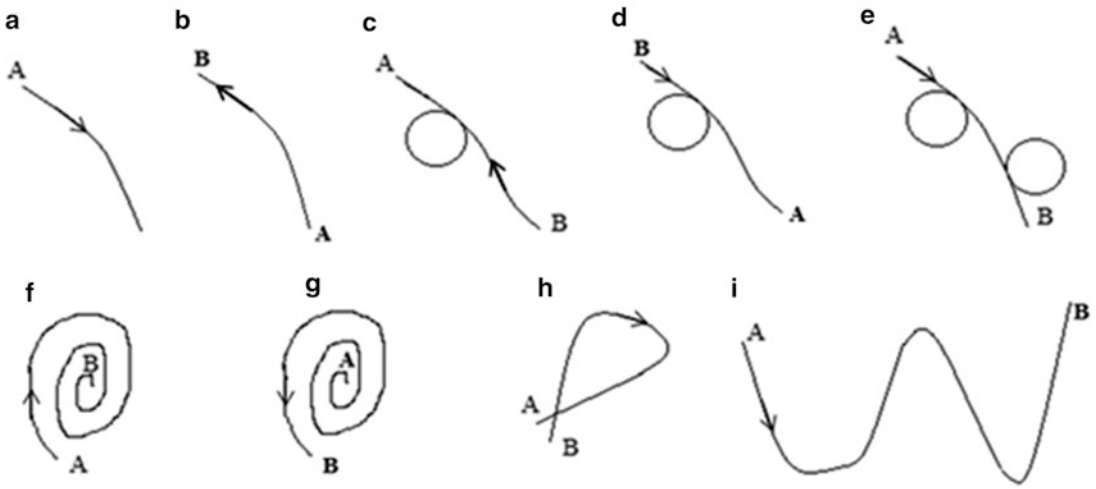


Fig. 2 Trajectories of human motion; (a)-(d): normal trajectories; (e): loitering in normal trajectory; (f)-(i): abnormal trajectories

say that the subject is loitering around (e). If the subject is not loitering around then we calculate the speed of the subject. A threshold value for speed was set. Subject with speed lesser than the threshold was classified as showing "normal activity" else deemed as showing "abnormal activity". All abnormal activities can be used to raise alarm by flashing the screen to inform the guards to verify if everything is alright.

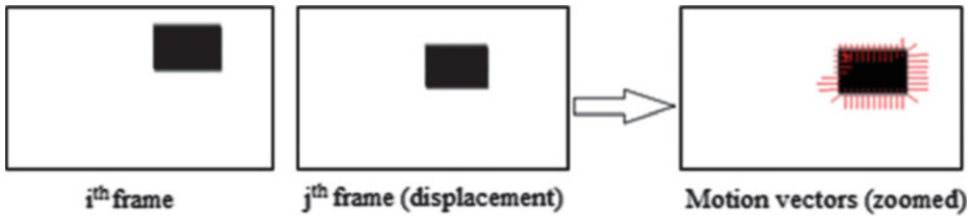


Fig. 3 Motion field

2.1 Motion field

The motion field is the projection of the 3D scene motion into the image. The motion field was calculated using Horn-Schunck algorithm [12] by estimating optical flow. Fig.3 shows the optical flow vectors for a small displacement of an object. On the basis of a threshold moving object can be recognized.

2.2 Algorithm for the proposed system

Trajectory Extraction.

1. Individual Frame extraction from the video sequence was done.
2. Converting the frame image from RGB to intensity image-The RGB image was converted to an intensity image (similar to gray scale).
3. Median filter was applied to perform noise reduction. Next we performed closing operation to fill the gaps between the blobs. It is defined as: In mathematical morphology, the closing of a set (binary image) A by a structuring element B is the erosion of the dilation of that set,

$$A \cdot B = (A \oplus B) \ominus B \tag{1}$$

where *oplus* and $\ominus$ denote dilation and erosion, respectively.
4. The next operation that was performed is erosion operation to thin out pieces of unwanted objects.
5. The optical flow was compared with threshold value to obtain blob of moving subject. Opening operation removed small objects, while closing removed small holes.
6. The centroid of the blob was obtained and a matrix was created to store the centroid coordinates. This locus of the centroids is plotted and used for trajectory analysis.

Restricted area entrance classification. Here we consider an area or multiple areas in the zone to be restricted. If the trajectory of the subject passes through the restricted area in the zone, we classify the activity to be not legal. Now this restricted area in real life could be a vault located in a bank. Normally a person is not supposed to enter such an area at night time. So any intrusion in such an area is classified as abnormal and this activity should raise an alarm. So here if the wandering trajectory of the person enters a user defined rectangular (or any shape) prohibited area, then the screen is flashed to indicate a warning.

Trajectory analysis. A trajectory is the path that a person moves as a function of time. The trajectory in a scene is often used to analyze the activity or behavior of the person. Moreover, the loitering behavior can be easily inferred by analyzing the trajectories. We used trajectory to estimate what kind of behavior is being dealt with. We broadly classified the trajectories into two categories: normal and abnormal as shown in Figure 2. Figure (a-b) show normal walking in any direction. Figure (c-d) show normal walking with at most one loop in any direction. Fig. e shows normal walking with looping which is a loitering activity. Figure (f-g) show spiraling in and out, Figure h shows closed circular path and Figure i shows helical path, all of which are abnormal trajectories.

For classification of the trajectories we have computed MHI. We first obtained the MHI of the centroid in each frame in which the action was taking place. The MHI can be calculated as follows:

$$M(x,y,t)_\tau = \begin{cases} \tau & \text{if} \quad B(x,y)_t = 1 \\ max(M(x,\ y,\ t-1)_\tau - 1, 0) & \text{if} \quad B(x,y)_t = 0 \end{cases} \tag{2}$$

Where B represents the binary image obtained after thresholding. We choose τ in such a way that it covers the full extent of action. We then calculated hu-moments of some representative MHIs. We used the first three hu-moments to classify the trajectories. We observed that the hu-moments are sufficiently far off to establish distinctness of the trajectories. We define a similarity measure as follows:

$$I(A,B) = \sum\nolimits_{i=1\ldots7} \left| m_i^A - m_i^B \right| \tag{3}$$

where,

$$m_i^A = sign\ (h_i^A).log_{10}\ h_i^A \tag{4}$$

$$m_i^B = sign\ (h_i^B).log_{10}\ h_i^B \tag{5}$$

Here A and B represent the two MHIs between which we want to calculate the similarity measure. 'm' represents the moment values [13]. We scaled the moments because the later moments are very small and their values have very low significance value. So this difference was computed between the MHI of the trajectory that is to be recognized and each of the representative MHIs. The one representative MHI which gives the least value of difference is declared to resemble the trajectory of the subject. So if the trajectory qualifies as normal, then we subject it to the next level which is the speed estimation.

Speed Estimation. After the trajectory of the subject has been classified as normal, we ensure that the person is not pacing. We can calculate the speed of person based on pixel distance and frame rate. Consider (p, q) be the centroid of the person at the start of the trajectory in frame i and (r, s) be the centroid of the person at the end of trajectory in frame j. So, total pixel distance travelled D_p is

$$D_p = \sqrt{(p-r)^2 + (q-s)^2} \tag{6}$$

and, speed estimated using pixel distance S_p is

$$s_p = \frac{D_p}{T} \tag{7}$$

where T is total time taken to cover D_p is given by

$$T = \frac{total\ no.\ of\ frames\ travelled}{frame\ rate} \tag{8}$$

This speed has been estimated with the help of pixel distance and hence, is not the actual speed. We have extracted multiplication factor 'k' to get actual speed, S_A based on following training procedure: First we manually measure time for known distance to compute actual speed in training sequence S_{at}. Next, we compute the speed based on pixel distance using equation (7) and calculate the multiplication factor 'k_t' as

$$k_t = \frac{s_{at}}{s_p} \tag{9}$$

In order to find out the error with k factor, we repeated this procedure for different number of video clips. The mean value of multiplication factors (k_t), i.e., 'k' is used to find the exact speed of moving person

$$S_A = S_p > \tag{10}$$

The method suggested in section 2.2 was used to estimate the speed of moving target. A ratio of actual speed and the speed of target as obtained by algorithm was found as k_t where $k_t = k' \times 10^{-7}$. This value of k' is tabulated in table 1. The mean of k' is 7.6984 while the variance is 0.6329. Hence, it can be inferred that the technique is highly efficient in predicting speed of moving target with high level of accuracy. We defined the average of this k_t as mean multiplication factor 'k'. This 'k' when added to algorithm successfully estimated the speed of moving object in meter per second with high level of accuracy. It should be noted that the value of 'k' depends on orientation and location of camera. Thus, finding the value of 'k' is a part of installation of camera. Once the 'k' is determined for the location, the system will predict speed of person unless the position of camera or orientation is changed. The average running speed of human considered here is 5 meter/second. If the subject moves at a higher speed, we consider it as abnormal activity. After setting this threshold value the video clips were run and the results are compiled in Table 2.

Table 1 Tabulation of "k'"

Sl. No.	k'	Deviation from mean for k'
1	7.9787	-0.2803
2	7.2368	0.4616
3	8.6930	-0.9946
4	6.8340	0.8644
5	7.7898	-0.0914
6	8.3456	-0.6472
7	8.1212	-0.4228
8	7.6543	0.0441
9	7.7981	-0.0997
10	6.5321	1.1663

3 Result and Discussion

The proposed method was verified by 180 video clips and the results are mentioned in Table 2.

1. For the first activity, in one of the instances, a fast moving object touched the tip of prohibited zone (but the centroid didn't) and screen didn't flash red, but high speed was immediately detected and alarm was raised. So the speeding object couldn't mislead the system. The system is robust in detecting normal trajectories.
2. In a few clips, where the subject followed the "loitering motion in normal path", the system identified it as "closed path" as hu- moments obtained for normal loitering and closed path were close but instances of such false detections are very low.
3. The method is highly efficient in detecting a fast moving subject and estimated his speed with high level of accuracy.
4. The method is again highly efficient in detecting objects moving in closed path.

Table 2 Evaluation of Recognition Results

Sl. No.	Human Activity in clip	No. of clips	Successful response	Success %
1	Entering a prohibited zone	20	19	97.5
2	Moving in normal path (with no loitering)	40	40	100
3	Subject loitering in normal path	40	37	92.5
4	Subject running in normal trajectory	40	40	100
5a	Subject moving in a spiral trajectory	20	18	90
5b	Subject moving in a closed trajectory	20	20	100
5c	Subject moving in a helical trajectory	20	17	85

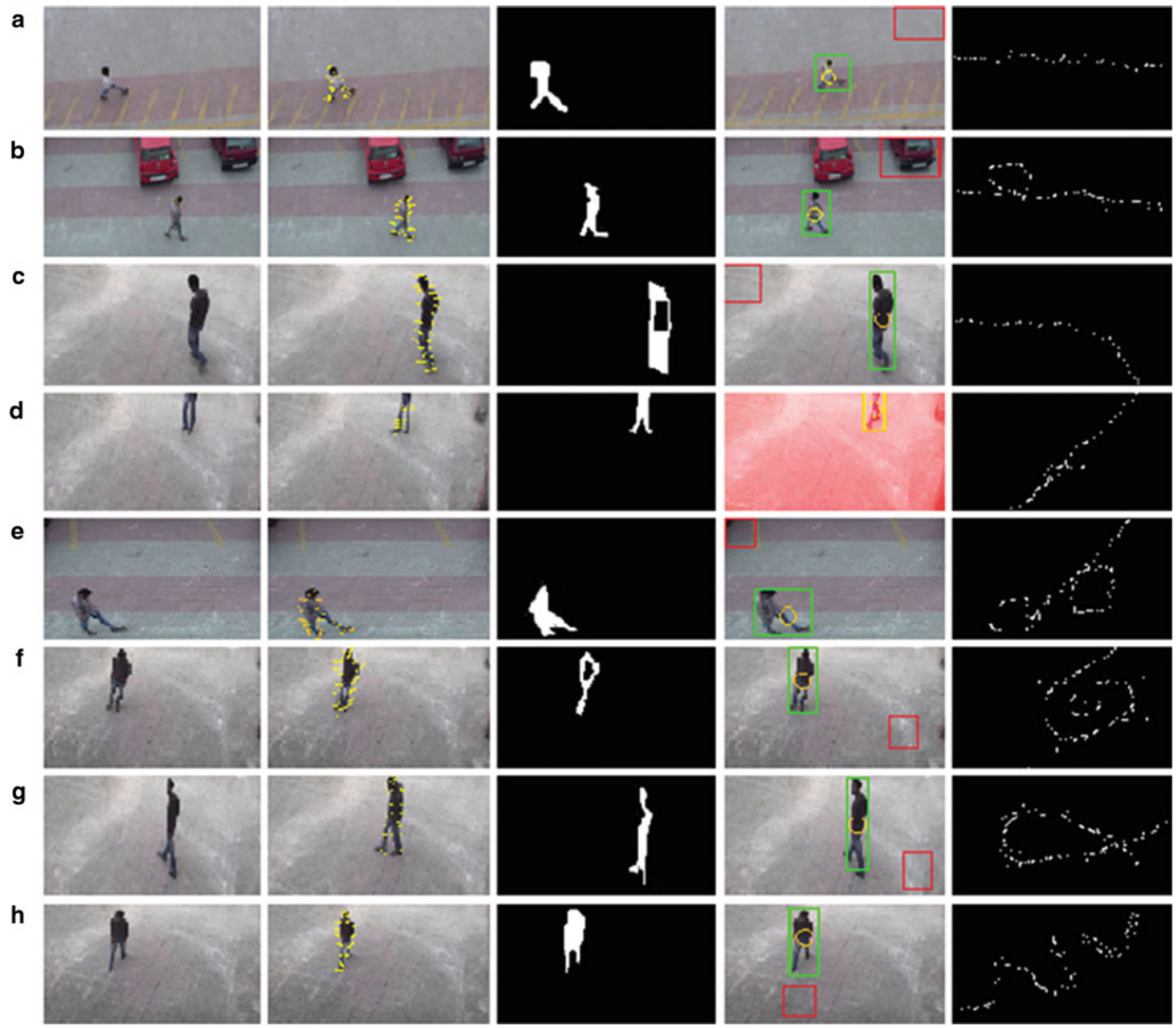


Figure 4 Column1: instances of video clips; Column2: motion vectors; Column3: extracted blob from the video sequences; Column4: the output of algorithm wherein green box with a yellow circle shows the tracked moving target and its centroid respectively; Column5: obtained trajectories

5. Spiral trajectory was detected with good accuracy. For some cases the spiral trajectory was wrongly classified as closed. But still the classified trajectory is abnormal and the resultant trajectory is classified as abnormal.

The proposed system was verified using video clips taken in real life conditions. The result is compiled in the Figure 4. In (a), (b) and(c) the subject passes the four levels hierarchical testing and classified as showing “normal activity”. In (d)

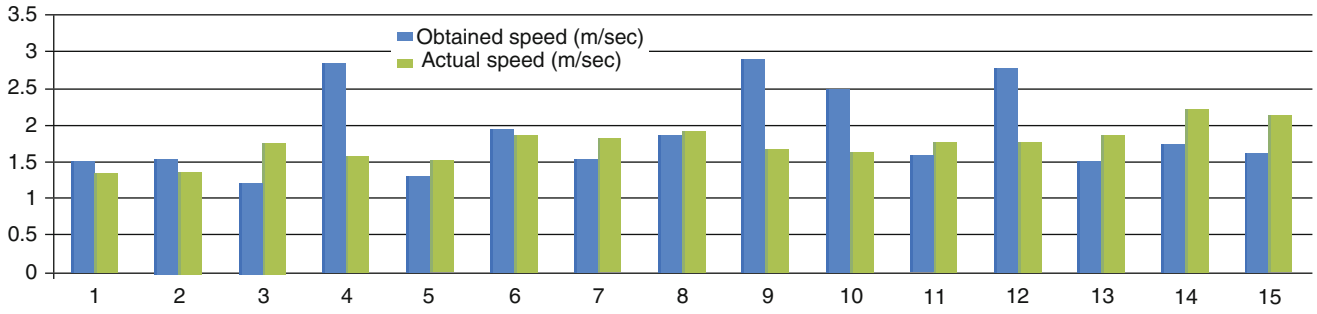


Figure 5 Comparison between obtained and actual speeds

subject enters prohibited area leading to an immediate flashing of screen (red) to indicate alarm. In (e) the subject fails to satisfy third level of hierarchical system and its activity is classified as abnormal. In (f), (g) and (h) subject fails the test at second level. The proposed algorithm was also used to judge subject's actual speed using equation-10. The graph showing actual velocity vs. calculated velocity is shown in Figure 5.

4 Conclusions

We obtained the trajectory of a moving person and used it determine (i) if the person was entering a prohibited area and (ii) whether the person's trajectory is abnormal or not. The algorithm was also successful in calculating the speed of the moving person in m/s to ensure that the person is not pacing in the zone, where such an activity is prohibited. All the above experiments were performed using a static camera with different backgrounds in real life conditions. The algorithm is robust and produces accurate results. In future we will work on incorporating other improvements like using a dynamic camera. We believe that the system so developed will be successful in detecting abnormal activities, especially around severely vulnerable places like bank vaults, ATMs and military installations.

References

1. Collision, P.A, "*The application of camera based traffic monitoring systems*", IEEE Seminar on CCTV and Road Surveillance, pp. 8/1–8/6, 1999
2. Remagnino P, Tan T, Baker K, "*Agent orientated annotation in model based visual surveillance*", IEEE Int. Conference on Computer Vision, Bombay, India, pp.857–862, 1998.
3. A K Singh Kushwaha, Maheshkumar H Kolekar, A Khare, "*Vision based method for object classification and multiple human activity recognition in video surveliance system*", CUBE Int. Information Technology Conference, 47-52, India, 2012
4. A K Singh Kushwaha, O Prakash, A Khare, Maheshkumar H Kolekar, "*Rule based human activity recognition for surveillance system*", 4th IEEE Int. Conf. on Intelligent Human Computer Interaction, India, 2012
5. Zhang, Y, Liu Z, "*Irregular behavior recognition based on treading track*", Int. Conf. on Wavelet Analysis and Pattern Recognition, Beijing, China, vol. 3, pp. 1322–1326, 2007
6. Qingzhang C, Rongjie W U, Yunfeng N I, Ruohong H, Zhehu W, "*Research on Human Abnormal Behavior Detection and Recognition*", Intelligent Video Surveillance, Journal of Computational Information Systems, vol 9(1), pp. 289–296, 2013.
7. Maheshkumar H Kolekar, S Sengupta, "*Hidden markov model based video indexing with discrete cosine transform as a likelihood function*, IEEE INDICON conference, pp. 157-159, 2004
8. Maheshkumar H. Kolekar, "*Bayesian Belief Network Based Broadcast Sports Video Indexing*", Int. Springer Journal of Multimedia Tools and Applications, 54:27-54, 2011.
9. Dore, A, Soto, M, Regazzoni, C.S, "*Bayesian Tracking for Video Analytics*", IEEE Signal Processing Magazine, vol 27 (5), pp. 46-55, 2010.
10. A. F. Bobick, J. W. Davis, "*The Recognition of Human Movement Using Temporal Templates*", IEEE trans. on pattern analysis and machine intelligence, vol. 23, no. 3, 2001.
11. Maheshkumar H. Kolekar, K. Palaniappan, S. Sengupta and G. Seetharaman "*Semantic Concept Mining based on Hierarchical Event Detection for Soccer Video Indexing*", Int. Journal on Multimedia, vol-4 (5), pp. 298-312, 2009
12. Berthold K. Horn, Brian G. Schunck, "*Determining Optical Flow*", Techniques and Applications of Image Understanding, SPIE 0281,319, 1981.
13. M. K. Hu, "*Visual Pattern Recognition by Moment Invariants*", IRE Trans. on Information Theory, 1962

Special Session: From Boolean Problems to the Internet of Everything

Design and Implementation of Novel Algorithms for Frequent Pattern Trees

R. Siva Rama Prasad, N.S. Kalyan Chakravarthy, and D. Bujji Babu

1 Introduction

During the past decade, the entire business scenario has been tremendously changed, because of rapid change in the business policies and the organization standards. The ultimate change of their objectivity is to retain customers and the business. The innovations and globalization have generated great opportunities in business and choices in the universal marketplace for firms and customers. The organizations began understanding their customers and their interest to fulfill the requirements of them. As a part of understanding the customers entirely, the organizations are collecting various details of the customers for analysis purpose. The collected data is primarily stored in the fashion of a database, and it allocates a unique identification number to the customer. This kind of primary information is useful to the organizations in maintaining the relationship with the customers. In Data Mining Association rule learning is a popular and well researched technique for discovering interesting relations between variables in large databases. Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence [1][2][3]. Association rules can be extracted using two familiarized algorithms named as Apriori algorithm and FP-Growth algorithm[15][16][17]. The FP-Growth algorithm is completely depends on fp-tree[4]. The previous fp-tree node is labeled only with its current support count, which consumes more time while traversing to extract association rule. In this paper we are more concentrated on design and implementation of novel algorithms for frequent pattern trees. By using the proposed algorithms the traversal time is reduced. In this paper we present six efficient algorithms.

A. ***Frequent Item or Frequent Item set***
An item is said to be a frequent, if it is purchased by minimum number of customers. In other words in any transactional database DB contains any item repeated occurred then the particular item is frequent item. If set of items are repeatedly occurred then that set is called a frequent item set. Generally, these frequent items can be determined using the statistical approach mode. If the data set has only one mode then that is unimodality, if it has two modes the bimodality, if it has three modes then it is called trimodality.

B. ***Association Rule :***
Let D be the database of transactions and ITEMSET = {I1,I2,I3,.. …, In} be the set of items. TR is a transaction includes one or more items in ITEMSET (i.e., TR $\subseteq$ ITEMSET). An association rule has the form P $\Rightarrow$ Q, where P and Q are non-empty sets of items that is P $\subseteq$ ITEMSET, Q $\subseteq$ ITEMSET such that P $\cap$ Q = Ø.

2 Literature Review

A. ***Association Rule Mining(ARM) :***
Agarwal et al.(1993)[1] reported frequent item set mining and association rule mining first time. One of the first algorithms proposed for association rule mining was the AIS algorithm. The Apriori algorithm employs the downward closure property that is if an item set is not frequent any superset of it cannot be a frequent. FP-Growth was developed by Grahne & Zhu and it uses an extra array-based structure to decrease the number of traversals of the tree that are required during the analysis. This saves time during general traversal of the tree and also enables direct initialization of the next level of the

R.S.R. Prasad (✉)
Dept. of I.B Studies, Acharya Nagarjuna University, Guntur, A.P, India
e-mail: raminenisivaram@yahoo.co.in

N.S.K. Chakravarthy
QIS College of Engineering & Technology, Ongole, A.P, India
e-mail: suryaetg@yahoo.com

D.B. Babu
Dept. of Computer Science and Engineering, Prakasam Engineering College, Kandukur, A.P, India
e-mail: bujjibict@gmail.com

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_115,

FP-Tree(s) [A. Ceglar & J. F. Roddick, 2006; G. Grahne & J. Zhu, 2003]. Another FP-Growth based approach is TD-FP-Growth proposed by Wang et. al., and is a top-down variation to the base FP-Growth approach. This approach is said to alleviate the need or demand to generate/build conditional pattern bases and physical projections of the trie [A. Ceglar & J. F. Roddick, 2006; K. Wang et. al., 2002[5][6][7][8][9]].

B. ***Trees :***

James Joseph Sylvester introduced the term graph in 1878 [10]. Arthur Cayley conducted a study on a particular type of graphs which are called Trees. The study causes several implications in theoretical chemistry. These results were published by George Polya in 1935. Denes Konig wrote the first text book on the graph theory in 1936. In 1962, AVL Trees was invented by G.M.Adelson, Velskii and E.M.Landis[11]. Heinrich Heesch, in 1969 published a computerised problem solving method. Rudolf Bayer and Edward.M first described the B Trees in 1972.

3 Research Objective

The main objective of the research is to reduce the tree traversal time of a frequent pattern tree. Design and implementation of novel algorithms for new frequent pattern tree with new naming approach. To prove how the novel algorithms are effectively working with an example.

4 Proposed Algorithms

A. Algorithm for fp-tree construction:

Algorithm : Novel Fp_tree construction
Input : *Trans_list* database DB
Output : Frequent pattern tree

1. Read Transaction_list into *Trans_list* from database DB.
2. Let *Freq_items* be the set of frequent items based on threshold value in *Trans_list*.
3. Sort *Freq_items* on descending order of *support_count* and let the result be *F_Sort_List*.
4. Create a root node of an FP_Tree T for *Trans_list* and label it as 'NULL'.

For each transaction *Trans* in *Trans_list* do the following

a) Select the frequent items from *Trans* and the result be *F_items*.
b) Sort *F_items* according to the order of *F_Sort_List* and the result be h|R, where h is the first (head) element and R is the remaining elements of the list.

Call **insert_tree(h|R,T)**.

B. Proposed Algorithm for insert_tree with Two Level Node Labeling

Algorithm : insert_tree with Two Level Node Labeling
Input : each transaction and Tree
Output : Frequent pattern tree with two level node labels

1. If T has a child N such that N.item-name = h.item-name, then increment N 's count by 1;
2. else
   {
   create a new node N
   N's count initialized to 1,
   }
3. N's pred data = T's present data
4. N's parent link linked to T
5. N's node-link linked to the nodes with the same item-name via the node-link structure.
6. If P is nonempty, then call insert tree(R, N) recursively.

C. Novel Algorithm to extract n frequent items from FP_growth tree

Algorithm : NovelFP-growth
Input : Tree, Key element a and n.
Output : n frequent patterns associated with the given key element

Procedure NovelFP-growth(Tree, a,n)
{
//a is the key item
//n indicates the no of items associated with a particular item.
Step 1: call Traversal_SPP_and_MPP(Tree, a,n)
Step 2: return(freq pattern set(SPP) ∪ freq pattern set(MPPset) ∪ (freq pattern set(SPP) × freq pattern set (MPPset)))
}

D. Proposed Algorithm NovelFP-growthSpp

Note : X represents a pattern in single predecessor path and Y represents a pattern in Multi predecessor path.
Algorithm : NovelFPgrowth_to_traverse_SinglePredecessor Path
Input : Tree, Key element a and n.
Output : Traverses in single predecessor path and returns the patterns associated with the key

NovelFP-growthSpp(Tree, a,n)
{
Step 1: let SPP be the Single Predecessors path of n-1 nodes part of Tree;
Step 2: for each combination (denoted as X) of the nodes in the path SPP do
Step 3: generate pattern X ∪ a with support = minimum support of nodes in X;
Step 4: let freq pattern set(SPPSet) be the set of patterns so generated;
}

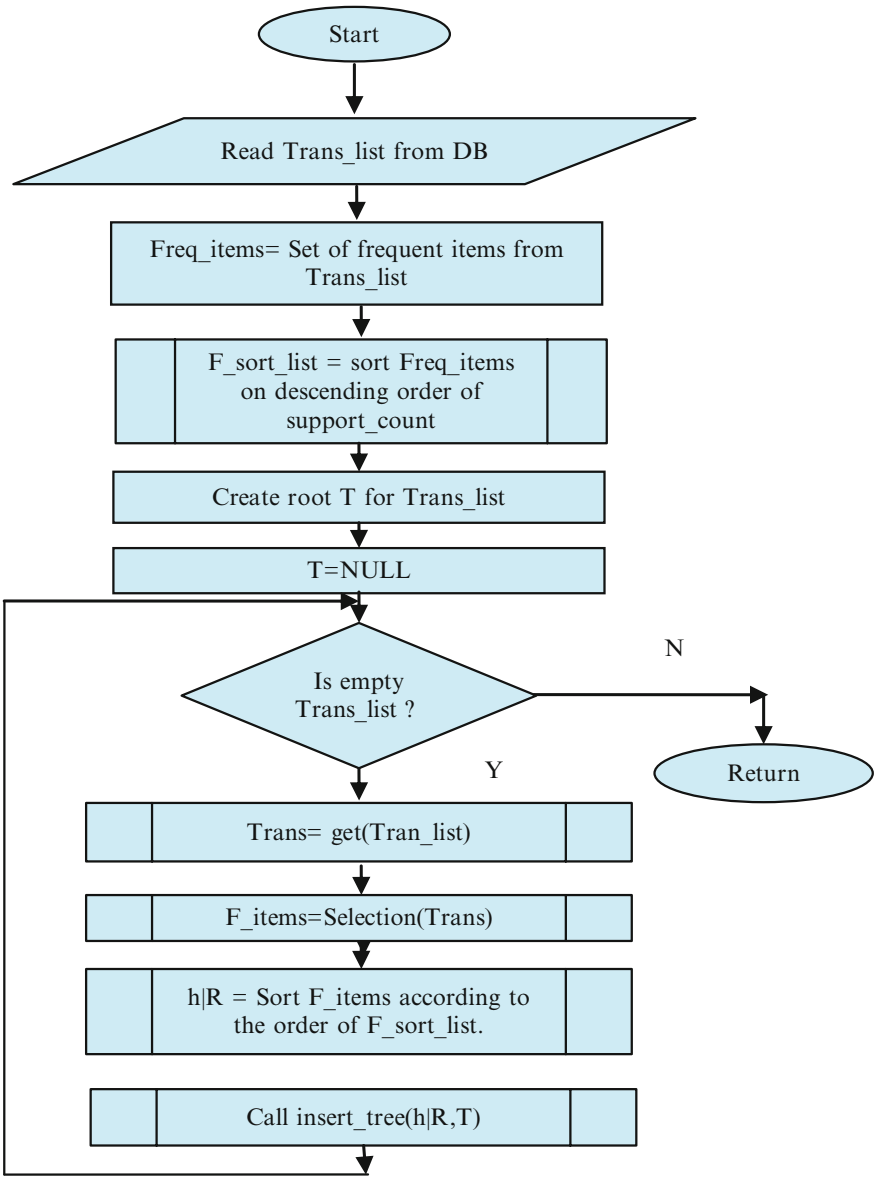


Figure: 1 Flow chart for Novel Fp_tree construction Algorithm

E. Proposed Algorithm NovelFP-growthMpp

Algorithm : NovelFP-growth_to_traverse_MultiPredecessor Path

Input : Tree, Key element a and n.

Output : Traverses in multi predecessor path and returns the patterns associated with the key

NovelFP-growthMpp(Tree, a,n)

{

1. let MPP be the Multi Predecessors Path part of Tree with i=1;
2. for each item a_i in MPP with i <= n-1 do

{

3. generate pattern Y = ai ∪ a with support = ai.support;
4. i=i+1

}

5. construct Y's conditional pattern-base and then Y's conditional FP-tree Tree Y;
6. if Tree Y ≠ Ø then
7. call NovelFP-growth(Y, a, n);
8. let freq pattern set(MPPSet) be the set of patterns so generated;

}

F. Proposed Algorithm Traversal_SPP_and_MPP

Algorithm : Traversal_SPP_and_MPP

Input : Tree, Key element a and n.

Output : Traverses either in single predecessor path or in multi predecessor path and returns the patterns associated with the key

//Freq pattern set SPPSet and MPPSet are global variables.

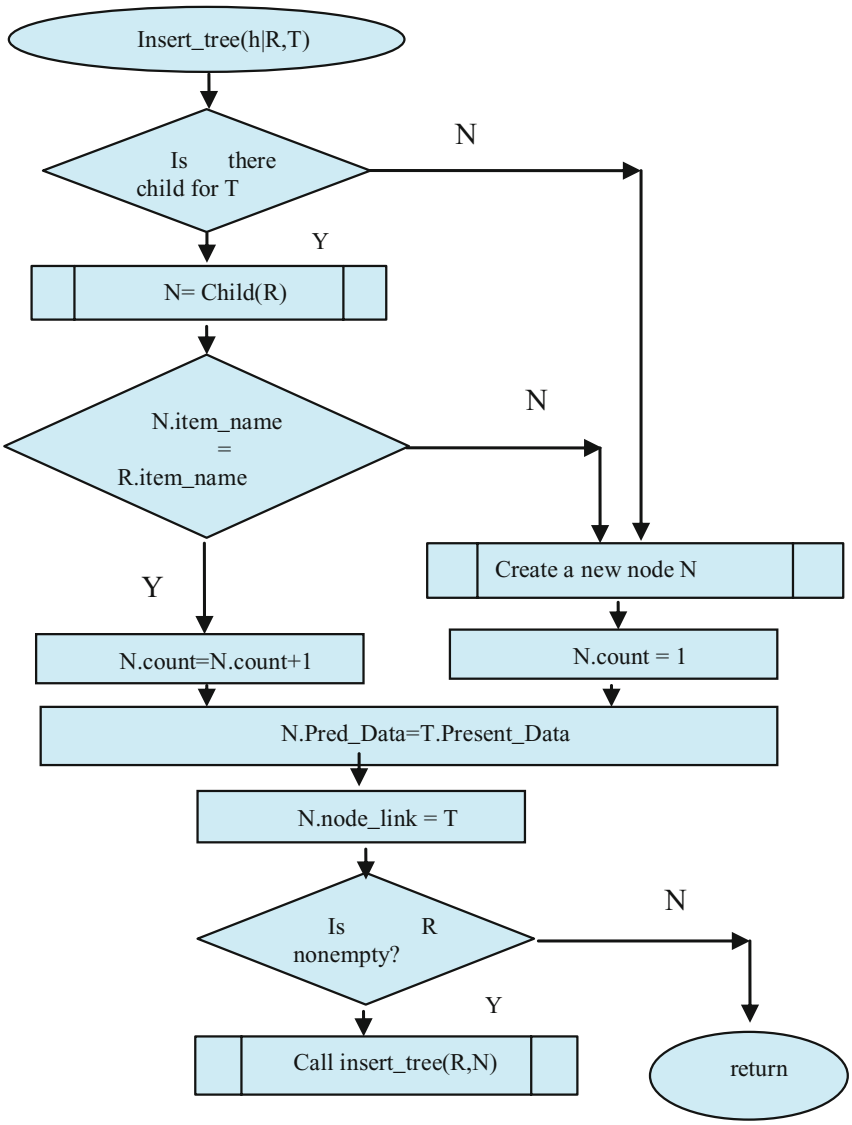


Figure: 2 Flow chart for insert_tree with Two Level Node Labeling Algorithm

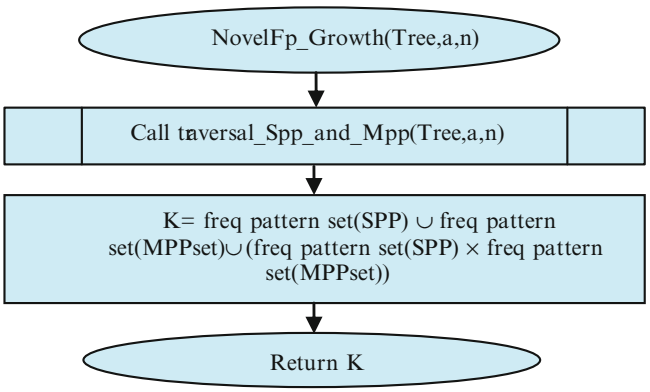


Figure: 3 Flow chart for NovelFP-growth Algorithm

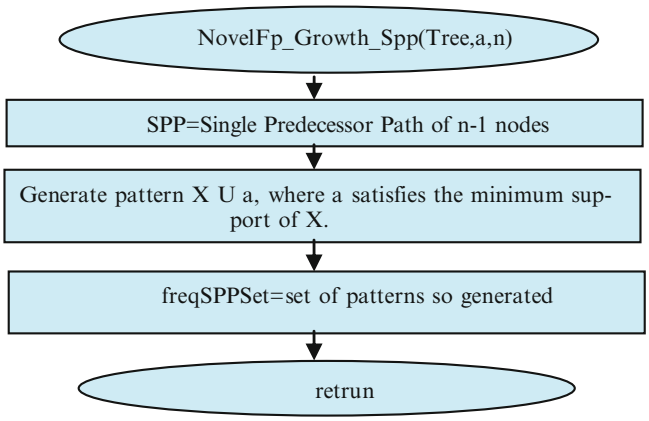


Figure: 4 Flow chart for NovelFP- growth_to_traverse _SinglePredecessorPath

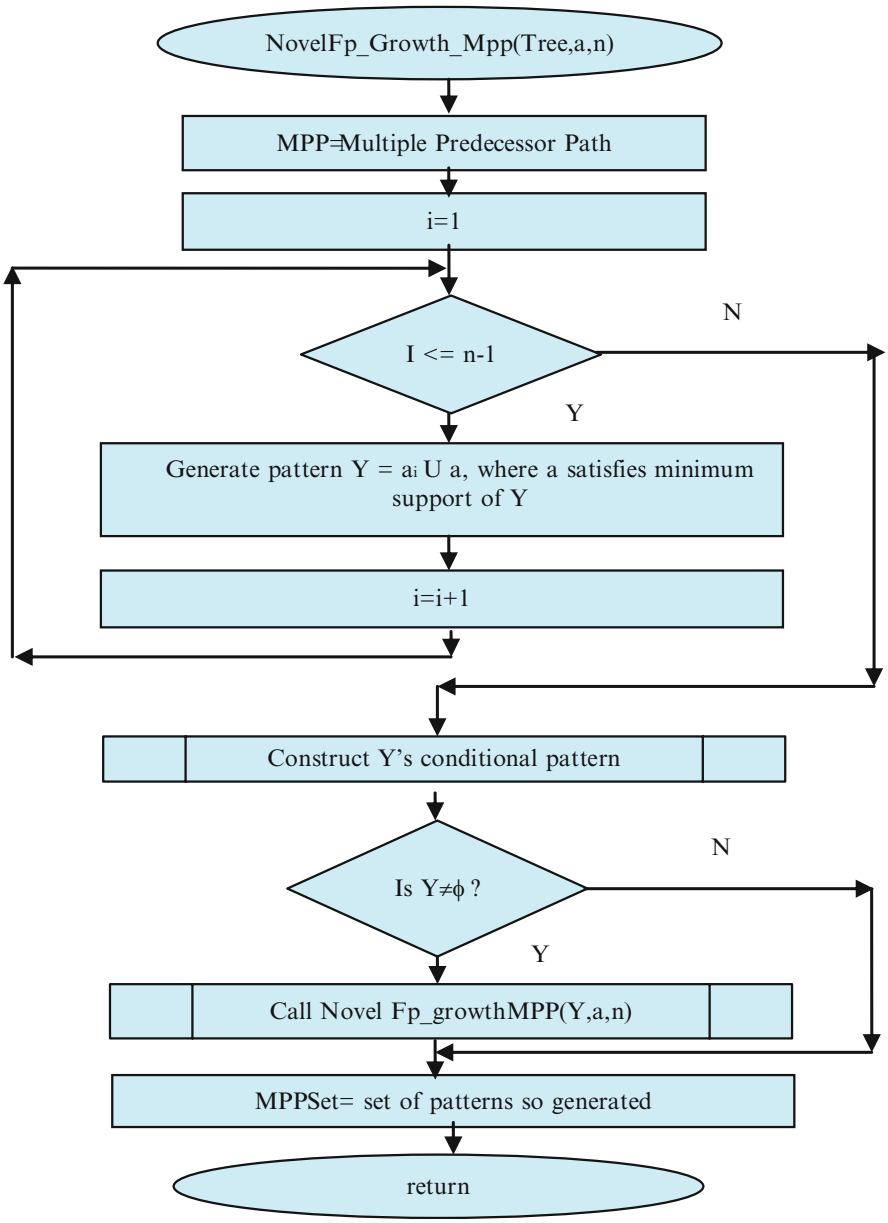


Figure: 5 Flow chart for NovelFP-growth_to_traverse_ MultiPredecessorPath Algorithm

Procedure Traversal_SPP_and_MPP(Tree, a,n)
{
//n indicates the no of items associated with a particular item.
Step 1: if Tree contains a single Predecessors path then
Call NovelFP-growthSpp(Tree, a,n)
else
Call NovelFP-growthMpp(Tree, a,n)
Step 2: return(freq pattern set(SPPSet) ∪ freq pattern set(MPPset)
∪ (freq pattern set(SPP) × freq pattern set (MPPset)))
}

5 Performance and results

Theorem : The time complexity of The Two-level node labeling algorithm(s) is O(n) with reduction of one unit of time in worst case also.

Proof : The proposed Two Level Node Labeling algorithm constructs a frequent pattern tree which is similar to binary tree. Hence, the time complexity of traversal of frequent pattern tree, which is formed with the proposed algorithm, is similar to the traversal complexity of a binary

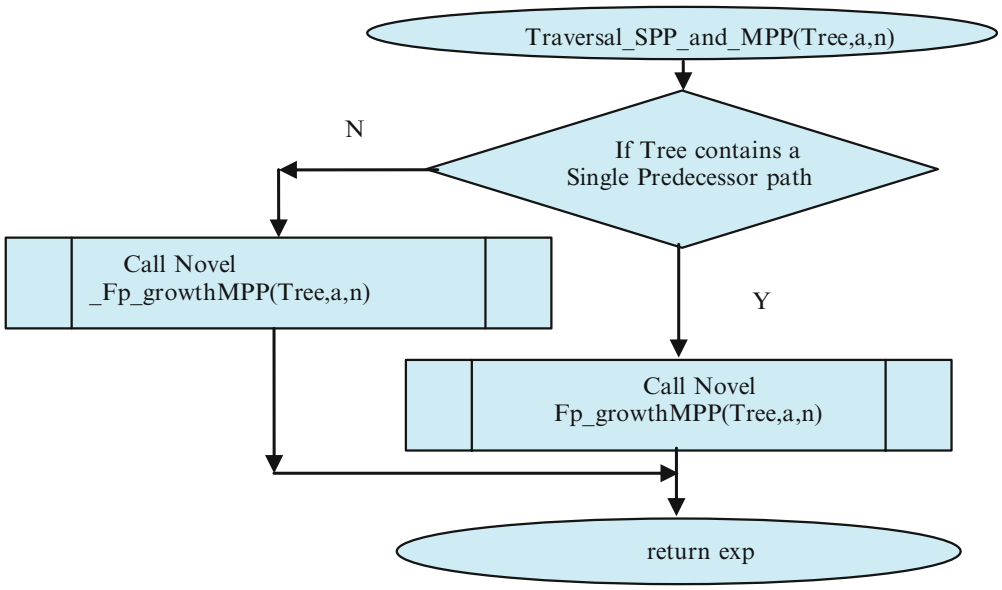


Figure: 6 **Flow chart for Traversal_SPP_and_MPP Algorithm**

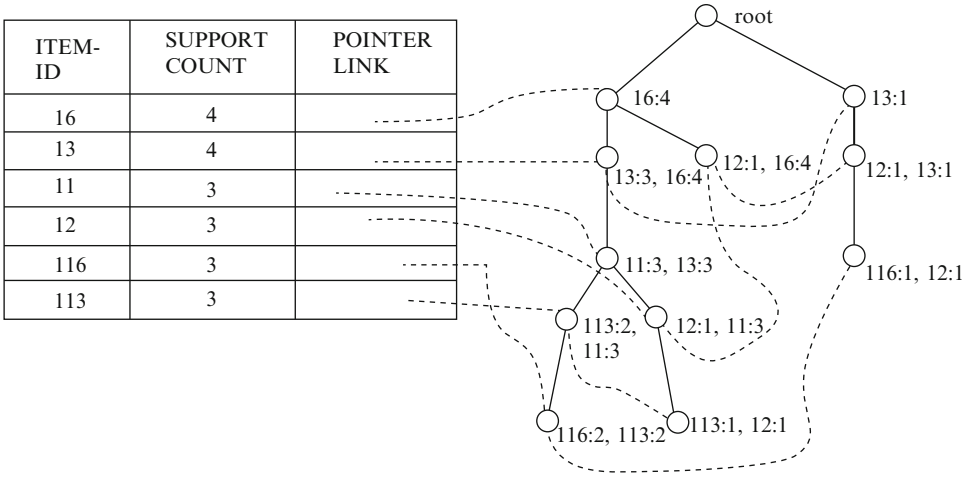


Figure: 7 **Frequent item Transactional Data Base**

Table:1 **Transactional Data Base**

TID	ITEMS PURCHASED
T100	I6, I1, I3, I4, I7, I9, I13, I16
T200	I1, I2, I3, I6, I12, I13, I15
T300	I2, I6, I8, I10, I15
T400	I2, I3, I11,I17, I16
T500	I1, I6, I3, I5, I9, I16, I13, I14

Table:2 **Frequent item Transactional Data Base**

TID	ITEMS PURCHASED
T100	I6, I3, I1, I13, I16
T200	I6, I3, I1, I2, I13
T300	I6, I2
T400	I3, I2, I16
T500	I6, I3, I1, I13, I16

tree. Since the time complexity of binary tree traversal algorithm is O(n) [13][14], the time complexity of frequent pattern tree is O(n). But the proposed algorithm reduces one level in the tree during traversal.

With taking the training data set as in the table:1 [12]. We get the frequent transactional data base as shown in the table:2 with considering the minimum support value as 3. The final frequent pattern tree will appears as shown in figure:7 by using the proposed algorithms.

The traversal time for the predecessor nodes in frequent pattern trees[18][19][20][21] using the traditional frequent pattern tree approach is only O(n) for the n^{th} level node. The proposed algorithm also constructs a binary tree. Hence the time complexity of the newly constructed tree is also O(n). But, one unit of time is less than the time of traditional approach in worst case. The proposed algorithm takes a fewer number of node comparisons to determine the predecessor nodes and its path than that of traditional approach in best and average cases.

6 Conclusion

In this paper, we presented the design and implementation issues of six novel algorithms. One algorithm constructs a frequent pattern tree. Four algorithms are used to label each node and another algorithm extracts the frequent patterns from the frequent pattern tree. In the Example figures the solid line between the nodes represents the relation between the nodes and the dashed line indicates the pointer link between the same nodes, useful to maintain the cumulative node count in the data structure. We also presented the flow charts of each algorithm.This novel approach reduce one level tree traversal of the tree in the worst case also.

Acknowledgements We are so grateful to Sri. Dr.Kancharla Ramaiah garu,Secretary and correspondent of Prakasam Engineering College, kandukur, for extending his marvelous encouragement and support to do the research with providing the research environment. Last but not least, we are very much thankful to all the authors and co-authors of the reference papers for providing us knowledge about clouds, cloud environment and the data mining techniques particularly about association rule learning process and algorithms.

References

1. Agarwal, R., and Srikanth,R., “Fast Algorithms for mining association rules,” In Proc. Of the International Conference on VLDB-94, Sept.1994,pp.487-499.
2. T.Mitchell.” Machine learning,” Mc Graw Hill, Boston,M.A,1997.
3. J.Han and m.Kamber. “Data Mining: Concepts and Techniques,” Morgan Kaufmann Publishers, San Francisco, 2001.
4. Pang-ning-Tan,Vipin Kumar,Michael Steinbach.” Introduction to Data Mining” Pearson 2007. ISBN 978-81-317-1472-0.
5. Bodon. F, “A Survey on Frequent Itemset Mining”, Technical report, Budapest Univ. Of Technology and Economics, 2006.
6. Cheung D, V.T Ng, A. Fu, and Y.Fu. “Efficient mining of association rules in distributed databases”. *IEEE Trans. Knowledge and Data Engineering*, pp 1-23, 1996
7. Rupali Haldulakar and Prof. Jitendra Agrawal, “Optimization of Association Rule Mining through Genetic Algorithm”, International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 3 Mar 2011, pp. 1252-1259.
8. Manish Saggar, Ashish Kumar Agarwal and Abhimunya Lad, “Optimization of Association Rule Mining using Improved Genetic Algorithms”IEEE 2004.
9. Anandhavalli M, Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K., “Optimized association rule mining using genetic algorithm”, Advances in Information Mining, ISSN: 0975–3265, Volume 1, Issue 2, 2009, pp-01-04.
10. J. J. Sylvester (1878) "On an application of the new atomic theory to the graphical representation of the invariants and covariants of binary quantics, — with three appendices," *American Journal of Mathematics, Pure and Applied*, (1) : 64-90. The term "graph" first appears in this paper on page 65.
11. Adelson-Velskii and E. M. Landis, 1962. "An algorithm for the organization of information". *Proceedings of the USSR Academy of Sciences* 146: 263–266. (Russian) English translation by Myron J. Ricci in *Soviet Math. Doklady*, 3:1259–1263, 1962.
12. D.Bujji Babu,Dr.R.Sivarama Prasad and Y.Umamaheswararao “Efficient Frequent Pattern Tree Construction”,*International Journal of Advanced Computer Research, Volume -4 Number-1issue-14 March-2014*.
13. http://www.geeksforgeeks.org/618
14. Massachusetts Institute of Technology (MIT), "Master Theorem: Practice Problems and Solutions", http://www.csail.mit.edu/~thies/6.046-web/master.pdf.
15. M. Zaki, S. Parthasarathy, M. Ogihara, and W. Li.New Algorithms for Fast Discovery of Association Rules. *Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD’97)*, 283–296. AAAI Press, Menlo Park, CA, USA 1997.
16. G. Grahne and J. Zhu. Efficiently using prefix-trees in mining frequent itemsets. In *FIMI’03, Workshop onFrequent Itemset Mining Implementations*, November 2003.
17. F. Masseglia, F. Cathala, and P. Poncelet. Psp : Prefix tree for sequential patterns. In Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD’98) Nantes France LNAI, pages 176–184, 1998.
18. Adelson-Velskii and E. M. Landis, 1962. "An algorithm for the organization of information". *Proceedings of the USSR Academy of Sciences* 146: 263–266. (Russian) English translation by Myron J. Ricci in *Soviet Math. Doklady*, 3:1259–1263, 1962.
19. F. Masseglia, F. Cathala, and P. Poncelet. Psp : Prefix tree for sequential patterns. In Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD’98) Nantes France LNAI, pages 176–184, 1998.
20. G. Grahne and J. Zhu. Efficiently using prefix-trees in mining frequent itemsets. In *FIMI’03, Workshop onFrequent Itemset Mining Implementations*, November 2003.
21. C. Borgelt. Recursion Pruning for the Apriori Algorithm. *Proc. 2nd IEEE ICDM Workshop onFrequent Item Set Mining Implementations (FIMI 2003, Brighton, United Kingdom)*. CEUR Workshop Proceedings 126, Aachen, Germany 2004.http://www.ceur-ws.org/Vol-126/
22. Leung, C. K. S., Mateo, M. A. F., & Brajczuk, D. A. (2008). A tree-based approach for frequent pattern mining from uncertain data. Lecture Notes in Computer Science, 5012, 653–661.

Using Symbolic Functional Decomposition to Implement FSMs in Heterogenous FPGAs

Piotr Szotkowski, Mariusz Rawski, and Paweł Tomaszewicz

1 Introduction

Implementation of finite state machines in FPGA devices is traditionally a two-step process: first, the FSMs states are assigned binary values and its state and next-state variables are encoded to binary values; then the resulting binary function is mapped into the elements of the FPGA structure. [3] Previous research [1] shows that one of the best methods for implementation of binary functions in FPGAs is the method of functional decomposition. In case of the traditional two-step approach to implementation of FSMs, the effectiveness of binary next state and output function synthesis highly depends on the encoding chosen for the machines states. Unfortunately, the number of possible minimal-bit encodings grows factorially with the number of states; additionally, other previous research [6] shows that the method of functional decomposition yields the best results when states are encoded using some redundant bits – and, obviously, the number of sensible bit widths also grows with the number of states.

2 The Symbolic Functional Decomposition Method

To address the shortcomings of the two-step approach, the symbolic functional decomposition method for implementation of FSMs in FPGAs was proposed in [4, 5] and the algorithms for its implementation were proposed in [7, 8]. This method does not encode the FSM states to binary values before performing the decomposition of a binary function; instead it peforms a symbolic decomposition, maintaining the multi-value representation of the state and next-state variables, effectively encoding them partially on every decomposition step only to the extent that is required – and optimal – for the given step.

Results of implementation of benchmark FSMs, described in [6], show that thanks to adjusting the encoding to the decomposition process this method often yields significant improvements over the traditional, two-step approaches.

3 Applying the Symbolic Functional Decomposition Method to Heterogenous FPGAs

The abovementioned publications contain results tailored to homogenous FPGA structures – logic elements (look-up table cells, or LUT cells) able to directly implement binary functions of a given number of inputs and outputs, such as 5/2 (five-input, two-output) or 6/1 cells. The first step of the symbolic functional decomposition method is to select the U and V sets of inputs – and this selection directly depends on the LUT size of the target architecture (the number of inputs of the directly-implementable binary function).

However, modern FPGA structures consist of *slices*: logic elements able to implement a variety of binary function architectures. For example, a single slice of Virtex-5 can directly implement an 8/1 function, two 7/1 functions, four 6/1 functions, or four 5/2 functions. This means that all steps of the symbolic functional decomposition need to be adjusted to target heterogenous structures.

For example, the U and V partition step can no longer be optimised to target a single, optimal number of inputs for the G function; G functions of various numbers of inputs can fit efficiently on a heterogenous slice. Similarly, the β_{QU}, β_{QV} and β_G construction algorithms can target multiple input and output counts, with various combinations yielding good results (again, thanks to a single slice accomodating functions of various input/output counts).

P. Szotkowski (✉) • M. Rawski • P. Tomaszewicz
Institute of Telecommunications, Warsaw University of Technology, Warsaw, Poland
e-mail: p.szotkowski@tele.pw.edu.pl; rawski@tele.pw.edu.pl; p.tomaszewicz@tele.pw.edu.pl

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_116,

4 Experimental Results

As described in [6], the ideas and algorithms behind the symbolic functional decomposition method were implemented in an academic tool for logic synthesis – *art décomp*. This tool performs decomposition of finite state machines described in KISS format by iterative application of the symbolic functional decomposition method, and can now target heterogenous FPGA structures.

Table 1 compares the decomposition results of benchmark finite state machines into logical elements of the Virtex-5 FPGA structure. The first two columns contain the benchmark FSM name and the number of its states, while the rest contains the number of slices yielded by each approach and the number of bits used to encode the FSMs states.

The *[slices]* columns compare the number of Virtex-5 slices required to implement the given FSM after describing it in VHDL and enforcing different state encodings: the *one-hot* encoding (the *1-hot* column), encodings obtained using the *Jedi* and *Nova* state-encoding algorithms [2, 9], and a random encoding assignment (the *rand* column). Finally, the *a.d.* column contains the number of slices obtained by decomposing the given FSM using the *art décomp* software.

Table 1 FSM Synthesis for the Virtex-5 Structure

		[slices]					[bits/state]		
FSM	states	1-hot	Jedi	Nova	rand	a.d.	min	a.d.	1-hot
bbara	10	5	5	5	5	2	4	4	10
bbsse	16	9	9	10	9	8	4	6	16
bbtas	6	3	3	3	3	1	3	4	6
beecount	7	5	4	5	4	2	3	4	7
dk14	7	6	6	6	6	3	3	3	7
dk17	8	4	1	1	1	2	3	3	8
donfile	24	4	1	1	1	3	5	7	24
ex2	18	8	8	8	8	2	5	5	18
ex3	9	6	5	6	6	1	4	4	9
ex5	8	5	4	4	4	2	3	3	8
ex6	8	3	3	4	4	3	3	4	8
ex7	9	6	6	6	6	1	4	4	9
keyb	19	3	4	4	3	5	5	5	19
lion9	9	17	12	14	14	1	4	2	9
mark1	15	5	6	5	5	4	4	6	15
mc	4	6	6	6	6	2	2	3	4
opus	10	2	2	2	2	3	4	4	10
pma	24	6	5	5	5	6	5	9	24
s208	18	14	13	13	14	6	5	9	18
s27	6	6	6	6	6	1	3	3	6
s386	13	3	3	3	4	7	4	4	13
s420	18	9	10	9	9	6	5	9	18
s510	47	6	6	6	6	19	6	18	47
s8	5	32	32	31	32	1	3	3	5
sse	16	3	3	3	3	6	5	9	16
styr	30	9	9	10	9	22	5	9	30
tbk	32	27	27	25	25	20	5	6	32
tma	20	41	39	38	39	4	5	6	20
train11	11	11	10	10	10	1	4	3	11

The last three columns (grouped under *[bits/state]*) show how many bits were used to encode the FSM states. *Jedi*, *Nova* and random methods are minimal-width encodings – the number of bits is the upper bound of base-two logarithm from the number of states; their number of bits per state is in the *min* column. On the other end of the encoding size spectrum, the *one-hot* encoding uses as many bits as there are states (the *1-hot* column) for an s-state FSM all of the codes are s bits wide, consist of $s - 1$ zeroes and a single 1 (hence the *one-hot* name), with the 1 at a different position for every state.

The *a.d.* column shows the number of bits used to encode states of a given FSM by the algorithms used in *art décomp* – as can be seen, this number is usually closer to the minimal width than to what *one-hot* uses, but quite often larger than the minimum. Note how for the *lion9* and *train11* FSMs the number is actually smaller than what the minimal-width encodings use – this is because *art décomp* treats the state and next-state variables as symbolic and is thus able to perform an implicit state minimization process, encoding indistinguishable states using the same code.

5 Summary

While not always resulting in the smalles possible slice counts, Table 1 shows that in most tested cases the symbolic functional decomposition method yields better or comparable results to approaches encoding the states and then implementing the resulting binary function in the target FPGA structure.

Neither of the pre-encoding approaches is significantly better than the others, and they either use the minimum possible number of bits (*Jedi*, *Nova* and the random assignment) or as many bits as there are states (*one-hot*). The tested cases suggest that the best results are often obtained when the number of bits used for encoding is larger than the minium, but smaller than the number of states. Additionally, in the symbolic functional decomposition method – by obtaining the final encoding partially during the decomposition process – states that dont provide information relevant for the operation of the FSM are naturally encoded with the same codes, effectively leading to implicit minimization of the FSM. The results obtained so far (and presented in Table 1) confirm large potential of the symbolic functional decomposition method for efficient implementation of finite state machines in heterogenous FPGA structures.

References

1. J. A. Brzozowski and T. Łuba. Decomposition of Boolean functions specified by cubes. *Journal of Multiple-Valued Logic and Soft Computing*, 9(4):377–417, 2003.
2. B. Lin and A. R. Newton. Synthesis of multiple level logic from symbolic high-level description languages. *Proceedings of the IFIP TC 10/WG 10.5 International Conference on Very Large Scale Integration*, pages 187–196, 1989.
3. T. Łuba, G. Borowik, and A. Kraśniewski. Synthesis of finite state machines for implementation with programmable structures. *Electronics and Telecommunications Quarterly*, 55(2):183–200, 2009.
4. M. Rawski. The novel approach to FSM synthesis targeted FPGA architectures. *Proceedings of IFAC Workshop on Programmable Devices and Systems PDS 2004*, pages 169–174, 2004.
5. M. Rawski, H. Selvaraj, T. Łuba, and P. Szotkowski. Application of symbolic functional decomposition concept in FSM implementation targeting FPGA devices. *Sixth International Conference on Computational Intelligence and Multimedia Applications ICCIMA 2005*, pages 153–158, 2005.
6. P. Szotkowski. *Symbolic Functional Decomposition Method for Implementation of Finite State Machines in FPGA Devices*. PhD thesis, 2010.
7. P. Szotkowski and M. Rawski. Improvements to symbolic functional decomposition algorithms for FSM implementation in FPGA devices. *Electronics and Telecommunications Quarterly*, 55 (2):335–354, 2009.
8. P. Szotkowski, M. Rawski, and H. Selvaraj. A graph-based approach to symbolic functional decomposition of finite state machines. *Systems Science*, 35(2):41–47, 2009.
9. T. Villa and A. Sangiovanni-Vincentelli. NOVA: State assignment of finite state machines for optimal two-level logic implementation. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 9(9):905–924, 1990.

Efficient Functional Decomposition Algorithm Based on Indexed Partition Calculus

Mariusz Rawski, Paweł Tomaszewicz, and Piotr Szotkowski

1 Introduction

Functional decomposition is one of the best logic synthesis method targeting FPGAs. The most popular decomposition algorithms may be roughly divided into ones using *binary decision diagrams* (BDD) for Boolean function representation and ones using *cube* representation of the decomposed function. The advantage of the former is an efficient representation of functions with a large number of input variables. However, these algorithms face the problem of representing multiple-output and not fully specified functions [8, 9]. The latter algorithms allow simple representation of multiple-output and not fully specified functions. Most algorithms based on this representation use the blanket calculus for function manipulation [1]. The blanket calculus allows construction of very efficient decomposition algorithms [7, 10]. However the computational complexity of blanket manipulations makes it impossible to efficiently apply these algorithms to large Boolean functions.

In [5] a concept of *indexed partition* has been presented, allowing to manipulate Boolean functions described by cubes in similar way as blanket calculus but requiring much less computational and memory complexity. It combines the advantages of both blanket calculus and information relationships and measures [2].

2 Indexed Partition Calculus

Blankets, as well as information sets, are the way of expressing the compatibility relation COM on Boolean function's cubes. This also can be done using the concept of compatibility or *incompatibility graph*.

M. Rawski (✉) • P. Tomaszewicz • P. Szotkowski
Institute of Telecommunications, Warsaw University of Technology, Warsaw, Poland
e-mail: rawski@tele.pw.edu.pl

Operations on blankets have direct analogue in operations on incompatibility graphs. The concept of *indexed partitions* is used to model such operations on incompatibility graphs.

Definition 1. An *indexed partition is a set of ordered dichotomies* $\{B_i, P_i\}$ *on set S, such that* B_i *are disjoint subsets of S* and

$$\bigcup_i B_i = S. \tag{1}$$

In [1] a theorem based on the concept of cubes has been proposed which describes the existence of serial decomposition. This theorem can be re-expressed using the concept of indexed partitions.

Theorem 1. *Let* δ_V, δ_U, *and* δ_F *be indexed partitions induced on the function's F input cubes by the input subsets V and U, and outputs of F, respectively.*

If there exists an indexed partition δ_G *on the set of function F's input cubes such that* $\delta_V \leq \delta_G$, *and* $\delta_U \cdot \delta_G \leq \delta_F$, *then F has a serial decomposition with respect to (U,V).*

In practice, even large Boolean functions of many input variables are often specified by relatively small number of cubes. Since the size of indexed partition depends on the number of cubes, the test from Theorem 1 can be generally much computationally simpler than the test based on blankets [5].

3 Functional Decomposition Algorithm

The algorithm presented in this paper is based on the method of decomposition of function sets presented in [4]. However the blanket calculus has been substituted by indexed partition calculus.

The cube representation can easily describe a multi-output, not fully specified Boolean function using the

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_117,

espresso format. Unfortunately, it can be very inefficient when single outputs of the multi-output function depend on different input variables. The method proposed in [4] allows decomposing single-output Boolean function sets instead of one multi-output function.

The logic multi-output circuit of n input variables X can be represented as a set of Boolean functions $f_1(X_1), \ldots, f_m(X_m)$, where X_i is a sub-set of X. In the general case, each single function may depend on different subset of input variables and can be described by a different set of cubes. The application of serial functional decomposition based on blanket calculus is impossible in this case. The proposed modified serial functional decomposition algorithm allows decomposing function sets described by separate truth tables, even if each single function depends on different subset of input variables and truth tables consist of different sets of cubes.

When decomposing a set of Boolean functions $f_1, \ldots, f_m$ it is necessary to find such sub-function G that satisfies the decomposition condition for each function: $f_i = h_i(U_i, G(V))$.

The decomposition algorithm consists of the following steps:

1. input variable partitioning an appropriate input support V has to be selected for function G; in case functions $f_1, \ldots, f_m$ are specified by different truth tables, the V set has to be selected in such a way that it is a subset of input variables of each decomposed function,
2. the computation of δ_{V_i}, δ_{U_i} and δ_{f_i} for each function f_i - since truth tables of each function may consist different cubes, the indexed partitions δ_{V_i}, δ_{U_i} and δ_{f_i} for each function may be different,
3. the construction of δ_{G_i} in functional decomposition algorithm the construction of this indexed partitions corresponds to the construction of the multi-valued function G, since we want to find the decomposition with the same function G for all functions f_i, such δ_{G_i} should be created that satisfy the decomposition condition for all these functions and correspond to the same function G at the same time.

This decomposition method allows decomposing a set of Boolean functions described by separate truth tables. The decomposed function may depend on different input variables and can be described by different set of cubes. The application of the proposed decomposition algorithm allows for efficient creation of sub-functions common to all decomposed functions. This allows the designer to apply a much wider range of synthesis strategies.

Table 1 Results of Synthesis for 6-input LUTs

			Imfs+Lutpack		CombDec	
example	inputs	outputs	LUTs	levels	LUTs	levels
alu4	14	8	453	5	116	8
apex4	9	19	732	5	172	3
ex1010	10	10	1059	5	159	4
ex5	8	63	108	3	114	2
misex3	14	14	446	5	116	6
pdc	16	40	171	5	127	4
spla	16	46	263	4	153	5
$\sum$			3232	32	957	32

4 Results

For the purpose of evaluation of the efficiency of the proposed decomposition algorithm, a software tool *CombDec* has been created. The tool implements multilevel logic synthesis based on the concept of decomposition of function sets. To solve the input variable partitioning problem, the heuristic method presented in [6] has been applied.

The tool has been used to decompose selected benchmark examples into 6-input LUT networks. The resulting LUT networks were all verified using the combinational equivalence checker implemented in *ABC*.

Table 1 presents the comparison of the results obtained by the presented approach and the results reported in [3], obtained by technology mapping and resynthesis procedures (Imfs+Lutpack) implemented in the *ABC* tool. For both tools the number of 6-input LUTs in the resulting networks and the delay calculated as the depth of the networks have been provided.

It can be noticed that the method proposed in this paper produces networks of much greater area quality (calculated as the number of 6-LUTs) than those generated by the *ABC* procedures. The delay of networks obtained by both methods is similar.

5 Conclusions

There are many methods for decomposition of multi-output Boolean functions. These methods mostly use cube or BDD-based representation of the decomposed function. However, the common characteristics of these methods is the requirement that the decomposed multi-output Boolean function is represented by a single truth table in the case of cube representation or a single BDD.

The proposed decomposition method allows decomposing a set of Boolean functions described by separate truth tables. Application of indexed partition allows performing multi-level logic synthesis for multi-output, not fully specified Boolean functions of large number of input variables. The obtained results are of high quality in terms of the area calculated as the number of LUTs, as well as the delay of resulting network.

References

1. J.A. Brzozowski and T. Łuba. Decomposition of boolean functions specified by cubes. *Journal of Multiple-Valued Logic and Soft Computing*, 9:377–417, 2003.
2. L. Jóźwiak. Information relationships and measures: an analysis apparatus for efficient information system synthesis. In *EUROMICRO'97. New Frontiers of Information Technology., Proceedings of the 23rd EUROMICRO Conference*, pages 13–23, September 1997.
3. A. Mishchenko, R. Brayton, and S. Chatterjee. Boolean factoring and decomposition of logic networks. In *Computer-Aided Design, 2008. ICCAD 2008. IEEE/ACM International Conference on*, pages 38–44, November 2008.
4. M. Rawski. Decomposition of boolean function sets. *Electronics and Telecommunications Quarterly*, 53(3), 2007.
5. M. Rawski. Application of indexed partition calculus in logic synthesis of boolean functions for FPGAs. *International Journal of Electronics and Telecommunications*, 57(2), 2011.
6. M. Rawski. Heuristic algorithm of bound set selection in functional decomposition for heterogeneous FPGAs. In *Systems Engineering (ICSEng), 2011 21st International Conference on*, pages 465–466, 2011.
7. M. Rawski, T. Łuba, Z. Jachna, and P. Tomaszewicz. The influence of functional decomposition on modern digital design process. *Design of Embedded Control Systems*, pages 193–203, 2005.
8. T. Sasao, Y. Iguchi, and M. Matsuura. Comparison of decision diagrams for multiple-output logic functions. In *International Workshop on Logic and Synthesis*, pages 4–7, 2002.
9. T. Sasao and M. Matsuura. BDD representation for incompletely specified multiple-output logic functions and its applications to functional decomposition. In *IN PROC. DESIGN AUTOMATION CONF*, pages 373–378, 2005.
10. H. Selvaraj, P. Sapiecha, M. Rawski, and T. Łuba. Functional decomposition - the value and implication for both neural networks and digital designing. *International Journal of Computational Intelligence and Applications*, 6(1):123–138, 2006.

Rule Induction Based on Logic Synthesis Methods

Grzegorz Borowik, Andrzej Kraśniewski, and Tadeusz Łuba

1 Introduction

The effectiveness of the commonly used data mining methods is constrained by both unsolved theoretical problems and application related difficulties resulting from their computational complexity. In particular, the existence of NP-hard problems [15] and incomplete data make it difficult to efficiently process large databases. The existing data mining tools cannot perform the necessary calculations in a reasonable time, being also constrained by limited memory of available computers. These difficulties have been experienced by performing experiments with a number of data mining methods and systems on the publicly available data, such as those collected in the UC Irvine Machine Learning Repository (UCI database) [23].

The experiments with the state-of-the-art Dermatology database (UCI database that derives supportive medical diagnosis from 366 patients and uses 34 parameters (attributes)) demonstrate the inefficiency of the existing algorithms for data mining. One of the most widely known data mining systems, Rough Set Exploration System 2 (RSES2) [22], cannot even perform the attribute reduction (feature extraction) for this database. Another drawback of the algorithms that are used in the existing systems for attribute reduction, including RSES2, is the problem with the processing of incomplete data, which can easily be observed when performing experiments with another UCI database – Trains.

Our recent studies have shown that the above described problems with data mining result not only from deficiencies of the tools and computation procedures, but to a large extent are caused by immaturity and unsuitability of their underlying algorithms. One of the reasons for their low efficiency lies in the commonly used basic mathematical model for transformations of logical formulas [11], which exploits only the fundamental Boolean properties. With another model [3], so far used only in logic synthesis of digital circuits, the efficiency of medical data analysis can be significantly improved – the computations can be accelerated by a factor of several thousands. For example, the prototype implementation of the Boolean function complement algorithm, developed by the authors, in the process of attribute reduction computes all the reducts for the UCI Dermatology database in less than 4 minutes, whereas, as mentioned earlier, RSES2 cannot produce the result (the "insufficient memory" message is displayed).

We have found that similar problems occur for the rule induction – the procedure that plays a crucial role in knowledge discovery from databases; the commonly used rule induction algorithms are not effective. In the following section, we show how the rule induction can effectively be dealt with using algorithms developed for the synthesis of logic circuits.

2 Rule induction

Rule induction is one of the most important tasks in data mining. In data mining systems, the decision rules induced from the training data (objects and their known classification) are used to classify new objects, i.e. to assign each new object to an appropriate decision class. The classification is based on matching the description of that object with the decision rules (Fig. 1). Being more specific, the induced rules are used to determine whether or not the object satisfies the conditions defined by the subset of parameters belonging to a given decision class.

The objective of an efficient rule induction procedure is to generate a minimal set of rules which are as simple as possible (have as few parameters as possible). The problem of finding a minimal set of rules that covers a given set of example objects (in training data) and classifies them correctly is, however, NP-complete.

G. Borowik (✉) • A. Kraśniewski • T. Łuba
Warsaw University of Technology, Institute of Telecommunications, Warsaw, Poland
e-mail: G.Borowik@tele.pw.edu.pl; andrzej@tele.pw.edu.pl; luba@tele.pw.edu.pl

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_118,

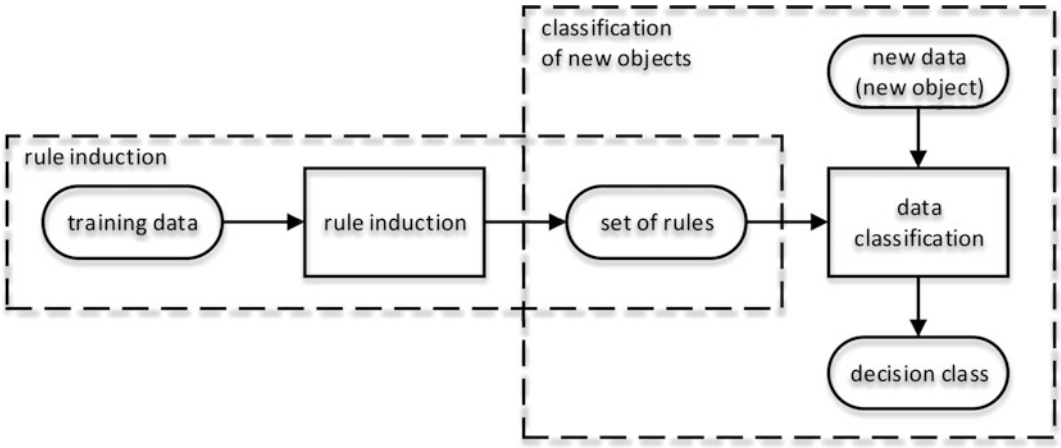


Fig. 1 Rule induction for data classification

In the literature, the problem of rule induction is often transformed to the problem of minimal coverage [2, 12, 17, 19, 20]. The best known methods for solving this problem are based on sequential covering. In each step of rule induction through sequential covering, a single object is considered, for which a decision rule is created and the other objects covered by that rule are removed. As a result, a set of rules is induced that covers the considered set of examples (set of training data). In [14], a more detailed scheme of this algorithm can be found. A somewhat different method is used in the LEM2 algorithm proposed by Grzymala-Busse [10], adopted in the earlier mentioned RSES2 system.

Our approach to the induction of decision rules is based on different ideas. We have proved that the problem of rule induction is equivalent to the problem of minimization of an appropriately constructed Boolean function. Therefore, to solve this problem, we can apply the algorithm for Boolean function expansion and Boolean function complement, being part of the Espresso logic synthesis system [4, 21].

The proposed procedure, executed for each decision class D_k, comprises the following steps:

1. Construction of the discernibility matrix for object u_i belonging to decision class D_k
2. Pre-induction of all rules corresponding to object u_i
3. Induction of all minimal rules corresponding to the objects of decision class D_k
   a. Finding the cover for decision class D_k
      To obtain a minimal set of rules – not necessarily exact (of the smallest cardinality) – representing decision class D_k, an appropriate cover has to be found. The cover is represented by a binary matrix M of n columns (n is the cardinality of the set of all minimal rules corresponding to the objects in D_k) and k rows (k is the number of objects in class D_k). When the resultant rule r_i covers object u_j, element (i, j) of the matrix equals 1, otherwise it equals 0.
   b. Finding the minimal set of rules for decision class D_k
      The minimal set of rules representing decision class D_k is determined by finding the minimal column cover of matrix M.

For this procedure, the problem of decision rule induction for a given decision class D_k is equivalent to the problem of minimization of Boolean function $f = (F, R)$, where the vectors of set F (ON-set) correspond to the objects in class D_k, and set R (OFF-set) is used to obtain the discernibility matrix [4]. Thus, the computational complexity of the rule induction can be estimated as being similar to the computational complexity of the Boolean function minimization. What determines the combinatorial explosion of this problem is, therefore, the computation of all minimal column covers of M. The complexity of this computation is determined by the rapidly growing (with an increase in the number of attributes) cardinality of the minimal rule family. Thus, the induction of decision rules for a large database – as proven by our preliminary experimental studies – must rely on heuristic algorithms. An efficient highly-heuristic Boolean function complement algorithm that has been successfully applied to solve the feature extraction problem [3] can be used for the calculation of a minimal rule for an individual object u_i.

The difference between the conventional approach to rule induction and our method is illustrated for the binary decision system in Table 1.

Using LEM2 [10], we obtain the following decision rules:

$$(a_1, 0)\ \&\ (a_2, 1)\ \&\ (a_3, 0)\ \&\ (a_4, 0) \rightarrow (d, 1),$$
$$(a_1, 1)\ \&\ (a_2, 1)\ \&\ (a_3, 0)\ \&\ (a_4, 0)\ \&\ (a_5, 0) \rightarrow (d, 1),$$
$$(a_1, 1)\ \&\ (a_2, 1)\ \&\ (a_3, 1) \rightarrow (d, 1),$$
$$(a_1, 0)\ \&\ (a_2, 1)\ \&\ (a_3, 0)\ \&\ (a_4, 1)\ \&\ (a_5, 0) \rightarrow (d, 1),$$
$$(a_1, 1)\ \&\ (a_2, 1)\ \&\ (a_3, 0)\ \&\ (a_4, 0)\ \&\ (a_5, 1) \rightarrow (d, 0),$$
$$(a_1, 0)\ \&\ (a_2, 1)\ \&\ (a_4, 1)\ \&\ (a_5, 1) \rightarrow (d, 0),$$
$$(a_1, 0)\ \&\ (a_2, 0)\ \&\ (a_4, 0) \rightarrow (d, 0),$$

with the corresponding logic expressions:

for decision class "equal to 1"
$d = \overline{a}_1 a_2 \overline{a}_3 \overline{a}_4 + a_1 a_2 \overline{a}_3 \overline{a}_4 \overline{a}_5 + a_1 a_2 a_3 + \overline{a}_1 a_2 \overline{a}_3 a_4 \overline{a}_5$,
for decision class "equal to 0"
$\overline{d} = a_1 a_2 \overline{a}_3 \overline{a}_4 a_5 + \overline{a}_1 a_2 a_4 a_5 + \overline{a}_1 \overline{a}_2 \overline{a}_4$.

Our method produces the following decision rules:

$(a_1, 1)$ & $(a_3, 1) \rightarrow (d, 1)$,
$(a_1, 0)$ & $(a_4, 0)$ & $(a_5, 1) \rightarrow (d, 1)$,
$(a_2, 1)$ & $(a_5, 0) \rightarrow (d, 1)$,
$(a_2, 0) \rightarrow (d, 0)$,
$(a_1, 1)$ & $(a_3, 0)$ & $(a_5, 1) \rightarrow (d, 0)$,
$(a_4, 1)$ & $(a_5, 1) \rightarrow (d, 0)$,

with the corresponding logic expressions:

for decision class "equal to 1"
$d = a_1 a_3 + \overline{a}_1 \overline{a}_4 a_5 + a_2 \overline{a}_5$,
for decision class "equal to 0"
$\overline{d} = \overline{a}_2 + a_1 \overline{a}_3 a_5 + a_4 a_5$.

This example illustrates the key difference between the proposed procedure and the conventional rule induction methods: the rules induced using the advanced logic synthesis procedures are simpler (more general). This means that we ultimately obtain better classification of data.

Table 1 Binary decision system

	a_1	a_2	a_3	a_4	a_5	d
1	1	1	0	0	0	1
2	0	1	0	0	0	1
3	1	1	1	0	1	1
4	0	1	0	0	1	1
5	0	1	0	1	0	1
6	0	1	1	1	1	0
7	0	0	0	0	0	0
8	1	1	0	0	1	0
9	0	1	0	1	1	0
10	0	0	1	0	0	0

3 Experimental study

The evaluation of the effectiveness of the proposed rule induction method relies on cross-validation. The idea is to randomly split the available database into two parts: the training database and the test database [19]. The training part is used to generate the set of rules using the proposed rule induction procedure and the test part is used to evaluate the rules.

In our experiments, the "quality" of the set of rules is evaluated using two measures:

- accuracy: the percentage of objects in the test database for which the rules produce the correct result (for which the object is assigned to the correct decision class),
- coverage: the percentage of all objects, included in the training database or test database, for which the rules produce some result (correct or incorrect decision class).

The outcomes of the experiments performed on large medical databases are shown in Table 2. As can be seen,

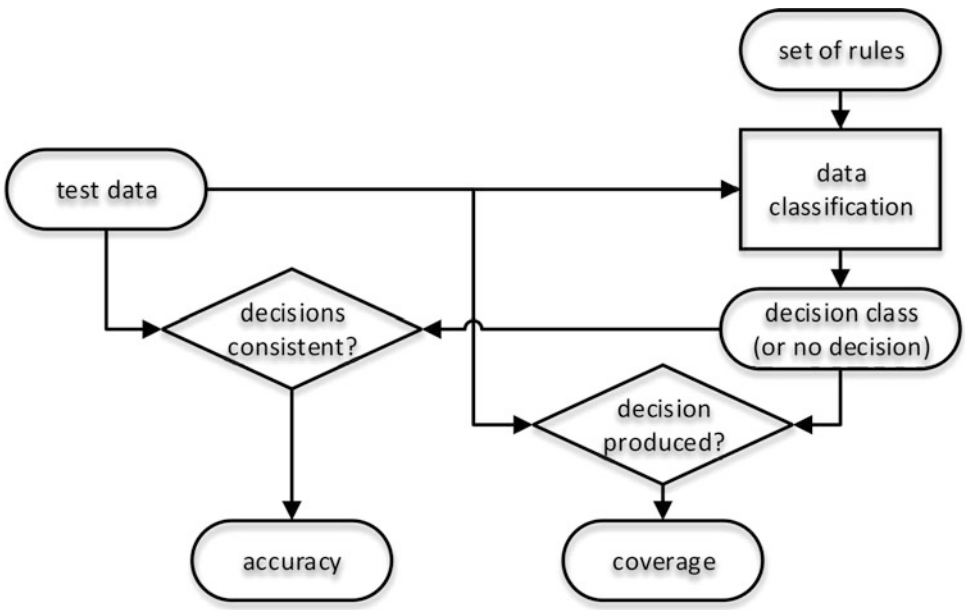


Fig. 2 Evaluation of the proposed rule induction method

Table 2 Experimental results

Algorithm	Database	Accuracy	Coverage
RSES2	Spambase	67.20%	93.10%
our method	Spambase	85.40%	95.10%
RSES2	House	84.90%	89.60%
our method	House	90.10%	100.00%
RSES2	Breast cancer	87.10%	87.10%
our method	Breast cancer	87.90%	93.20%
RSES2	Indian Liver Patient Dataset	28.84%	48.30%
our method	Indian Liver Patient Dataset	66.00%	97.00%
RSES2	Dermatology Data Set	87.77%	92.20%
our method	Dermatology Data Set	78.00%	99.80%
RSES2	**average**	**71.16%**	**82.06%**
our method	**average**	**81.48%**	**97.02%**

the proposed approach significantly improves the efficiency of the rule induction compared with the RSES2 system. The average rule accuracy is 10% higher and the rule coverage is 15% higher.

The presented results show the advantages of using logic synthesis methods for rule induction and clearly justify the need for more research in this field, so that to further exploit the opportunities discovered by our preliminary studies.

References

1. An, A., Cercone, N.: Rule Quality Measures for Rule Induction Systems: Description and Evaluation. Computational Intelligence 17(3), 409–424 (2001), DOI: 10.1111/0824-7935.00154
2. Andersen, T., Martinez, T.: Learning and generalization with bounded order rule sets. In: Proc. of 10th Int. Symp. on Computer and Information Sciences. pp. 419–426 (1995)
3. Borowik, G., Łuba, T.: Fast Algorithm of Attribute Reduction Based on the Complementation of Boolean Function. Klempous, R., Nikodem, J., Jacak, W., Chaczko, Z. (eds.) Advanced Methods and Applications in Computational Intelligence, Topics in Intelligent Engineering and Informatics, vol. 6, pp. 25–41. Springer International Publishing (2014), DOI: 10.1007/978-3-319-01436-4-2
4. Brayton, R.K., Hachtel, G.D., McMullen, C.T., Sangiovanni-Vincentelli, A.: Logic Minimization Algorithms for VLSI Synthesis. Kluwer Academic Publishers (1984)
5. Bruha, I.: Quality of decision rules: definitions and classification schemes for multiple rules. Nakhaeizadeh, G., Taylor, C. (eds.) Machine Learning and Statistics, pp. 107–131. Wiley and Sons (1997)
6. Clark, P., Boswell, R.: Rule induction with CN2: Some recent improvements. Kodratoff, Y. (ed.) Machine Learning — EWSL–91, Lecture Notes in Computer Science, vol. 482, pp. 151–163. Springer Berlin Heidelberg (1991), DOI: 10.1007/BFb0017011
7. Cohen, W.W.: Fast effective rule induction. In: Proc. of the Twelfth International Conference on Machine Learning. pp. 115–123. Morgan Kaufmann (1995)
8. Domingos, P.: Rule Induction and Instance-based Learning: A Unified Approach. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence – Volume 2. pp. 1226–1232. IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1995)
9. Džeroski, S., Lavrač, N.: Rule induction and instance-based learning applied in medical diagnosis. Technology and Health Care 4(2), 203–221 (1996)
10. Grzymala-Busse, J., Wang, A.: Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values. In: Proc. of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences (JCIS'97). pp. 69–72. Research Triangle Park, NC (Mar 1997)
11. Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: A tutorial (1999)
12. Maimon, O., Rokach, L.: Data Mining and Knowledge Discovery Handbook. Springer, 2 edn. (2010)
13. Mańkowski, M.: Decision rule generalization using complement of Boolean function (in Polish). B.Sc. dissertation, Warsaw University of Technology (2014)
14. Mitchell, T.: Machine Learning. Mac-Graw Hill, Boston (1997)
15. Papadimitriou, C.H.: Computational complexity. Academic Internet Publ. (2007)
16. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers (1991)
17. Skowron, A., Rauszer, C.: The Discernibility Matrices and Functions in Information Systems. Słowiński, R. (ed.) Intelligent Decision Support, Theory and Decision Library, vol. 11, pp. 331–362. Springer Netherlands (1992), DOI: 10.1007/978-94-015-7975-9-21
18. Smyth, P., Goodman, R.: Rule induction using information theory. Knowledge Discovery in Databases. AAAI/MIT Press (1991)
19. Stefanowski, J.: Rule induction algorithms for knowledge discovery (in Polish). Monograph Series, 361, Poznan University of Technology Publishing House, Poznan (2001)
20. Stefanowski, J., Vanderpooten, D.: A General Two-Stage Approach to Inducing Rules from Examples. Ziarko, W.P. (ed.) Rough Sets, Fuzzy Sets and Knowledge Discovery, pp. 317–325. Workshops in Computing, Springer London (1994), DOI: 10.1007/978-1-4471-3238-7-37
21. Espresso – multi-valued PLA minimization, http://embedded.eecs.berkeley.edu/pubs/downloads/espresso
22. RSES – Rough Set Exploration System, http://logic.mimuw.edu.pl/~rses/
23. UC Irvine Machine Learning Repository, http://archive.ics.uci.edu/ml/

Simpler Functions for Decompositions

Bernd Steinbach

1 Introduction

There are two basic approaches to find the structure of a combinatorial circuit for a given behavior. These are either covering methods or decomposition methods. One of the decomposition methods is the bi-decomposition. The basic idea of the bi-decomposition is that the given function is built by an AND-gate, an OR-gate, or by an XOR-gate of two inputs. If one of these gates splits the given function into two simpler functions, a complete circuit structure must be found after a certain number of decomposition steps.

However, two complicate functions, which control the inputs of one of these gates, can be merged by the gate into a simpler output function. In this case, the decomposition does find a circuit structure with a finite number of gates.

The bi-decomposition [1, 3] ensures the simplification in each decomposition step because it restricts to subfunctions of the decomposition which depend on less variables than the given function. The known bi-decomposition approach even detects existing subfunctions which depend on the smallest number of variables. Of course, such very simple subfunctions should be utilized in a decomposition, because they reduce the number of required decomposition gates and contribute to short path lengths.

There are Boolean functions for which no strong bi-decomposition exists. In such a case, a weak bi-decomposition finds one simpler subfunction and extends the other subfunctions to a lattice of functions. A complete circuit structure can be synthesized for each Boolean function only using the AND-, the OR-, and the XOR-bi-decomposition, and additionally the weak AND-, and the weak OR-bi-decomposition.

The drawback of this approach is that weak bi-decompositions are needed to reach the completeness and each bi-decomposition step adds one gate to the path of the decomposed function. An interesting question is whether the weak bi-decomposition can be substituted by a bi-decomposition into simpler subfunctions which depend on the same number of variables as the given function to decompose. This paper combines the knowledge of two other approaches to answer this question.

One of these approaches is generalization of lattices of Boolean functions. The independence of a function from a single variable can be detected by a simple derivative of the Boolean Differential Calculus [4]. The $2^{2^{n-1}}$ functions of $\mathbb{B}^n$ which are independent of a single variable x_i belong to the well-known lattice. In [5, 7] the existence of more general lattices are introduced in which a vectorial derivative overtake the role of the simple derivative. Functions of such lattices depend on all variables but are simpler than other functions of $\mathbb{B}^n$. It is possible to utilize these new found lattices for the bi-decomposition?

Another source to find simpler functions utilizes the Specialized Normal Form (SNF) [6]. The SNF is a unique ESOP representation of a Boolean function and the number of cubes in the SNF indicates the complexity of the function [8]. It arises the question about the relation of these two approaches and the possibilities to utilize such information for the bi-decomposition. The latest results of the research in this field are summarized in this paper.

To make the paper self-contained, Section 2 gives the needed definitions of derivatives, Section 3 introduces the basic principle of the SNF, and Section 4 very briefly explains the bi-decomposition approach. The results of the main analysis are explored in Section 5. This analysis studies the relations between the complexity provided by the SNF and dependencies of all functions of $\mathbb{B}^4$ regarding all directions of change. An example in Section 6 demonstrates achievable benefits of the suggested extended approach of the bi-decomposition before Section 7 concludes the paper.

B. Steinbach (✉)
Freiberg University of Mining and Technology, Institute of Computer Science, D-09596 Freiberg, Germany

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_119,

2 Simple and Vectorial Derivative

The simple derivative of a Boolean function $f(\mathbf{x})$ with regard to the variable x_i describes for which patterns of the remaining variables the change of the x_i-value causes the change of the function value.

Definition 1. *Let $f(\mathbf{x}) = f(x_1, \ldots, x_i, \ldots, x_n)$ be a Boolean function of n variables, then*

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = f(x_1, \ldots, x_i, \ldots, x_n) \oplus f(x_1, \ldots, \overline{x}_i, \ldots, x_n) \quad (1)$$

is the (simple) derivative of the Boolean function $f(\mathbf{x})$ with regard to the variable x_i.

The simple derivative $\frac{\partial f(\mathbf{x})}{\partial x_i}$ is again a Boolean function. If

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = 0 \quad (2)$$

then the function $f(\mathbf{x})$ is independent of the variable x_i. From Definition (1) follows the welcome property that the result function of the simple derivative $\frac{\partial f(\mathbf{x})}{\partial x_i}$ does not depend on the variable x_i anymore.

$$\frac{\partial}{\partial x_i}\left(\frac{\partial f(\mathbf{x})}{\partial x_i}\right) = 0 \quad (3)$$

holds for all Boolean functions $f(\mathbf{x})$ of $\mathbb{B}^n$.

The vectorial derivative of function $f(\mathbf{x})$ has a similar meaning like the simple derivative. The difference is that in the case of the vectorial derivative several variables change their values at the same point in time.

Definition 2. *Let $\mathbf{x}_0 = (x_1, x_2, \ldots, x_k)$, $\mathbf{x}_1 = (x_{k+1}, x_{k+2}, \ldots, x_n)$ be two disjoint sets of Boolean variables, and $f(\mathbf{x}_0, \mathbf{x}_1) = f(x_1, x_2, \ldots, x_n) = f(\mathbf{x})$ a Boolean function of n variables, then*

$$\frac{\partial f(\mathbf{x}_0, \mathbf{x}_1)}{\partial \mathbf{x}_0} = f(\mathbf{x}_0, \mathbf{x}_1) \oplus f(\overline{\mathbf{x}}_0, \mathbf{x}_1) \quad (4)$$

is the *vectorial derivative of the Boolean function $f(\mathbf{x}_0, \mathbf{x}_1)$ with regard to the variables of $\mathbf{x}_0$.*

The vectorial derivative $\frac{\partial f(\mathbf{x}_0, \mathbf{x}_1)}{\partial \mathbf{x}_0}$ is also a Boolean function, but differently to the simple derivative, a vectorial derivative depends in general on all variables $(\mathbf{x}_0, \mathbf{x}_1)$ like the given function $f(\mathbf{x}_0, \mathbf{x}_1)$. However, a vectorial derivative is also simpler than the given function, because:

$$\frac{\partial}{\partial \mathbf{x}_0}\left(\frac{\partial f(\mathbf{x}_0, \mathbf{x}_1)}{\partial \mathbf{x}_0}\right) = 0 \quad (5)$$

holds for all Boolean functions $f(\mathbf{x}_0, \mathbf{x}_1)$.

3 Specialized Normal Form - SNF

The Specialized Normal Form was found in a research for minimal Exclusive-OR Sum Of Products (ESOPs) in [6]. The number of cubes in the SFN allows us to distinguish several complexity classes of functions in $\mathbb{B}^n$. Further subclasses were detected in [8] using the Hamming distance δ between the cubes of an SNF.

The SNF utilizes the following algebraic property of the exclusive-or operation ($\oplus$) and the Boolean variable x:

$$x = \overline{x} \oplus 1 \quad (6)$$

$$\overline{x} = 1 \oplus x \quad (7)$$

$$1 = x \oplus \overline{x}. \quad (8)$$

These three formulas show that each element of the set $\{x, \overline{x}, 1\}$ has isomorphic properties. For each variable in the support of the Boolean function f, exactly one left-hand side element of (6), (7), or (8) is included in each cube of an ESOP of f. An application of these formulas from the left to the right doubles the number of cubes and is called expansion. The reverse application of these formulas from the right to the left halves the number of cubes and is called compaction.

The procedure to construct the SNF utilizes one more property the Boolean function f, a cube C, and the exclusive-or operation:

$$f = f \oplus 0 \quad (9)$$

$$0 = C \oplus C \quad (10)$$

$$f = f \oplus C \oplus C. \quad (11)$$

From these formulas follows that two identical cubes can be added to or removed from any ESOP without changing the represented function. The SNF can be defined using two simple algorithms based on the properties mentioned above.

The `expand()` function in line 3 of Algorithm 1 expands the cube C_j with regard to the variable V_i into the cubes C_{n1} and C_{n2} based on the fitting formula (6), ..., (8). Algorithm 1 realizes this expansion for all variables of all cubes of a given ESOP. Assuming n variables in the given ESOP, this Algorithm distributes the information about each given cube to 2^n cubes, similar to the creation of a hologram

Algorithm 1 Exp(*f*)

Input: any ESOP of a Boolean function f
Output: complete expansion of the Boolean function f with regard to all variables of its support

```
for all variables V_i of the support of f do
    for all cubes C_j of f do
        ⟨C_n1, C_n2⟩ ← expand(C_j, V_i)
        replace C_j by ⟨C_n1, C_n2⟩
    end for
end for
```

Algorithm 2 R(*f*)

Input: any ESOP of a Boolean function f containing n cubes
Output: reduced ESOP of f containing no cube more than once

```
for i ← 0 to n − 2 do
    for j ← i + 1 to n − 1 do
        if C_i = C_j then
            C_i ← C_{n−1}
            C_j ← C_{n−2}
            n ← n − 2
            j ← i
        end if
    end for
end for
```

of an object. Algorithm 2 removes all pairs of cubes using the formulas (9), (10), and (11) so that a unique ESOP of the Boolean function *f* remains.

Using Algorithms Exp(*f*) and R(*f*) it is possible to create a special ESOP having a number of remarkable properties which are specified and proven in [6].

Definition 3 (SNF(*f*)). *Take any ESOP of a Boolean function f. The ESOP resulting from*

$$SNF(f) = R(Exp(f)) \tag{12}$$

is called the Specialized Normal Form (SNF) of the Boolean function.

4 Bi-Decomposition

A bi-decomposition (see left part of Figure 1) decomposes a function $f(\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c)$ into two subfunctions $g(\mathbf{x}_a, \mathbf{x}_c)$ and $h(\mathbf{x}_b, \mathbf{x}_c)$. Both subfunctions are simpler than the given function $f(\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c)$ due to the missing variables $\mathbf{x}_b$ in the subfunction $g(\mathbf{x}_a, \mathbf{x}_c)$ and the missing variables $\mathbf{x}_a$ in the subfunction $h(\mathbf{x}_b, \mathbf{x}_c)$.

There are three types of by bi-decompositions shown in the left part of Figure 1. It is a property of the function $f(\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c)$ whether a bi-decomposition exists with regard to one of the decomposition-gates. An empty set of variables $\{\mathbf{x}_c\}$ and the split of the set of all variables $\{\mathbf{x}\}$ into the subsets $\{\mathbf{x}_a\}$ and $\{\mathbf{x}_b\}$ of the same size contributes best to the synthesis of a circuit by bi-decomposition. However, only few functions have this welcome property.

A necessary condition of the bi-decomposition [4] is that both the set of variables $\{\mathbf{x}_a\}$ and the set of variables $\{\mathbf{x}_b\}$ contains at least one variable. In the limit case of single variables in the sets $\{\mathbf{x}_a\}$ and $\{\mathbf{x}_b\}$, we can assume $x_i = \mathbf{x}_a$ and $x_j = \mathbf{x}_b$; nevertheless both subfunctions of the bi-decomposition are simpler than the given function $f(x_i, x_j, \mathbf{x}_c)$ because:

$$\frac{\partial g(x_i, \mathbf{x}_c)}{\partial x_j} = 0 \quad \text{and} \quad \frac{\partial h(x_j, \mathbf{x}_c)}{\partial x_i} = 0. \tag{13}$$

Unfortunately, there are functions for which no bi-decomposition exists. Le [2] additionally suggested for such cases the weak bi-decomposition. The function to decompose must hold a certain condition [4] for the weak AND-bi-decomposition (Figure 1 (d)) and the weak OR-bi-decomposition (Figure 1 (e)). Only the subfunction $h(\mathbf{x}_c)$ is simpler due to the missing variables $\mathbf{x}_a$. The function $g(\mathbf{x}_a, \mathbf{x}_c)$ of a weak bi-decomposition depends on the same

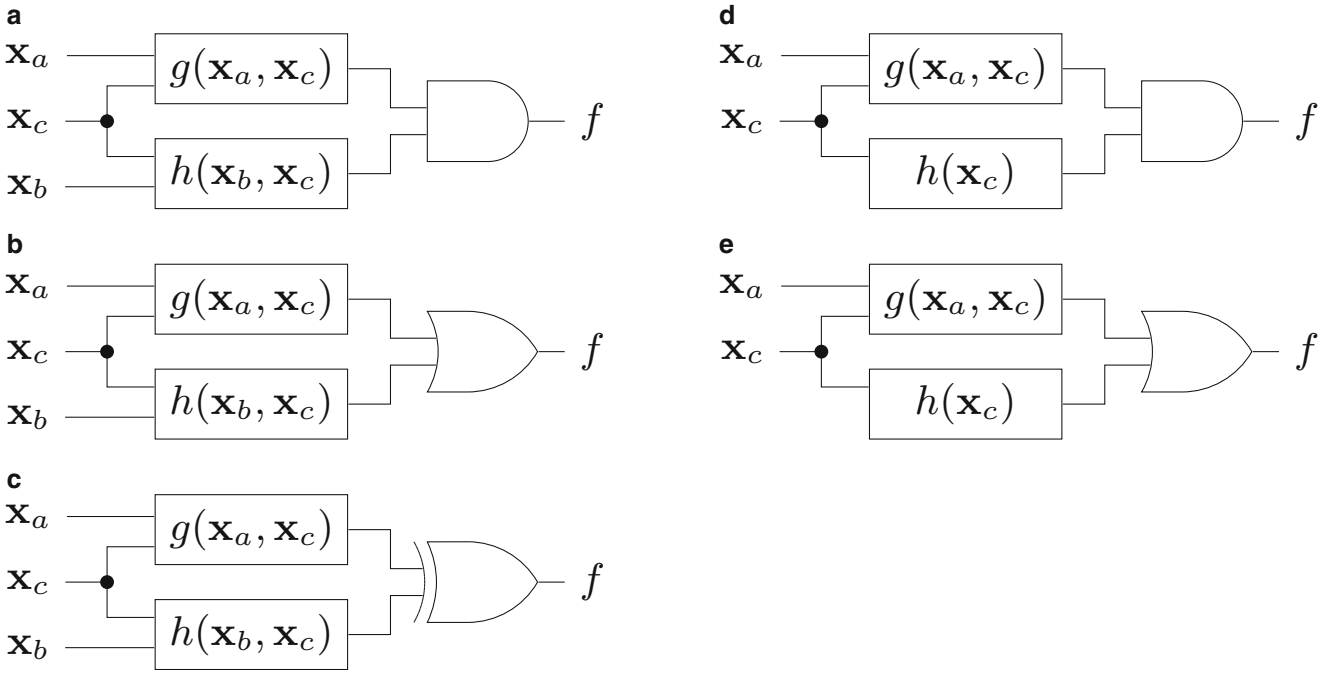


Fig. 1 Circuit structures of bi-decompositions: (a) AND-bi-decomposition; (b) OR-bi-decomposition; (c) XOR-bi-decomposition; (d) weak AND-bi-decomposition; (e) weak OR-bi-decomposition

variables as the given function $f(\mathbf{x}_a, \mathbf{x}_c)$, however, this function can be chosen from a larger lattice of functions.

A weak XOR-bi-decomposition can be realized for each pair of functions $f(\mathbf{x}_a, \mathbf{x}_c)$, $h(\mathbf{x}_c)$. However, the subfunction $g(\mathbf{x}_a, \mathbf{x}_c)$ can be more complicated than the given function $f(\mathbf{x}_a, \mathbf{x}_c)$. For that reason the weak XOR-bi-decomposition is excluded from the synthesis approach by bi-decomposition.

The recursive application of one of the three types of the bi-decomposition together with the weak OR-bi-decomposition and weak AND-bi-decomposition enables a complete multilevel design of each function. This completeness follows from Theorem 1 found by Le [2].

Theorem 1. *If the function* $f(\mathbf{x}_a, \mathbf{x}_c)$ *is neither weakly OR-bi-decomposable nor weakly AND-bi-decomposable with regard to a single variable* $x_i = \mathbf{x}_a$, *then the function* $f(x_i, \mathbf{x}_b)$ *is disjointly XOR-bi-decomposable with regard to the single variable* x_i *and the set of variables* $\{\mathbf{x}_b\} = \{\mathbf{x}_c\}$.
The proof of Theorem 1 is also given in [4].

5 Experimental Results

The key of the bi-decomposition is that each created subfunction is either independent of at least one variable x_i, which can be checked by (2); or the created subfunction belongs to a larger lattice than the given function $f(\mathbf{x})$. After several extensions, such a lattice contains a function $f(\mathbf{x})$ that also holds (2). The larger the number of variables from which all functions of the lattice are independent the simpler circuits can be synthesized by bi-decomposition.

In [7, 5] was shown, that

1. there are lattices of Boolean functions which do not depend on a certain number of variables;
2. the constant value 0 of the simple derivative (1) of such a function with regard to a respective variable indicates this property;
3. the constant value 0 of a vectorial derivative (1) also indicates a simpler function;
4. there are $2^n - 1$ directions of change in $\mathbb{B}^n$;
5. the number of independent directions of change for lattices in $\mathbb{B}^n$ is restricted to n.

From these findings arises the question whether vectorial derivatives can be utilized to find simpler subfunctions of a bi-decomposition. An alternative measure of the complexity of a Boolean function $f(\mathbf{x})$ is the number of cubes in the SNF ($f(\mathbf{x})$) [6, 8]. An experiment allows us to evaluate these properties form a more general point of view.

For that reason we calculated for all 65,536 Boolean functions $f(\mathbf{x})$ of $\mathbb{B}^4$ the number of cubes in the SNF($f(\mathbf{x})$) and all simple and vectorial derivatives. Table 5 summarizes these experimental results as follows:

- column 1 contains the numbers of functions of one complexity class of the SNF;
- column 2 lists the numbers of cubes of the SNF class as measure of the complexity;

Table 1 Evaluation of all functions of $\mathbb{B}^4$ regarding the SNF and vectorial derivatives

number of functions	number of cubes in the SNF	functions with $\frac{\partial f}{\partial \mathbf{x}} = 0$ and $\lvert\mathbf{x}\rvert = 1$	$\lvert\mathbf{x}\rvert = 2$	$\lvert\mathbf{x}\rvert = 3$	$\lvert\mathbf{x}\rvert = 4$	number
1	0	4	6	4	1	1
81	16	0	0	0	0	16
	16	1	0	0	0	32
	16	2	1	0	0	24
	16	3	3	1	0	8
	16	4	6	4	1	1
324	24	0	0	0	0	96
	24	1	0	0	0	96
	24	0	1	0	0	48
	24	1	1	1	0	48
	24	2	1	0	0	24
	24	2	2	2	1	12
1,296	28	0	0	0	0	832
	28	1	0	0	0	416
	28	0	0	1	0	32
	28	1	0	1	1	16
648	30	0	0	0	0	640
	30	0	0	0	1	8
648	32	0	0	0	0	320
	32	1	0	0	0	160
	32	0	1	0	0	96
	32	1	1	1	0	48
	32	0	3	0	0	16
	32	1	3	3	0	8
3,888	34	0	0	0	0	3,888
6,732	36	0	0	0	0	6,064
	36	1	0	0	0	32
	36	0	1	0	0	480
	36	0	0	1	0	128
	36	1	0	1	1	16
	36	0	2	0	1	12
7,776	38	0	0	0	0	7,776
9,234	40	0	0	0	0	8,704
	40	0	1	0	0	240
	40	0	0	1	0	192
	40	0	0	0	1	56
	40	0	1	2	0	24
	40	0	3	0	0	16
	40	0	6	0	1	2
14,472	42	0	0	0	0	14,416
	42	0	0	0	1	56
12,636	44	0	0	0	0	12,144
	44	0	0	1	0	288
	44	0	1	0	0	192
	44	0	2	0	1	12
5,184	46	0	0	0	0	5,136
1,944	48	0	0	0	0	1,776
	48	0	0	1	0	96
	48	0	1	0	0	48
	48	0	1	2	0	24
648	50	0	0	0	0	640
24	54	0	0	0	0	16
	54	0	0	0	1	8

- column 3 enumerates the numbers of simple derivatives which are equal to 0 for the evaluated functions (number of independent variables);
- in $\mathbb{B}^4$ there are six vectorial derivatives with regard to two variables; column 4 specifies how many of these vectorial derivatives are equal to 0;
- in $\mathbb{B}^4$ there are four vectorial derivatives with regard to three variables; column 5 specifies how many of these vectorial derivatives are equal to 0;
- in $\mathbb{B}^4$ there is one vectorial derivative with regard to all four variables; a value 1 in column 6 specifies that this vectorial derivative is equal to 0;
- the most right column 7 gives the numbers of functions with the properties introduced above.

As could be expected, functions of $\mathbb{B}^4$ which do not depend on all four variables have small values of $|\mathrm{SNF}(f(\mathbf{x}))|$. Column 3 of Table 5 shows the number of variables the evaluated functions do not depend on. The complete evaluation of all Boolean functions of $\mathbb{B}^4$ reveals that there are simple functions which depend on all four variables, but are independent of the common change of more than one variable; e.g., 48 functions with $|\mathrm{SNF}(f(\mathbf{x}))| = 24$ and one vectorial derivative with regard to two variables which is equal to 0, 32 functions with $|\mathrm{SNF}(f(\mathbf{x}))| = 28$ and one vectorial derivative with regard to three variables which is equal to 0, 8 functions with $|\mathrm{SNF}(f(\mathbf{x}))| = 30$ and one vectorial derivative with regard to all four variables which is equal to 0, and many more.

An interesting result is that there are also simple functions, which depend on all variables, and which also depend on all other directions of change. An example is the function $f(\mathbf{x}) = x_1\, x_2\, x_3\, x_4$ with $|\mathrm{SNF}(f(\mathbf{x}))| = 16$ for which neither any simple derivative nor any vectorial derivative is equal to 0.

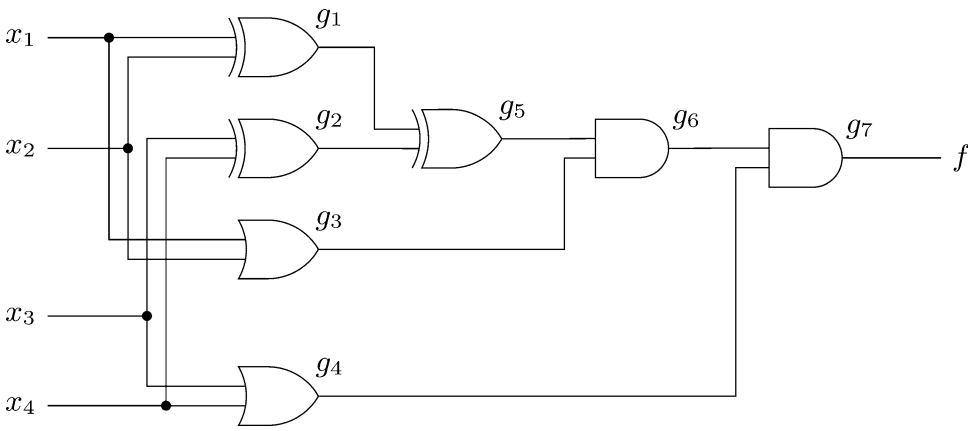


Fig. 2 Circuit structure of the function (14) designed by bi-decomposition controlled by the independence of variables.

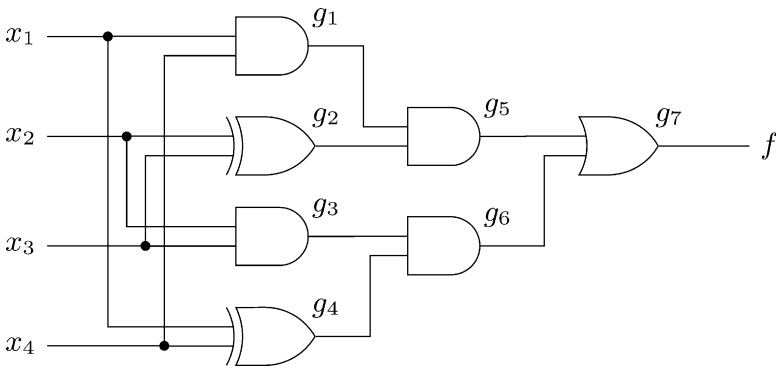


Fig. 3 Circuit structure of the function (14) designed by bi-decomposition controlled by $|\text{SNF}(f)|$.

6 Example of a Bi-Decomposition Controlled by $|\text{SNF}(f)|$

A simple example shows the advantage of the bi-decomposition controlled by $|\text{SNF}(f)|$. The function (14) is a symmetric function. It is known that there is no classical bi-decomposition for this function.

$$f = x_1\, x_2\, x_3\, \overline{x}_4 \vee x_1\, x_2\, \overline{x}_3\, x_4 \vee x_1\, \overline{x}_2\, x_3\, x_4 \vee \overline{x}_1\, x_2\, x_3\, x_4 \tag{14}$$

The classical bi-decomposition approach requires two weak AND-bi-decompositions (realized by the AND-gates g_6 and g_7 of Figure 2) before an XOR-bi-decomposition (realized by the XOR-gate g_5 of Figure 2) can be applied. Due to the weak AND-bi-decompositions there is no balanced path length. The shortest path contain only two gates (g_4 and g_7) and the longest path contains even four gates (g_1, g_5, g_6, and g_7). It should be mentioned that $|\text{SNF}(f)| = 40$ and the first weak AND-bi-decomposition increases the complexity: $|\text{SNF}(g_6)| = 44$, but utilizes the independence: $\frac{\partial g_6(\mathbf{x})}{\partial(x_3, x_4)=0}$.

The new bi-decomposition controlled by $|\text{SNF}(f)|$ decomposes the same function $f(\mathbf{x})$ (14) with $|\text{SNF}(f)| = 40$ into two subfunctions g_5 and g_6 (see Figure 3). Each of these two subfunctions depends on all four variables. Hence, the condition of the classical bi-decomposition is not achieved. However, these two subfunctions are simpler measured by the number of cubes in the SNF:

$$|\text{SNF}(g_5)| = |\text{SNF}(g_6)| = 24 \tag{15}$$

and one vectorial derivative of theses functions with regard to two variables is equal to 0. The number of gates remains 7, but the benefit of the circuit structure of Figure 3 is that all paths contain the same number of only two gates. Hence, the proof of concept of the bi-decomposition controlled by $|\text{SNF}(f)|$ is achieved.

7 Conclusions

The experimental results in Table 5 confirm that not only zero-functions of simple derivatives but also zero-functions of vectorial derivatives indicate simple functions. The number of cubes in the SNF is an additional indicator for a more

general bi-decomposition approach. The comparison of the circuit structures of the same function (14) using the classical bi-decomposition in Figure 2 and the new extended bi-decomposition in Figure 3 confirm that the proof of concept of the bi-decomposition controlled by $|\mathrm{SNF}(f)|$ is achieved.

References

1. Bochmann, D., Dresig, F., and Steinbach, B.: *A New Decomposition Method for Multilevel Circuit Design*. European Design Automation Conference, Amsterdam, The Netherlands, 1991, pp. 374–377.
2. Le, T.Q. Testbarkeit kombinatorischer Schaltungen—Theorie und Entwurf (in German). Dissertation thesis, Technical University Karl-Marx-Stadt, Karl-Marx-Stadt, Germany, 1989.
3. Mishchenko, A., Steinbach, B., and Perkowski, M.: *An Algorithm for Bi-Decomposition of Logic Functions*. in: Proceedings of the 38th Design Automation Conference 2001. June 18-22, 2001, Las Vegas (Nevada), USA, 2001, pp. 103–108.
4. Posthoff, Ch. and Steinbach, B.: Logic Functions and Equations - Binary Models for Computer Science. Springer, Dordrecht, The Netherlands, 2004.
5. Steinbach, B.: *Generalized Lattices of Boolean Functions Utilized for Derivative Operations*. in: Materiały konferencyjne KNWS'13, Łagów, Poland, 2013, pp. 1–17.
6. Steinbach, B. and Mishchenko, A.: *SNF: A Special Normal Form for ESOPs*. in: Proceedings of the 5th International Workshop on Application of the Reed-Muller Expansion in Circuit Design (RM 2001), August 10–11, 2001, Mississippi State University, Starkville (Mississippi), USA, 2001, pp. 66–81.
7. Steinbach, B. and Posthoff, Ch.: *Derivative Operations for Lattices of Boolean Functions*. in: Proceedings Reed-Muller Workshop 2013, Toyama, Japan, 2013, pp. 110–119.
8. Steinbach, B. and De Vos, A.: The *Shape of the SNF as a Source of Information*. in: Steinbach, B. (Ed.): Boolean Problems, Proceedings of the 8th International Workshops on Boolean Problems, September 18–19, 2008, Freiberg University of Mining and Technology, Freiberg, Germany, 2008, pp. 127–136.

Node Demand Reverse Deduction (DRD) Technology for Water Supply Networks

Ronghe Wang, Zhixun Wang, Junhui Ping, Jilong Sun, and Chaohong Xiao

1 Introduction

Historically, demand driven techniques have made significant contributions to scientific progress, engineering design, and the operation and management of the water distribution networks. However, in the actual operation and management of water supply networks, it is unrealistic to constantly measure the water usage for every node or every user; therefore it is impossible to get the accurate flow volume through the nodes. With the demand driven analytical techniques, There are significant limitations. Worldwide, people are still using manual operation, resulting in a huge waste of energy and water leakage. Even more serious is that people do not know where the leakages are happening. Therefore, leakage inspection and detection are forever a burden for the water companies and the only way to save energy and water resources.

Wu (2009) proposed a pressure-driven hydraulic calculation method; however it still employs Germanopoulos' equations of pressure fitting methodology. Ferrante (2013) took into consideration the effect of consumption in his pipeline leakage control equation, which greatly improved the accuracy of calculations based on flow. Walski (2012) based on node flow analysis proposed a new concept that flow is based on water pressure, volume, controllable pressure, uncontrollable pressure among other factors. Even though the above methods made improvement over the constant flow assumption, the calculation method is still flow driven.

In our mission to improve leak detection techniques, we have exhausted almost all conceivable methods. Kang (2012) through sensitivity analysis, using data gathered from monitoring systems, proposed water distribution network sensitivity analysis to perform pipe burst and leakage analysis. Martini (2012), based on a series of tests, proposed a way to identify pipe burst through pipe vibration monitoring. Wang (2014b) proposed a method based on temperature changes on the ground with infrared detection technology.

This paper, drawing inspiration from Nobre's (2011) method to extract manhole ground elevations in the drainage system, and Schwendel's (2012) evaluation of interpolation methods for obtaining terrain elevations, proposed a node demand reverse deduction (DRD) technology for water supply networks. Using the demand and hydraulic grades (HG) at the nodes, the leakage volume can be calculated and leakage location can be identified, Wang (2014a).

2 Water Supply Network Node Demand Reverse Deduction (DRD) Techniques

With the advancement in computations, communications, and intelligent sensing technology, the costs have drop significantly on pressure monitoring equipment, data gathering and transmission. Thusly, wide layout pressure monitoring network for water distribution systems is going to be very feasible. The data collected from pressure monitoring can be used to greatly improve the accuracy of water supply scheduling and management, and accurately determine the amount and location of leakage.

2.1 Monitored Water Pressure Data Processing

Due to the dynamic nature of flow through a distribution network, changes in water pressure and the concentration of dissolved gas in water, there will be momentary fluctuations in the pressure data. Therefore, it is critical to obtain a stable

R. Wang (✉) • J. Ping • J. Sun • C. Xiao
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
e-mail: Wang.ronghe@sz.tsinghua.edu.cn

Z. Wang
ESRI, 380 New York Street, Redlands, CA 92373, USA
e-mail: Andw90@hotmail.com

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_120,

pressure data. Through experiments and analysis (Lu, 2010), water pressure will be sampled 5-10 times every minute, and the moving average value is obtained in 2-5 minutes. If there is a sudden change in pressure for the moving average value, the data will be sent immediately, otherwise data will be transmitted in 15 minute intervals. This method allows for timely adjustments of the pump schedule and quick identification of pipe burst and leakage locations.

2.2 Node Hydraulic Grade Calculation

Based on the principles of digital terrain model (DTM), the use of finite discrete monitoring point (x,y) coordinates, and monitored water pressure data, the HG at each node can be calculated through a triangulated irregular network (TIN) and Lawson interpolation methods. The basic approach is to construct a large triangle or polygon with boundary pressure sensors. All the nodes in the network should be covered by the triangle or polygon. For the HG to be calculated node, it is necessary to connect the node with the three vertices of the triangle which surrounds the node to form three new triangles, then using Lawson's local optimization criterion to optimize the triangle and to calculate the HG of the node.

2.3 Monitor Point Layout Methodology

Based on the principle of the node HG calculation, water pressure monitoring points are arranged following the principles listed below:

- Network boundary points: The polygon formed by all the boundary monitor points should cover all nodes where the pressure needs to be calculated.
- Water source points: Pump station discharge points, water tanks and all other water control points
- Junctions on trunk pipeline: All important nodes connected to the main pipeline.
- Points where the interpolation calculation and the steady state model simulation calculation produce drastically different HG results: First, use the traditional pipe network model to perform steady-state calculation to get the HG of each node. Then use the water pressure sensor positioning method mentioned above to perform triangular irregular network HG interpolation calculation to solve for the HG at each node. The node with the biggest difference in HG shall be marked as a monitor point. Perform the interpolation calculation again until the HG calculated at each node by the two methods meet the accuracy requirement.

2.4 Pipe friction coefficient calculations

The pipe friction coefficient plays an important role in accurately assessing the flow rate. It is cost prohibitive to install flow meter sensor at each pipe; therefore the best way is to categorize pipes by similar friction coefficient. The pipes can be classified by pipe materials, service years, diameters and locations. Then representative pipes can be chosen to be monitored. The number of sensors needed depends on the size of the network. There are two methods for calculating the pipe friction coefficients, namely real-time monitoring and periodic measurements.

- Real-time monitoring: For critical pipelines, the same measurement methodology for pressure as discussed in section 2.1 is adapted for flow measurement.
- Regular Periodic Measurement: It is both feasible and necessary to take regular measurements of pipe friction to improve network operation and management.

Using the Hazen-William head loss formula, the pipe friction coefficient is calculated as following:

$$H_i - H_j = \frac{10.67 L_{ij} q_{ij}^{1.852}}{C_{ij}^{1.852} D_{ij}^{4.87}} \tag{1}$$

Therefore the Haze-William constant C is:

$$C_{ij} = 3.59 \left(\frac{L_{ij}}{H_i - H_j} \right)^{0.54} D_{ij}^{-2.63} q_{ij} \tag{2}$$

Where H is the HG at a node (m); q is the pipe flow rate (m3/s); L is the pipe length between two nodes (m), D is the diameter of the pipe (m); subscript i and j indicates the start and stop of the pipe nodes.

2.5 Model Processing Technique

Pipe network model for node HG calculation based on pressure sensors is not limited by the size of the water distribution system. As long as the polygon formed by the boundary sensor points covers the node required, the HG can be calculated. However, the accuracy does suffer if there are insufficient number of sensors, to improve the accuracy of the calculation while minimizing the number of sensors, the following approach can be taken:

- Trim branch pipes: all the branch pipes can be trimmed.
- Remove pipes in the leaf-like districts: district supplied by only one main pipeline can be removed.

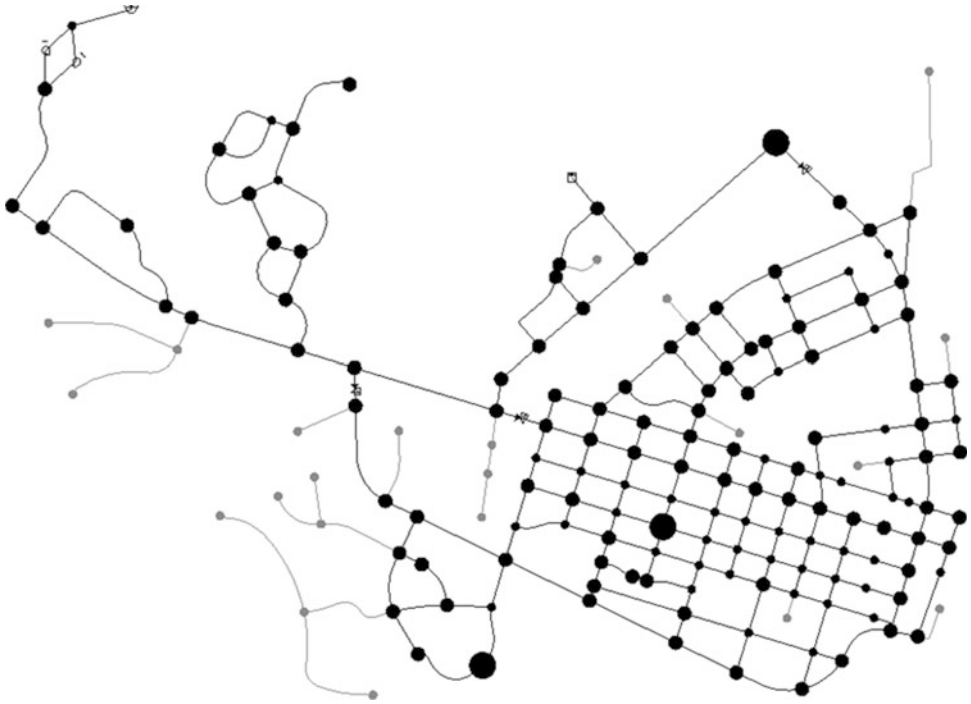


Fig. 1 Sample case model

2.6 Model calculation accuracy

The accuracy of HG has a large impact on the flow rate, especially in the case of small headloss in a short pipe segment. Because the flow rate unit is m^3/s in the calculation model, the flow accuracy is already 0.001 for the measurement unit of 1 L/s. Thusly, the recommended calculation accuracy of the HG in the model is 0.3 m, and the flow rate is 1 L/s. This is a tenth of the conventional model calibration accuracy of 3 m. Therefore, the accuracy of this calculation method is very high.

2.7 Node demand calculation

Using calculated node HG and dynamic or static head loss formula, the flow rate of each pipe can be solved by:

$$H_i - H_j = S_{ij} q_{ij}^n \tag{3}$$

$$q_{ij} = \left(\frac{H_i - H_j}{S_{ij}}\right)^{1/n} \tag{4}$$

According to the principle of mass balance, assume the flow towards the node as positive and away from the node as negative, the node demand can be calculated by summation of all flows in and out of the node:

$$Q_i = \sum_{j=1}^{m} q_{ij} \tag{5}$$

In the above equations, S is the pipe roughness; Q is the calculated node demand (m3/s); n is the head loss exponent; i and j indicates the node index, and m is the number of pipes connected to node i.

3 Examples and Applications

3.1 Sample Case

Fig. 1 presents a water distribution network composed of 250 pipes, 166 nodes, 1 water source, 2 pumps and 1 water tower. The pipe diameters range from 50 mm to 600 mm, total pipe length of 46.23 km, water supply capacity of 10000 m^3/day, and average flow rate of 85.37 L/s. The demand pattern is divided into 10 categories, including Class I, II, and III residential areas, government, business, general industry, chemical industry, military, school and agriculture.

Firstly, the system was run with steady state and 24-hour Extended Period Simulation (EPS) state. The simulation results were used as SCADA sensor data. The current pipe friction coefficients were used as is. Then the system was run with the following three scenarios by using the DRD calculation method:

- Pressure was monitored at all nodes of the entire network.
- Pressure was monitored at all nodes of the simplified network.
- Pressure is monitored at the minimum number of nodes while still meeting the accuracy criteria.

Table 1 SCADA monitoring data sample (Partial data)

	Pressure Sensors (m)				Flow Sensors (m^3/s)		
Date and Time	T-1	J-120	J-119	J-118	P-1	P-46	P-45
2/6/14 12:00 am	176.78	171.46	171.46	171.48	0.08	0.04	0.04
2/6/14 1:00 am	177.05	174.56	174.56	174.58	0.08	0.03	0.03
2/6/14 2:00 am	177.31	174.82	174.82	174.84	0.08	0.03	0.03

Table 2 pressure monitoring at all nodes of the entire network

Node	Measured HG (m)	Interpolated HG (m)	Calculated Node Demand (L/s)	Recalculated HG (m)	Interpolated HG – Measured HG (m)	Recalculated HG – Measured HG (m)
J-108	166.77	166.77	10.23	166.78	0.00	0.01
J-106	171.79	171.79	0.47	171.80	0.00	0.01
J-99	172.61	172.61	0.38	172.62	0.00	0.01
J-118	171.48	171.48	0.18	171.49	0.00	0.01

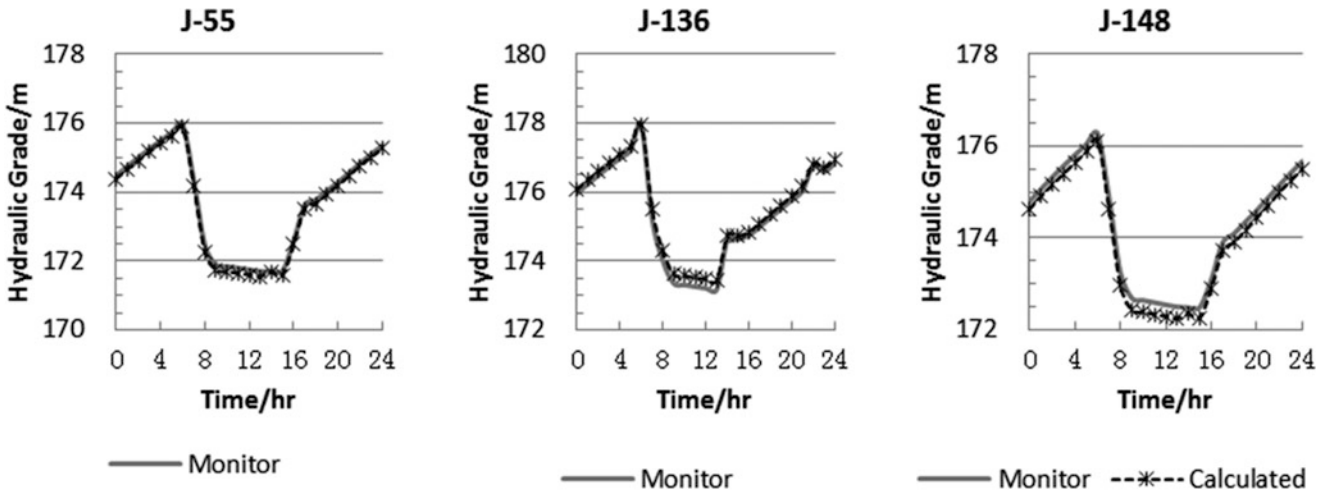


Fig. 2 Measured HG and Recalculated HG from 24 Hour EPS Analysis

In Fig. 1, the gray lines represent pipelines that can be removed during the simplification process, the medium and the biggest points represent the sensor points for satisfying the accuracy requirement, the three large points represent nodes with 24hour EPS run, and the biggest top point indicates the leak node. SCADA signal formatting is presented in Table 1.

3.2 Calculation Results and Analysis

Pressure is monitored at all nodes of the entire network. All the node pressures and the pipe flows of the pipes connected to the water sources are monitored. The pipe friction coefficient are using the original values for the water distribution system. As shown in Table 2, column "measured HG" represents the monitored HG by the SCADA system for all the nodes; column "Interpolated HG" represents the calculated HG through triangulated irregular network (TIN) and Lawson interpolation methods based on the measured HG; column "Calculated Node Demand" represents the demand at each node calculated based DRD theory. With the new calculated node demands, the water distribution system was simulated. Column "Recalculated HG" represents the simulation HG results for all the nodes. For ease of analysis, the difference between the measured HG and the interpolated HG, and the difference between measured HG and the recalculated HG are shown as column "Interpolated HG – Measured HG" and column "Recalculated HG – Measured HG". The table was sorted by "Recalculated HG – Measured HG" from large to small. As the table demonstrates, the largest difference error is only 0.01 m.

Simplified network with Pressure Monitoring at every node. By removing branch pipes, the network can be simplified without disturbing the state of the network. The simplified pipe network model is shown in Figure 2, gray pipes and nodes are removed from the network. After

Table 3 Pressure monitoring at all nodes of the simplified network

Node	Measured HG (m)	Interpolated HG (m)	Calculated Node Demand (L/s)	Recalculated HG (m)	Interpolated HG – Measured HG (m)	Recalculated HG – Measured HG (m)
J-106	171.79	171.79	0.48	171.80	0.00	0.01
J-99	172.61	172.61	0.38	172.62	0.00	0.01
J-118	171.48	171.48	0.19	171.49	0.00	0.01

Table 4 Optimized pressure sensor placement on simplified network

Node	Measured HG (m)	Interpolated HG (m)	Calculated Node Demand (L/s)	Recalculated HG (m)	Interpolated HG – Measured HG (m)	Recalculated HG – Measured HG (m)
J-148	173.51	173.51	0.56	173.25	0.00	-0.26
J-52	173.77	173.77	1.51	173.51	0.00	-0.26
			...			
J-3	177.15	177.15	0.19	177.15	0.00	0.00
			...			
J-135	174.21	174.21	0.24	174.47	0.00	0.26
J-136	174.25	174.25	7.11	174.51	0.00	0.26

simplification, the pressure at each node is monitored, and the node demand and HG are calculated. The results are shown in table 3, and the largest difference error is 0.1 m.

3.3 Simplified network with Pressure Monitoring by Optimized Pressure Sensor Placement

In order to reduce the number of pressure sensors and still maintain an accuracy level of 0.3 m for nodal HG, the pressure sensor optimization method from section 2.3 is applied. Figure 2 presents the optimized pressure sensors layout with medium and large points. After optimization, the number of sensors is reduced to 97, a reduction of 69 sensors or a 41 % reduction in the number of sensors.

Table 4 shows the results of node demand and HG. As expected, even with a reduction in the number of sensors, the methodology still maintains a high level of accuracy.

3.4 24hour EPS run based on DRD

The optimized monitoring points identified by use of steady state conditions, as verified by experiment, are also suitable for all other changed hydraulic states. Using the monitored data, a 24 hour EPS analysis was performed based on node demand reverse deduction (DRD) technology and the result was compared with the measured HG. Figure 3 shows the two points (J-148 and J-136) that had the largest positive and negative difference between recalculated and measured HG and a point in the middle of the network that had the largest difference (J-55).

3.5 Determining the leakage rate and leakage location based on DRD

DRD technology can also be used to find the leakage rate and location for the water distribution systems. Using currently existing technologies, leakage rate detection has the accuracy of about 240 m^3/day (10 m^3/hr). This paper also uses this precision. Taking into consideration the error of calculation to be 0.001 L/s (86.4 m^3/day), it is determined the daily leakage amount to be 150 m^3/day (~240-86.4).

There are two methods for leakage calculation:

- Store the calculated nodal demand for the recent 48 hours. Calculate the water volume of every node for every 24 hrs. If the water volume between the first 24 hrs and the second 24 hrs is more than 150 m^3/day, then there is a possible leak at the currently node location.
- Calculate the billing period water volume for each node and compare it with the meter reading. If the difference is greater than 150* billing days, then there is a possible leak at the current location.

The experiment performed for this paper created a leak at node J-136 with a leakage rate of 10 m^3/hr (2.78 L/s). This node is shown in Figure 2 as the big point at the very top of the network. Steady state simulation results are shown in Table 5. Since the calculated water usage is much higher than standard water usage, and the difference is bigger

Table 5 J-136 Leakage Calculation.

Items	J-136
Model Actual Node Demand (L/s)	7.68
Water Leakage Rate (L/s)	2.78
Calculated Normal Water Usage (L/s)	7.10
Calculated Water Usage with Leak (L/s)	9.70
Calculated Normal Daily Water Usage (m^3/day)	613.44
Calculated Daily Water Usage with Leak (m^3/day)	838.25
Leak rate (m^3/day)	224.81
Leak or not?	Yes

than 150 m^3/day, there could be a possible water leakage at the node.

4 Conclusion

The node demand reverse deduction (DRD) technology for water supply network can be used to find the actual node demand of every node in the network. Using the calculated node demand, a realistic model of the water network can be built to calculate leakage rate, locate the leakage position, get information about the water distributed, get the usage patterns of every user, schedule, operate and management.

As demonstrated by the analysis presented in this paper, the establishment of a complete network of pressure sensors at every node can accurately calculate the demand at each node. Additionally, by simplifying the water distribution network by cutting off branch pipes will reduce the number of sensors while still preserving the accuracy of the water model. Furthermore, through the use of the pressure sensor placement strategy, the number of pressure sensors can further be reduced while still maintaining a high level of accuracy. Using a verified model as a basis and the calculated node demand to perform a 24 hr extended period simulation analysis, this paper suggests an absolute HG accuracy of 0.3 m between the calculated, flow rate accuracy of 1 L/s.

The inverse node water HG and water demand calculation methodology based on real time pressure sensor data proposed in this paper is drastically different from the traditional demand driven calculation technology. This proposed solution brings excitingly new applications of SCADA networking monitoring to the water supply industry.

References

1. Ferrante, M., Massari, C., Todini, E., Brunone, B., & Meniconi, S. (2013). Experimental investigation of leak hydraulics. Journal of Hydroinformatics, 15(3).
2. Kang, D., & Lansey, K. (2012). Novel Approach to Detecting Pipe Bursts in Water Distribution Networks. Journal of Water Resources Planning and Management, 140(1), 121-127.
3. Lu, S., Liu, Z., Lai, Y., et al. (2010). Real time Manometry For Burst Leakage Monitoring Along Large scale And Long Distance Pipelines, China Water & Wastewater, 26(6), 58-62
4. Martini, A., Troncossi, M., Rivola, A., & Nascetti, D. (2014). Preliminary investigations on automatic detection of leaks in water distribution networks by means of vibration monitoring. Advances in Condition Monitoring of Machinery in Non-Stationary Operations, 535-544.
5. Nobre, A. D., Cuartas, L. A., Hodnett, M., Rennó, C. D., Rodrigues, G., Silveira, A., ... & Saleska, S. (2011). Height Above the Nearest Drainage–a hydrologically relevant new terrain model. Journal of Hydrology, 404(1), 13-29.
6. Schwendel, A. C., Fuller, I. C., & Death, R. G. (2012). Assessing DEM interpolation methods for effective representation of upland stream morphology for rapid appraisal of bed stability. River Research and Applications, 28(5), 567-584.
7. Walski, T., & Giustolisi, O. (2012). An overview of water demand: Volume vs. Pressure based demands. In WDSA 2012: 14th Water Distribution Systems Analysis Conference, 24-27 September 2012 in Adelaide, South Australia (p. 1392). Engineers Australia.
8. Wang, R., Ping, J., Sun, J., et al. (2014a). Pipe network node demand calculation and scheduling method based on pressure monitoring, China Patent number 201410056986.4.
9. Wang, R., Sun, J., Yang, H., et al. (2014b). Underground pipe network leak detection methods, China Patent number 201410038082.9.
10. Wu, Z. Y., Wang, R. H., Walski, T. M., Yang, S. Y., Bowdler, D., & Baggett, C. C. (2009). Extended global-gradient algorithm for pressure-dependent water distribution analysis. Journal of Water Resources Planning and Management, 135(1), 13-22.

Generalized Spring Tensor Model: A New Improved Load Balancing Method in Cloud Computing

Shahrzad Aslanzadeh and Zenon Chaczko

1 Introduction

Over the last decade, business and academic requirements from technology perspective have changed substantially with a greater emphasis on more powerful computing techniques. In IT, much of these changes have been driven by prompt success in Internet improvement and economical IT infrastructure development, which resulted in novel structured computational models [1]. Cloud computing is one of these newly emerged paradigms for hosting and delivering services over the Internet.

In cloud computing mapping a proper load-balancing algorithm was always an important challenge. The load on the network can be forced by CPU load, memory load, bandwidth load and tasks load [2]. Therefore due to the extensive load volume, the load balancer should prioritize the information using the distributed and heuristic algorithms. Moreover, the load should be managed in a real time manner to prevent the overloaded pipelines and respond to the user requirements as quick as possible [3]. Reviewing the literature, different collections of algorithms have been proposed by researchers. Although the algorithms can optimize the load balancing methods, still there is a need for designing a method that can forecast the load patterns according to application types.

The purpose of this research is to architect a nature base heuristic load balancing algorithm, which can anticipate the fluctuation and magnitude of the load with various mathematical apparatus. The proposed approach can help to alleviate the problem of un-balanced load by visualizing the task dependencies pattern that can result in effective load management and resource monitoring.

S. Aslanzadeh (✉) • Z. Chaczko
University of Technology, Sydeny, Center of Real Time Information Network (CRIN), Sydeny, Australia
e-mail: Shahrzad.Aslanzadeh@uts.edu.au; Zenon.Chaczko@uts.edu.au

2 Problem Formulation

There is a need to develop an efficient load management tool in cloud computing that can monitor the load in an elastic and scalable cloud. The load monitoring tool must not only consider the optimization method to distribute the load effectively, but also it should have anticipatory characteristics that can perform an optimal decision making.

Different types of load balancing methods have been designed using static, dynamic and hybrid algorithms. However, there are limited numbers of examination on load balancing algorithms with dependent patterns. This could be due to the complex nature of the workflow tasks and its vague behavior in terms of resource management [4]. Despite of the limited number of the works, valued solutions have been proposed which shaped the research direction in workflow load scheduling and highlighted the main gaps and their possible solutions.

Today, workflow scheduling is considered to be a key tool for automating the e-business and e-science applications. Critical applications such as earthquake modeling, climate forecasting and online booking systems for hotels and aircrafts are the example of these groups [5]. Therefore due to the complex procedure of data processing in these applications there is a need to architect a comprehensive tool incorporated with the heuristic algorithms to predict the direction and magnitude of the load changes on workflow structured applications.

3 Proposed Solution: Generalized Spring Tensor Model (STeM)

In this research we aim to evaluate the fluctuation and magnitude of the load changes in cloud computing through generalized spring Tensor algorithm. The scope of the experiment will be limited to workflow tasks, where tasks and jobs have certain dependencies to each other.

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_121,

Generalized spring tensor (STeM) is part of the coarse gained models. It is composed of two main components [6–7]:

- Gaussian Network Model (GNM)
- Anisotropic Network Model (ANM)

GNM is designed to predict the magnitude of the load while ANM is concentrating on the direction of the fluctuation.

Basically GNM will be effective if the tasks are located in certain distances, or in other words they are connected to each other in somehow [8].

From mathematical point of view this connectivity can be explained with Kirchhoff Matrix, also illustrated in equation 1 where r_c is representing the cut-off distance.

$$\tau_{ij} = \begin{cases} -1 & if \quad i \neq j \cap r_{0,ij} \leq r_c \\ 0 & if \quad i \neq j \cap r_{0,ij} > r_c \\ \sum_{j,j\neq i}^{N} \tau_{ij} & if \quad i = j \end{cases} \tag{1}$$

ANM, however, was suggested as a coarse gained model which is using the simpler Hookian potentials to bypass the energy minimization needed for measuring the direction of the load [9].

ANM is using Hessian matrixes shown in equation 2 with $N \times N$ super elements, where each element is a 3×3 tensor and H_{ij} is the interaction tensor between i and j [10–11].

$$H_{ANM} = \begin{bmatrix} H_{1,1} & \cdots & H_{1,N} \\ \vdots & \ddots & \vdots \\ H_{N,1} & \cdots & H_{N,N} \end{bmatrix} \tag{2}$$

Both ANM and GNM models have their own advantages and disadvantages. They are considered as croase-gained models which do not require any energy minimization techniques. However as it is described earlier, ANM is focusing on direction of the fluctuation, while GNM is highlighting the magnitude of that.

Therefore to take the advantages of these two coarse-gained models and overcome their limitations, generalized spring tensor (STeM) was proposed, as a result of ANM and GNM combination. STeM can calculate the magnitude and direction of the load fluctuation in multi-dimensional environment base on nodes interactions [12].

To explore STeM model on cloud computing, a simple workflow application in a static structure should be considered. As shown in figure 1, each node in this workflow structure is representing a particular task, while combination of the tasks are representing group of jobs.

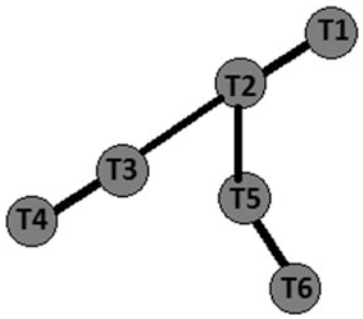


Fig. 1 Simple workflow job modeling

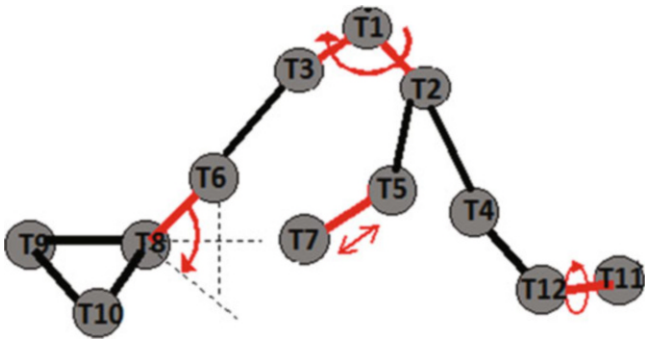


Fig. 2 Go-Like potentials main parameters on workflow job modeling

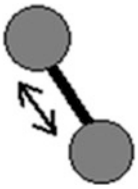
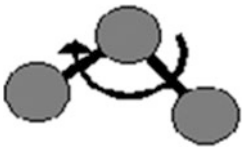
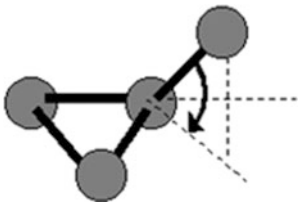
STeM algorithm is functioning base on Go-like model. In this model, it is recognized that each task has a defined location comparing to other neighbors while they are connected to each other by a single spring [13].

Therefore in early step of applying the STeM algorithm, the first derivative of the Go-like potential will be determined. "Chain connectivity", "Bond Angle", "Bond dihedral" and "Non-local" interaction between tasks are the main parameters that should be computed using equation 3.

$$\begin{aligned} V(X,X_0) = & \sum BondV_1(r,r_0) + \sum AnglesV_1(\theta,\theta_0) \\ & + \sum DihedralV_3(\Phi,\Phi_0) + \sum Non \\ & - LocalV_4\left(r_{ij},r_{0,ij}\right). \end{aligned} \tag{3}$$

Figure 2 is interpreting the main parameters of Go-like potential; bond, Angle, Dihedral and non-local interactions; on workflow tasks model. Details of each parameter are illustrated in table 1. It should be mentioned that, each task is composed of different threads with variety of commands and functionalities. Each of these threads should have priority numbers while they should start and finish by a certain time. Therefore in workflow model each task has its own depth, start-time and finish-time. Depth is depicting task's priorities in terms of execution, while start-time and finish-time show the period needed to lock a particular resource for

Table 1 Interpretation of Bond, Angle, Dihedral and non-local connections between task on workflow model

Model	Description
	*Bond*: This parameter is defining the chain connectivity between tasks in workflow models. It will mainly highlight the potential connections between a task and its neighbors [15].
	*Angle*: This parameter is defining the angles between tasks in workflow model. The angle can be interpreted as the interval between finishing time of one task comparing with starting time of the neighbor task.
	*Dihedral*: The parameter is describing the location of the tasks after they forced by external load. Torsion can disconnect the current connections between two nodes and it can substitute that with a new relation between non-neighboring nodes [16].
	Non-local interaction: The parameter is defining the task's connectivity with other tasks through non-local interactions. With this model it is implied that, the forces of changes between non-neighboring tasks can be calculated.

$$
H_{ij} = \begin{bmatrix} \frac{\partial^2 V_{1(r,r_0)}}{\partial X_i \partial X_j} & \frac{\partial^2 V_{1(r,r_0)}}{\partial X_i \partial Y_j} & \frac{\partial^2 V_{1(r,r_0)}}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V_{1(r,r_0)}}{\partial Y_i \partial X_j} & \frac{\partial^2 V_{1(r,r_0)}}{\partial Y_i \partial Y_j} & \frac{\partial^2 V_{1(r,r_0)}}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V_{1(r,r_0)}}{\partial Z_i \partial X_j} & \frac{\partial^2 V_{1(r,r_0)}}{\partial Z_i \partial Y_j} & \frac{\partial^2 V_{1(r,r_0)}}{\partial Z_i \partial Y_j} \end{bmatrix}
+ \begin{bmatrix} \frac{\partial^2 V_{2(\theta,\theta_0)}}{\partial X_i X_j} & \frac{\partial^2 V_{2(\theta,\theta_0)}}{\partial X_i \partial Y_j} & \frac{\partial^2 V_{2(\theta,\theta_0)}}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V_{2(\theta,\theta_0)}}{\partial Y_i \partial X_j} & \frac{\partial^2 V_{2(\theta,\theta_0)}}{\partial Y_i \partial Y_j} & \frac{\partial^2 V_{2(\theta,\theta_0)}}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V_{2(\theta,\theta_0)}}{\partial Z_i \partial X_j} & \frac{\partial^2 V_{2(\theta,\theta_0)}}{\partial Z_i \partial Y_j} & \frac{\partial^2 V_{2(\theta,\theta_0)}}{\partial X_i \partial Z_j} \end{bmatrix}
+ \begin{bmatrix} \frac{\partial^2 V_{3(\varphi,\varphi_0)}}{\partial X_i \partial X_j} & \frac{\partial^2 V_{3(\varphi,\varphi_0)}}{\partial X_i \partial Y_j} & \frac{\partial^2 V_{3(\varphi,\varphi_0)}}{\partial X_i Z_j} \\ \frac{\partial^2 V_{3(\varphi,\varphi_0)}}{\partial Y_i X_j} & \frac{\partial^2 V_{3(\varphi,\varphi_0)}}{\partial Y_i Y_j} & \frac{\partial^2 V_{3(\varphi,\varphi_0)}}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V_{3(\varphi,\varphi_0)}}{\partial Z_i \partial X_j} & \frac{\partial^2 V_{3(\varphi,\varphi_0)}}{\partial Z_i \partial Y_j} & \frac{\partial^2 V_{3(\varphi,\varphi_0)}}{\partial X_i \partial Z_j} \end{bmatrix}
+ \begin{bmatrix} \frac{\partial^2 V_{4\left(r_{ij},r_{0,ij}\right)}}{\partial X_i \partial X_j} & \frac{\partial^2 V_{4\left(r_{ij},r_{0,ij}\right)}}{\partial X_i \partial Y_j} & \frac{\partial^2 V_{4\left(r_{ij},r_{0,ij}\right)}}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V_{4\left(r_{ij},r_{0,ij}\right)}}{\partial Y_i \partial X_j} & \frac{\partial^2 V_{4\left(r_{ij},r_{0,ij}\right)}}{\partial Y_i \partial Y_j} & \frac{\partial^2 V_{4\left(r_{ij},r_{0,ij}\right)}}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V_{4\left(r_{ij},r_{0,ij}\right)}}{\partial Z_i \partial X_j} & \frac{\partial^2 V_{4\left(r_{ij},r_{0,ij}\right)}}{\partial Z_i \partial Y_j} & \frac{\partial^2 V_{4\left(r_{ij},r_{0,ij}\right)}}{\partial X_i \partial Z_j} \end{bmatrix} \tag{4}
$$

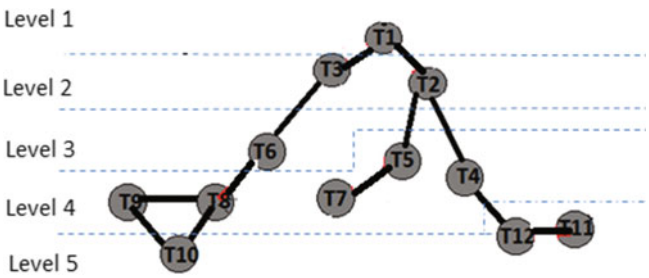


Fig. 3 Workflow job modeling in hierarchical time manner

task completion. Figure 3 is interpreting the workflow model in hierarchical time slots.

In next step, the second derivative of the obtained Go-like model will be calculated to determine the projection of the magnitude and direction of the load fluctuation on workflow load balancing [14]. This value can be obtained using equation (4).

Table 2 Depth, start-time and finish of the selected tasks in 8 time slots

Cloud-Let ID	Depth	Start-Time	Finish Time
100	0	0.1	0.21
1	1	0.21	13.34
40	1	0.21	14.08
22	2	13.48	24.7
78	2	45.36	56.17
79	3	56.17	61.13
85	4	61.13	66.47
82	5	66.47	77
96	5	66.47	77.45
97	6	77.45	85.71
98	7	85.71	95.13
99	8	95.31	102.22

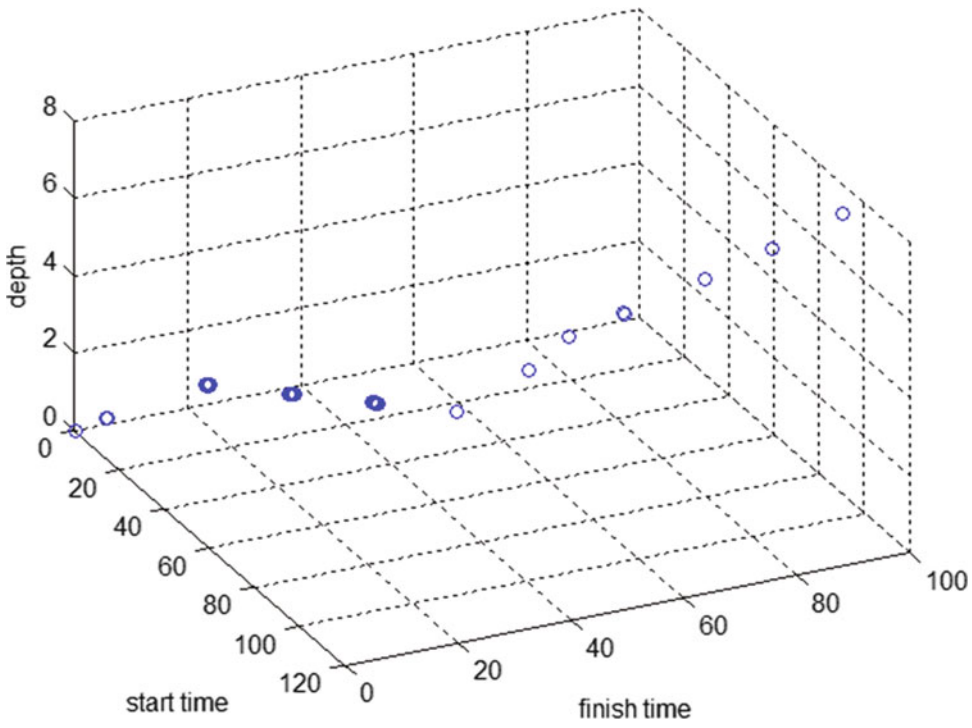


Fig. 4 Simulated workflow tasks with depth, start time, finish time

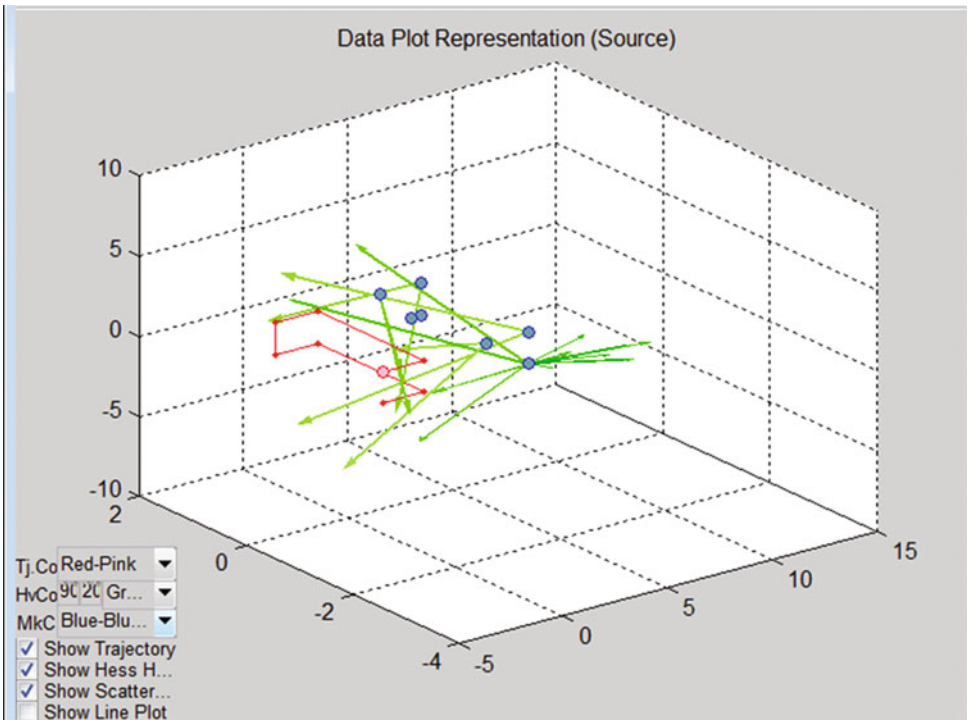


Fig. 5 Determinant values of analyzed Hessian for data set of 100 tasks in one set of job

4 Results and Discussion

In this work, we simulated a deployment of homogenous set of nodes to capture the global behavior of cloud load. Using Matlab, simulation was performed and evaluated for hundreds of nodes that represent the masses of the workflow tasks. Figure 3 is representing 100 workflow tasks, recognized by their depth, start-time, and finish-time. This workflow task has 8 main levels of time slots. The details of the first and the last time-slots of the tasks are shown in table 2.

Applying the STeM algorithm, figure 4 shows the determinant values of the analyzed Hessian for the data set of 100 tasks in one set of job. It shows the direction of the load on

each task. The magnitude of the load is shown with red color, illustrating the greater value of change. As the future work, several scenarios will be considered to evaluate the behavior of the load in cloud.

5 Conclusion

This research explores the adoption of STeM algorithm for visualizing the behavior of the load fluctuations on workflow tasks, in localized and globalized pattern. In our study, Elastic Network Model (ENM) analysis was selected to evaluate the dynamics of the cloud load. Amongst the popular algorithms of ENM, Generalized Spring Tensor (STeM) was adopted to satisfy our objectives.

The foundation of STeM algorithm is based on Gaussian Network Model (GNM) and Anisotropic Network Model (ANM) which is able to detect of the motions of tasks in cloud network. The approach helps us to identify and model the magnitude and direction of the load fluctuations at a single task in a workflow application model. Simulation result tested on 100 tasks demonstrates that STeM algorithm can visualize the tasks connectivity with both local and non-local interactions. The expected benefit of our proposed model shows that STeM algorithm can be applied in designing an effective, dynamic and autonomous load balancer that is able to support optimal decision making in critical situations.

References

1. Zhang, Q., Cheng, L. & Boutaba, R. 2010. Cloud computing: state-of-the-art and research challenges. Journal of Internet Services and Applications, 1, 7–18
2. M. Armbrust et al., Above the Clouds: A Berkeley View of Cloud Computing, tech. report EECS–28, Univ. of California, Berkeley, 2009.
3. Yike, G., Ghanem, M. & Rui, H. Does the Cloud need new algorithms? An introduction to elastic algorithms. Cloud Computing Technology and Science (CloudCom), IEEE 4th International Conference on, 3–6 Dec. 2012. 66–73.
4. Barrett, E., Howley, E. & Duggan, J. A Learning Architecture for Scheduling Workflow Applications in the Cloud. Web Services (ECOWS), Ninth IEEE European Conference on, 14–16 Sept. 2011. 83–90.
5. Gupta, A., Sarood, O., Kale, L. V. & Milojicic, D. Improving HPC Application Performance in Cloud through Dynamic Load Balancing. Cluster, Cloud and Grid Computing (CCGrid), 13th IEEE/ACM International Symposium on, 13–16 May 2013. 402–409.
6. Tu-Liang, L. & Guang, S. Generalized spring tensor models for protein fluctuation dynamics and conformation changes. Bioinformatics and Biomedicine Workshop,. BIBMW. IEEE International Conference on, 1–4 Nov. 2009. 136–143.
7. Xing, L., Karimi, H. A., Yang, L. W. & Bahar, I. Protein functional motion query and visualization. Computer Software and Applications Conference,. COMPSAC 2004.
8. Gregorcic, G. & Lightbody, G. 2007. Local Model Network Identification With Gaussian Processes. Neural Networks, IEEE Transactions on, 18, 1404–1423.
9. Chaczko, Z. & Aslanzadeh, S. C2EN: Anisotropic Model of Cloud Computing. Systems Engineering (ICSEng), 21st International Conference on, 16–18 Aug. 2011. 467–473.
10. Guang, Spring tensor model source code, 2012, http://www.cs.iastate.edu/~gsong/CSB
11. Aslanzadeh,S & Chaczko, Z. Article: Generalized Spring Tensor Algorithms: with Workflow Scheduling Applications in Cloud Computing. International Journal of Computer Applications 84(7):15–17, December 2013. Published by Foundation of Computer Science, New York, USA
12. Kaya, H.; Liu, Z.R.; Chan, H.S. Chevron Behavior and isostable enthalpic barriers in protein folding: Successes and limitations of simple Go-like modeling. Biophys. J. 2005, 89, 520–535
13. Gashler, M.; Martinez, T. (2011). "Tangent Space Guided Intelligent Neighbor Finding". Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'11). pp. 2617–2624.
14. Nakamura, H. K., Sasai, M. & Takano, M. 2004. Scrutinizing the squeezed exponential kinetics observed in the folding simulation of an off-lattice Go-like protein model. Chemical Physics, 307, 259–267.
15. Hamelryck T, Kent JT, Krogh A (2006) Sampling Realistic Protein Conformations Using Local Structural Bias. PLoS Comput Biol 2 (9): e131.
16. Blondel, A., & Karplus, M., 1998. "New formulation for derivatives of torsion angles and improper torsion angles in molecular mechanics: Elimination of singularities". Journal of Computational Chemistry 17 (9): 1132–1141.

Middleware Solution for Cross-Site Data Transfer

Zenon Chaczko, Shahrzad Aslanzadeh, and Mehdi Soltani

1 Introduction

Availability and cross-site data transfer is a reoccurring and a growing concern in software intensive systems. Cloud systems services can be turned offline due to system maintenance, communication infrastructure conservation, and power outages or due to possible service denial attacks. Fundamentally, reliable cross-site data transfer is determined by the time that the system is up and running correctly and efficiently; the length of time between failures and the length of time needed to resume operation after a failure. Availability of cross-site data transfer mechanisms need to be analyzed through the use of presence information, forecasting usage patterns and dynamic resource scaling [1].

The work presented in this paper aims to demonstrate and discuss a critical role of reliable cross-site data transfer mechanisms in improving and maintaining the availability in cloud systems. Application of middleware based solutions, across multiple availability zones, reduces the chance of failures that could simultaneously affect the services in cloud systems. Data transfer techniques, in the area of cloud computing, reduce costs associated with document management systems and maximizes availability of resources reducing the amount of downtime that affect businesses during outages. This article discusses possible ways to improve the performance of cloud networks by the introduction of cross-site transfer mechanism that uses the message-oriented middleware within the web service oriented model of software architecture.

While some companies moved around 80 % of their solution to cloud, there were still 20 % of their services which still have to run inside traditional network in their technology center due to often very large volumes of mission critical messages that need to be transferred from technology center to data center and vice versa as well as a shortage of technologies and equipment required. In some cases, after migrating services to data center in the cloud, reliability issues could start to show up. If not careful, companies for which a reliable cross–site data transfer is critical may even face a risk of losing their business.

2 Reliable Message Patterns

I have started doing some research around the web and read some books regarding SOA Patterns to find out the best solution my problem. The question is: How can services communicate reliably when implemented in an unreliable environment? During my research I came across a very great book, SOA Design Patterns by Thomas Erl, which changed my view about Service Oriented Architecture. I found Reliable Messaging [Little, Rischbeck, Simon] pattern in this book as an abstract to my problem so I have decided to start from this pattern [2, 3].

Problem:

Service communication cannot be guaranteed when using unreliable messaging protocols or when dependent on an otherwise unreliable environment.

Solution:

An intermediate reliability mechanism is introduced into the inventory architecture, ensuring that message delivery is guaranteed.

According to this pattern, an intermediate reliability mechanism needs to be introduced to guarantee the delivery of messages in unreliable environment. Hence, a decision was made to start with finding the right solution for an intermediate reliability mechanism. It was clear that a Messaging Queue middleware need to be introduced to solve this issue.

Z. Chaczko (✉) • S. Aslanzadeh • M. Soltani
University of Technology, Sydeny, Center of Real Time Information Network (CRIN), Sydeny, Australia
e-mail: Zenon.Chaczko@uts.edu.au; Shahrzad.Aslanzadeh@uts.edu.au; Mehdi.Soltani@uts.edu.au

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_122,

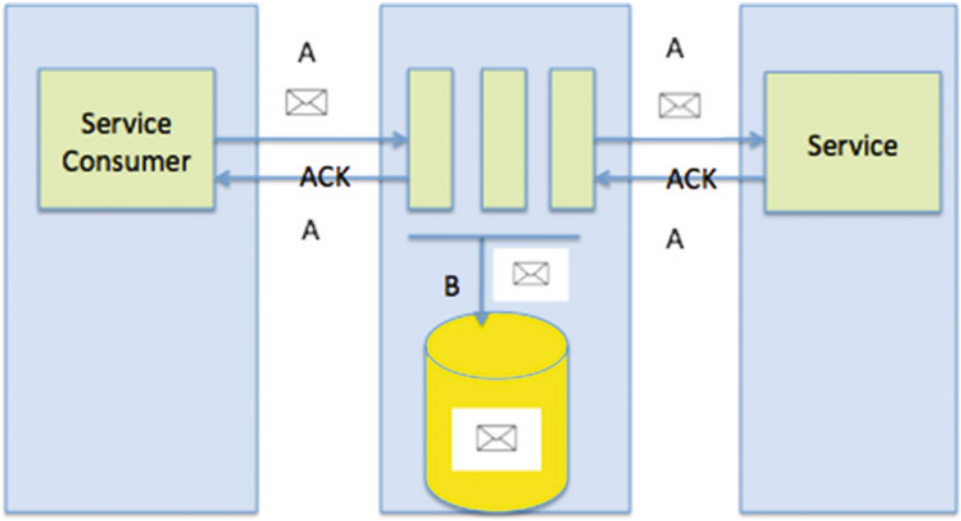


Fig. 1 Reliable Messaging Pattern

3 Messaging Queue Middleware

Message queues provide an asynchronous communications protocol, meaning that the sender and receiver of the message do not need to interact with the message queue at the same time. Messages placed onto the queue are stored until the recipient retrieves them. These message queuing systems typically provide enhanced resilience functionality to ensure that messages do not get "lost" in the event of a system failure. They provide messaging patterns such as Store & Forward and Publish/Subscribe [4].

4 Store & Forward Messaging

In this model, when the client process calls an API to send a message to the server process, the API returns control to the calling thread before the message is sent. At that point, the transfer of the message across the network becomes the responsibility of the messaging technology. There may be various kinds of communications interference, the server machine may simply be down, or a firewall may be slowing down the transfer. Also, even though the message may have reached the target machine, the target process may currently be down. While all of this is going on, the client process is oblivious. Critical resources like threads (and it's allocated memory) are not being held waiting for the call to complete. This prevents the client process from losing stability as a result of having many threads and all their memory used up waiting for a response from the other machine or process [5].

5 1.4 Publish / Subscribe

In this style, the sender of the message often does not know about the specifics of those that wish to receive the message. This additional loose coupling comes at the cost of subscribers explicitly opting-in to receiving messages. Subscribers need to know about which endpoint is responsible for a given message. This information is usually made available as part of the contract, specifying to which endpoint a subscriber should send its request. As a part of the subscription message, a subscriberpasses its "return address", the endpoint at which it wants to receive messages. Keep in mind that the publisher may choose to store the information about which subscriber is interested in which message in a highly available manner. This would allow multiple processes on multiple machines to publish messages to all subscribers, regardless if one had received the subscription message or not. Subscribers don't necessarily have to subscribe themselves. Through the use of the Return Address pattern, one central configuration station could send multiple messages to each publisher specifying which subscriber endpoints to subscribe to which message. Another option that can be used is for multiple physical subscribers to make themselves appear as one single logical subscriber. This makes it possible to load balance the handling of messages between multiple physical subscribers without any explicit coordination on either the part of the publisher or the part of any one subscriber. All that is needed is for all subscribers to specify the same return address in the subscription message [6]. Publishing a message involves having the message arrive at all endpoints which had

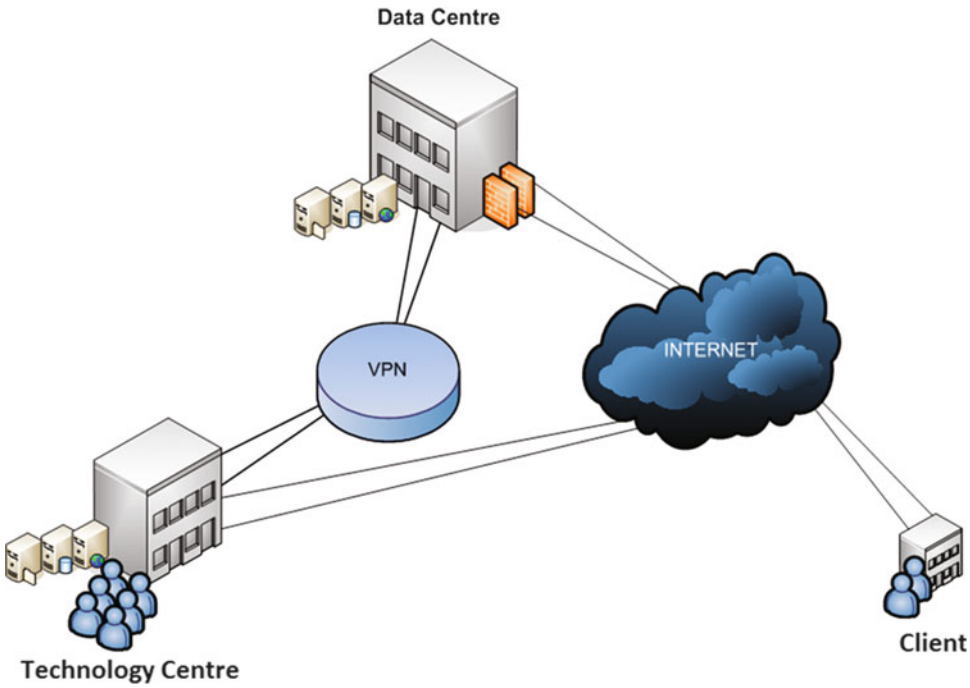


Fig. 2 Simple architecture of load balancing algorithm

previously subscribed to that type of message. Messages which are published often represent events - things that have happened, for instance Order Cancelled, Product Out of Stock, and Shipping Delayed. Sometimes, the cause of an event is the handling of a previous command message, for instance Cancel Order. A publisher is not required to publish a message as a part of handling a command message although it is the simplest solution. Since many command messages can be received in a short period of time, publishing a message to all subscribers for every command message multiplies the incoming load and, as such, is a less than optimal solution. A better solution would have the publisher roll up all the changes that had occurred in a given period of time into a single published message. The appropriate period of time is dependent on the Service Level Agreement of the publisher - its commitment to the freshness of the data. Another advantage of publishing messages on a timer is that that activity can be offloaded from the endpoint/server processing command messages effectively scaling out over more servers. Examples of commercial implementations of this kind of message queuing software also known as message-oriented middleware include [7]:

- IBM's WebSphere MQ (formerly MQ Series)
- Oracle Advanced Queuing (AQ)
- Java Message Service (JMS)
- Microsoft Messaging Queue (MSMQ)

There are a number of open source choices of messaging middleware systems, including:

- JBoss Messaging
- JORAM
- Apache ActiveMQ
- RabbitMQ
- ZeroMQ

6 System Architecture

The options we had regarding the Architecture of the system were Broker Style Architecture and Bus Style Architecture [8].

The company used Microsoft BizTalk as integration middle-ware for years and they were really keen to use BizTalk for this solution as well. BizTalk is a man in the middle and it introduces the Broker style architecture which wasn't fit into the proposed solution for the following reasons:

- We need to deploy this middleware in all sites which cost lots of money in terms of licensing.
- We are not really using many features of BizTalk for this solution like Orchestration, BAM and BRE which makes the BizTalk over complicated solution for our scenario.
- Broker style architecture is not a right choice for this solution as we need to deploy this in distributed servers so something like Bus style architecture should work better.

7 Bus Architecture

While the Broker style architecture introduces the central box in the middle which all communications goes through, a Bus isn't necessarily a physical entity and it will distributed across several servers. This is exactly what we need for our solution as we need to distribute our process across several

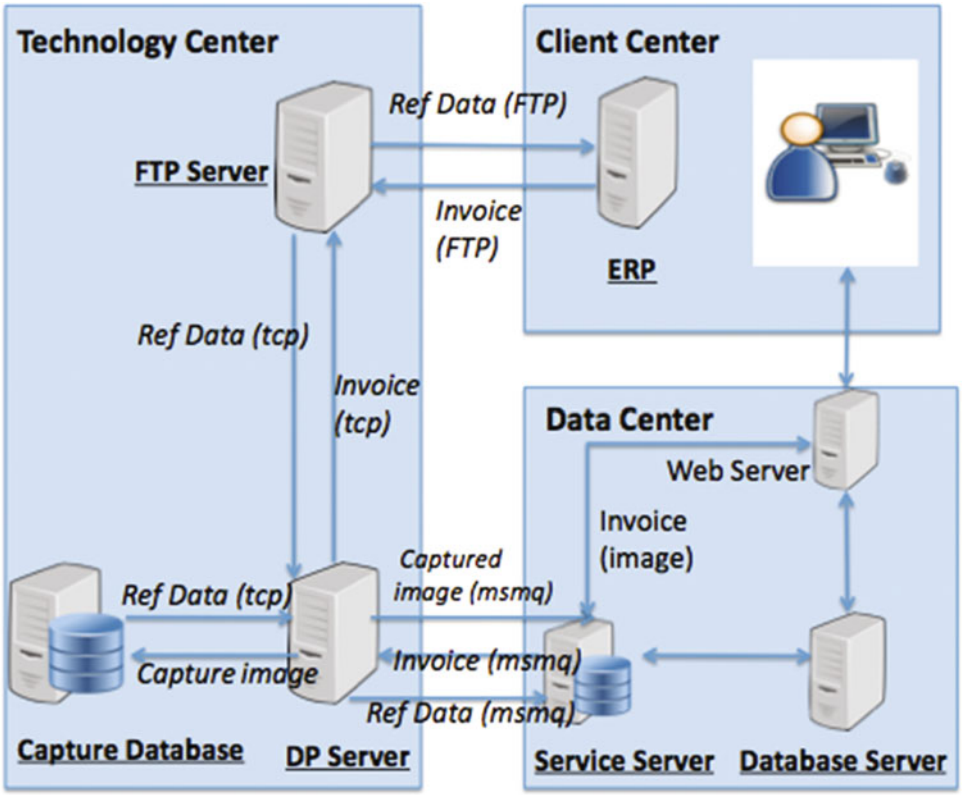


Fig. 3 High Level Data Flow Diagram

servers and sites [9]. After some research a decision was made to choose the NServiceBus as an open source ESB (Enterprise Service Bus) for our solution. The diagram in Fig 2 shows the high level cloud architecture including the Technology Centre, Data Centre and Client sites. The diagram in Fig. 3 shows one of the client implementations which utilize this solution for transferring data between different sites.

8 Statistical Analysis

The final solution has been successfully deployed into production and it is currently running for several months on production environment without any major issues. The result is very satisfactory since not even a single message was lost after deploying this solution to production, while previously for many months average loses of messages during transferring between test site and data center were reaching around 9.8 %. This means that 4710 out of 48056 messages were lost in total, during three months test period [10]. Considering mission critical data and messages, lots of support resources have been involved to investigate and recover the lost messages manually. Fig 4 shows the number of total as well as failed transactions in last 6 months. In first 3 months we were using legacy File Transfer solutions while late March we switched to new solution.

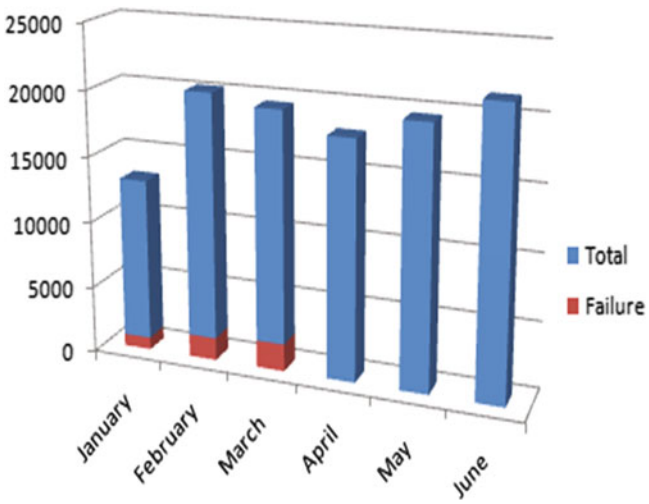


Fig. 4 Number of Total and Failed Transactions in Last 6 months

In two following pie charts, fig 5 and fig.6, the percentage of failure transactions has been shown before and after applying the solution:

We have been involved in lots of challenges to make this project successful. Much time has been spent evaluating existing vendor specific middleware products as well as open source solutions available on the market. It was found that the proposed solution addresses problems that are very common for many companies that are moving to cloud.

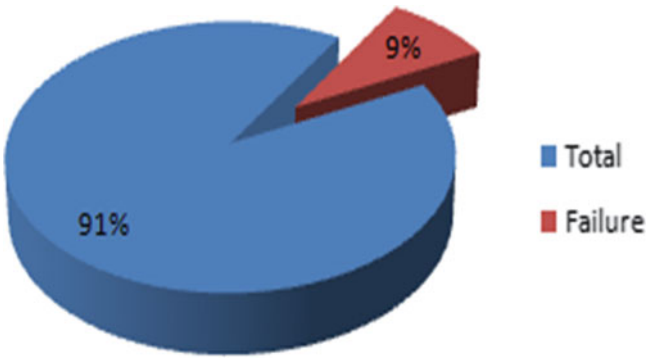


Fig. 5 Failure Percentage before Applying the Solution

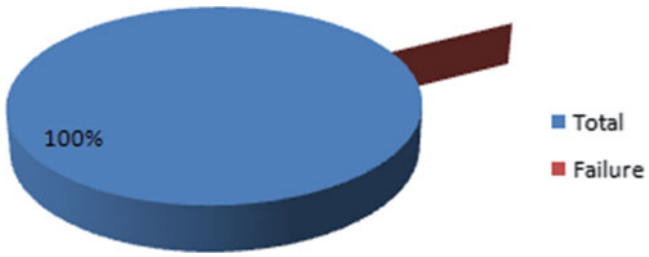


Fig. 6 Failure Percentage after Applying the Solution Conclusion

9 Conclusion

We have been involved in lots of challenges to make this project successful. Much time has been spent evaluating existing vendor specific middleware products as well as open source solutions available on the market. It was found that the proposed solution addresses problems that are very common for many companies that are moving to cloud. This paper demonstrated the applicability of using load balancing techniques to obtain measurable improvements in resource utilization and availability of cloud-computing environment. On that basis, the proposed approach could bring major cost advantages to cloud vendors who are concerned with utility costs and who are searching for efficiencies that can be relatively easily achieved. Various models and rules can be applied to load balancers, however these should be based on the scenario the load balancer will be applied for. The network structure or topology should be taken into account when creating the logical rules for the load balancer. This is due to the pricing of transfer between regions, availability zones and cloud vendors, which all constitute different pricing strategies. Message oriented architecture as a middleware model has been pointed out to improve load balancing in distributed networks. Based on messaging techniques XMPP allowed resources to be monitored and provide availability of cloud resources.

Reference

1. Armbrust, M. et al., “A view of cloud computing,” Commun. ACM, vol. 53, pp. 50-58, 2010.
2. Carter, R.L. et al., “Resource allocation in a distributed computing environment,” in Digital Avionics Systems Conference, 1998. Proceedings., 17th DASC, AIAA/IEEE/SAE, 1998, Vol. 1, pp. C32/1-C32/8.
3. Durkee, D., “Why Cloud Computing Will Never Be Free,” Queue, vol. 8, pp. 20-29, 2010.
4. Erl, T., SOA Design Pattern, Prentice Hall, 2009.
5. Hohpe, G., Woolf, B., Enterprise Integration Patterns – Designing, Building, And Deploying Messaging Solutions, Addison Wesley, 2007.
6. Reese, G., “Cloud Application Architectures”, O’Reilly Media, 1st Ed., Inc. Sebastopol, CA, US, 2009.
7. Ruixia, T. and Z. Xiongfeng, Z., “A Load Balancing Strategy Based on the Combination of Static and Dynamic,” in Database Technology and Applications (DBTA), 2010 2nd International Workshop on, 2010, pp.1-4.
8. Schantz, R.E. and Schidt, D.C., “Middleware for Distributed System: Evolving The Common Structure for Network-centric Applications”, Encyclopedia of Software Engineering, New York, Wiley & Sons, 2001, pages 801-813
9. Amazon, Amazon EC2 Pricing, Viewed on 25th of Nov. 2010, available at http://aws.amazon.com/ec2/pricing/, Amazon Web Services LLC, 2010
10. Win32 system message queues. “About Messages and Message Queues”. Windows User Interface. Microsoft Developer Network. http://msdn.microsoft.com/enus/library/ms644927(VS.85).aspx.

Autonomous Model of Software Architecture for Smart Grids

Zenon Chaczko, Shahrzad Aslanzadeh, and Alqarni Lulwah

1 Introduction

One of the major challenges of the 21st century is increased energy demand. As such, there is need to ensure energy efficiency which can be realised through the use of renewable energy and smart grid. However, the ability to run a supply-on-demand system that achieves maximum reliability has emerged as a major challenge. In addition, the inability to balance between largely uncontrolled demand and highly controlled supply has emerged as a major setback on power systems. However, the use of operational smart grid has the likelihood of mitigating some of the challenges faced in the energy sector. This is because it can allow greater penetration of variable energy resources realised through a flexible management system. The concept of smart grid has emerged as more reliable compared to traditional grid. Renewable energies can be integrated through smart software grids to achieve energy efficiency. One of the major cases that can elaborate the need for smart grid and renewable energy is the electrical blackout that took place in North America in 2003 affecting more than a hundred power plants and paralysing different activities. Through the use of smart grids, it becomes easy to meet demand needs. In addition, smart grids ensure efficiency driven response and reliability. This report paper explores smart software architecture for smart grids with respect to achieving energy efficiency.

In this paper, smart grid and energy efficiency have been explored. In addition, the writer explores application of clouds platforms to enhance smart software architecture for smart grids with respect to energy efficiency. Lastly, issues of privacy and security in smart software architecture for smart grids supported by clouds have been analysed.

Z. Chaczko (✉) • S. Aslanzadeh • A. Lulwah
Faculty of Engineering and IT, University of Technology, Sydney, Sydney, Australia
e-mail: Zenon.Chaczko@uts.edu.au; Shahrzad.Aslanzadeh@uts.edu.au; Luluqarni@gmail.com

2 Significance of Smart Grid and Energy Efficiency

Smart grid is the combination of traditional grid with modern technology, control and information technologies (Potter, Archambault & Westrick 2008). In other words, smart grid incorporates communication technologies and advanced sensors with the objective of ensuring effective use of assets, increased improved reliability, and energy efficiency as well as allowing consumers to have access to dimensional energy services. Yang (2012) adds that smart grid is the process of using computer-based remote control to automate power systems. Through smart grid technology, power can be generated through renewable energy resources (wind and solar) with the aim of meeting power demand without resource waste thus achieving energy efficiency. Potter, Archambault and Westrick (2008), observe that "the use of smarter grid operations allows for greater penetration of variable energy sources through the more flexible management of the system" (p.1). The implication made is that the use of smarter grids allows flexible management systems which can be achieved through demand side management which is a form of temporary storage technology.

Basically, the existing defining features of smart grids indicate that smart grid offer an interface between traditional assets incorporated in a power system and consumer appliances. As a result, a two-way communication is achieved which provides the consumer with many options and effective control of energy usage. Through the use of sensor networks, power systems become responsive to power stimulus such as power failure. Consequently, efficiency in operations is realised especially when handling interruptions across the power systems. Since smart grid incorporates traditional grid with communication technologies, power production is decentralised. For instance, Porter et al. (2008) suggest that advanced metering technologies coupled with improved communication enhance the use of two way metering and sensors which allow decentralised power production. This allows the

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_123,

optimisation of power flows across the existing power system transmission. Consequently, it improves systems reliability and capital expenditures thus realising energy efficiency.

Lu Liang, Li, Lin and Shen (2011) argue that the electrical blackout that was experienced in 2003 across North America could have been countered through the use of smart grids. This is because, compared to tradition grid, smart grids allow effective real-time diagnosis and load balance through automated outage management, self-activating, and self healing. In addition, energy efficiency is achieved through cost effective power generation, and effective transmission, as well as effective distribution. Some components such as smart meters are vital as they not only allow real-time information collection but they relay the collected information related to grid operations and transmission in real-time (Lu Liang, Li, Lin, & Shen 2011). Moreover, smart devices have the capability to give consumers increased control and awareness thus reducing energy costs and usage (Drake, Najewicz, Watts, & General Electric Company 2010). Effective deployment of smart grid technologies allow the integration of variable renewable power, reduction of carbon dioxide emission and management of electricity demand.

3 Software Architecture for Smart Grids

For smart software architecture to be in a position to support smart grid application, one of the major intrinsic components required is a cloud platform (Simmhan, Cao, Prasanna & Giakkoupis 2011). Basically, as earlier stated, smart power grids are incorporated with sensors and smart meters which enable two-ways communication. Therefore, software systems that operate demand response optimisation need a platform that can support generation of data and carryout computation. For example, in Los Angeles city, Department of Water and Power is required to continuously analyse streaming information and data from the city consumers. Therefore, Clouds are necessary as they have "advantages of scalable and elastic resources to build a software infrastructure to support such dynamic, always-on applications" (Simmhan, Giakkoupis, Cao & Prasanna 2011, p. 1). The implication made is that clouds offer a platform that can be used to develop smart software that can support smart grids hence energy efficiency. All data and information streaming from energy smart grids can be stored in clouds as it streams in large scales. In addition, they allow smart grids operations as clouds allow scalability.

To achieve energy efficiency in a smart grid, smart software architecture is designed. This is because the essence of a smart grid is to enhance effective communications and carry out effective transmission and distribution of power (Michaels & Donnelly 2010). However, smart grid is based on evolving infrastructures which require smart software architecture from time to time. Therefore, there is need to consider software architecture for smart grid to achieve efficiency. Balaraman (2012) suggests that energy efficient smart grid devices have to be built on smart software architecture. As a result, it would be possible to meet the high demand of energy needs at a lower cost. Given that there has been more renewable energy in most states such as UK and Germany, and energy can effectively be managed through upgrading of the existing emerging grids by replacing them with modern smart software architecture which allows efficiency. As a result, it would be possible to offset peak load demand and manage congestion in smart, effective and efficient way.

The increased demand for energy coupled by rapid growing use of new renewable energy has brought changes to power networks landscapes resulting in new challenges. However, smart grids tend to address these issues especially when smart software architectures are used to integrate software that enhance self-management on smart grids (Perez, Diaz & Gonzalez 2012). It is through these architectures supported by cloud platforms (Simmhan, Giakkoupis, Cao & Prasanna 2011) that renewable energy resources and traditional power grids are integrated with new technological elements and cloud computing infrastructures which allow scalability and autonomous operations along the power system networks. To achieve these, smart software architecture need to be built based on software architectural models which incorporate power network domain. The diagram below as provided by Perez, Diaz and Gonzalez (2012) is *Smart Grid Architectural Model which combines autonomous model and smart grid architectural model to form an Autonomous Smart Grid Architectural Model which enhance energy efficiency*.

Smart Grid software architecture present in clouds allows different stakeholders such as third party service providers, utilities, and consumers to interact. For example, smart meters installed in residential places are based on home area network (HAN) or the building area network (BAN) which gather data related to power usage and relay signals (Simmhan, Giakkoupis, Cao & Prasanna 2011a). In addition, information on power pricing and energy usage needs to be shared in real-time with power consumers via online portals. This can only be achieved through the use of cloud platforms. Collected data and information which is integrated from different sources need to be accessible and available to third party applications assuming that privacy concerns are met. This benefits smart software developers in developing intelligent applications which meet customer needs. Therefore cloud, either private or public, provides a platform from where data can be shared (Simmhan, Giakkoupis, Cao & Prasanna 2011). Moreover, clouds provide a platform from where third party applications can be put into place.

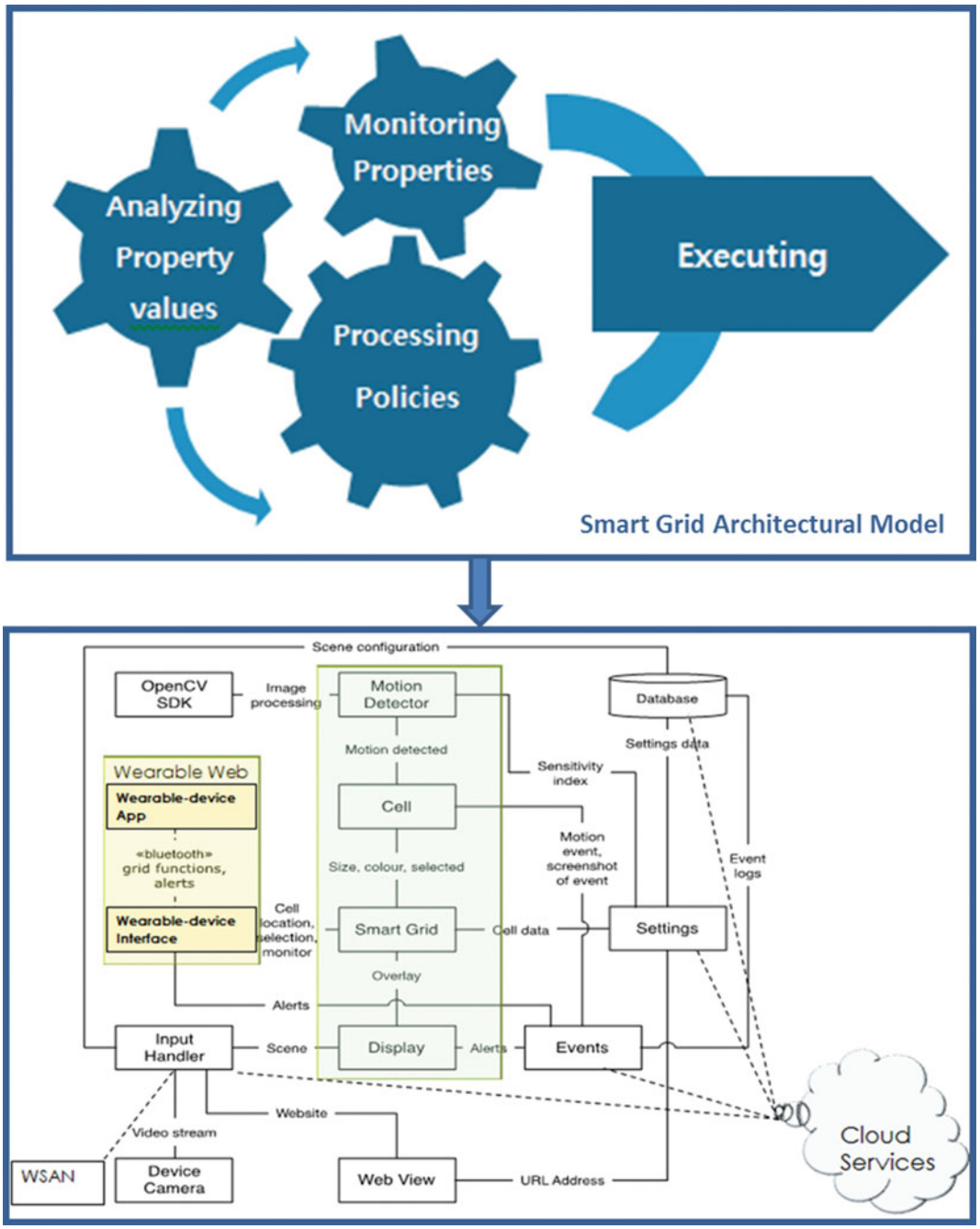


Fig. 1 Proposed Autonomous Smart Grid Architectural Model

One of the major software applications in smart grid architecture is the advanced meter infrastructure (AMI). AMI is defined as a software application that ensures that energy use is measured and the reads communicated over to the utility. In the modern world, companies are upgrading their traditional power grids to smart grids through the use of AMIs which are installed to allow two-way communications between consumers and the utilities in real-time (Simmhan, Cao, Prasanna, & Giakkoupis 2011). Smart meters increase the volume of data and information available to power systems and utilities. For instance, City of Los Angeles Department of Water and Power has been on the verge of achieving demand-response (DR) optimisation which can only be realised through smart meters built on smart software architecture and supported by cloud platform. Basically, smart enabled grids have the capability to discover faults in power systems. As a result, an average of 20 to 30 % of power is saved compared to traditional grids (Michaels & Donnelly 2010). AMI offer operating advantages such as outage management, reduced meter reading costs, and

granular visibility. To achieve highest peak demand savings, HAN technologies are adopted. Based on a research conducted in California in 2003, households can save 34.5 % through the use of enabling technology.

4 Security and Privacy Issues in Software Intensive Architectures for Smart Grids

One of the major challenges being faced in the incorporation of renewable energies into smart grid architecture is cost. This is because multiple technologies are used to ensure that integration which may be costly. Another challenge is security. As noted by Birman, Ganesh, and van Renesse (2012), security is important in ensuring a successful secured smart grid communications. To prevent issues related to insecurity, there is need to ensure report authentication and data integrity. Legacy is experienced when there a movement of utilities to smart grids which are based on smart grid software architecture. According to (Simmhan, Giakkoupis, Cao and Prasanna (2011a), not all utilities have the capacity to restructure all the power system. Therefore, co-existence of legacy is allowed. When this happens, privacy and security frameworks are expected to be compatible with existing and new applications. In other others, new applications should be integrated within existing applications without fear for privacy and security issues. However, vulnerabilities and threats that may emerge from insecure legacy smart devices need to be established. Therefore, smart grid applications need to be based on smart software architecture that allows room for privacy and security consideration during the development process. Other methods that can be applied include the use of Smart Grid applications which allow access through Web services, online websites, mobile applications, and local executables (Simmhan et al, 2011a). The applications could also be shared in form of virtual machine image thus ensuring energy usage information security.

A major concern has been raised with the ability of clouds to protect private data and information. As noted by Birman, Ganesh, and van Renesse (2012), current cloud platforms lack protective mechanisms on consumers' private information which has raised consumers concerns. Simmhan et al. (2011a) note that customers can hack communication channels or smart meter thus changing power usage report. In addition, third party software and utilities can access personal information and incorporate power usage data with the aim of forecasting future power demand and load curtailment response. As a result, privacy concerns as a result of security issues need to be addressed while designing smart software architecture for smart grids supported by cloud platforms. Therefore, power providers shifting to smart grids need to restore customers' confidence by addressing security issues. These concerns can best be addressed before being integrated to cloud platform which support smart grids.

5 Conclusions

This study covers only some selected issues related to the performance of the wireless communication, and how to make it more efficient and better utilized. There is still much work to be done in the future to improve the efficiency of interference cancellation. In this analysis, smart clouds and energy efficiency have been explored. Smart grid combines tradition grid with communicational technologies and advanced sensors to realize energy efficiency. Consumers can access usage information through smart meters. It also allows the integration of renewable energy resources into the power systems thus meeting energy demand without energy and resource wastage. Smart grid allow effective real-time diagnosis and load balance through automated outage management, self-activating, and self healing which encourage energy efficiency. Cloud platform is necessary in supporting smart software architecture for smart grids. Clouds provide elastic and scalable resources necessary for the design of smart software infrastructure necessary for smart grids. Data and information can easily be stored in cloud platforms and shared among utilities, third party software applications, and consumers. Because of the increased energy demand and use of renewable resources, energy efficient smart grid devices have to be built on smart software architecture which is supported by clouds. Advanced meter infrastructure has emerged as major smart software necessary in ensuring energy efficiency. Issues of security and privacy have emerged as the major setback of using cloud platforms to support smart software architecture necessary for smart grids.

References

1. Balaraman, S. 2012. *Key considerations for designing low cost, energy efficient smart grid devices*. [Online]. Available at: http://www.rtcmagazine.com/articles/view/102619 (accessed 19 January 2013).
2. Birman, P K., Ganesh, L., & van Renesse, R. 2011. *Running smart grid control software on cloud computing architectures*. [Online]. Available at: (accessed 19 January 2013). http://www.cs.cornell.edu/projects/quicksilver/public_pdfs/SmartGrid-final.pdf
3. Drake, J., Najewicz, D,, Watts,W, & General Electric Company. 2010. *Energy efficiency comparisons of wireless communication technology options for smart grid enabled devices*. [Online]. Available at: http://energypriorities.com/library/ge_zigbee_vs_wifi_101209.pdf (accessed 19 January 2013).
4. Michaels, H., & Donnelly, K. 2010. Architecting *the smart grid for energy efficiency*. Massachusetts Institute of Technology.

5. Pérez, J, Díaz, J & González, E. 2012. *Designing and simulating smart grids*. Technical University of Madrid (UPM), Spain
6. Potter, C W, Archambault, A & Westrick, K. 2008. *Building a smarter smart grid through better renewable energy information*. [Online]. Available at: http://c0402442.cdn.cloudfiles.rackspacecloud.com/static/ttcms/1.0.0.44/us/documents/publications/IEEE_Smart_Grid_Final.pdf (accessed 19 January 2013).
7. Simmhan, Y, Giakkoupis, M, Cao, B & Prasanna, V. 2011. *On using cloud platforms in a software architecture for smart energy grids*. University of Southern California, Los Angeles.
8. Simmhan, Y, Giakkoupis, M, Cao, B & Prasanna, V. 2011a. *An analysis of security and privacy issues in smart*. University of Southern California, Los Angeles.

Specification and Design Method for Big Data Driven Cyber Physical Systems

Lichen Zhang

1 Introduction

Cyber physical systems (CPS)[1] such as automobile and intelligent transportation systems, aerospace systems, medical devices and health care systems are receiving a lot of attentions recently. Those systems contain a large network of sensors distributed across different components, which leads to a tremendous amount of measurement data available to system operators. Cyber-physical systems involve a complex integration of physical and computational processes. Such systems usually integrate two distinct components - (i) a set of sensors that continuously produce streaming data (ii) a set of communication and computation systems that aggregate data and perform data analytics. Real-time data from sensors are the first-class citizens in cyber-physical systems (CPS) since they interconnect cyber elements and physical elements of CPS. A real-time data service, which maintains the freshness of data while satisfying other key performance goals such as timeliness of transactions and small resource consumption, is an essential architectural component in many CPS applications.

Big data in cyber physical systems data are subdivided into two types: *sensor data* and *derived data*. Sensor data are the data issued from sensors. Derived data are the data computed using sensor data. Their design of big data driven cyber physical systems needs appropriate concepts and tools which are not available under systemic or object oriented methods. UML, the most used nowadays, cannot, in its standard form, satisfy the requirements of such design.

In this paper, we propose a big data driven cyber physical system design method based on AADL [2], which can specify and model the requirements of big data driven cyber physical systems, implement these requirements on big data platforms, guarantees QoS in a highly scalable manner.

2 Big Data Driven Cyber Physical System Design

Features of big data can be used "4V" (Volume, Velocity, Variety, Virtual) to describe [3]: Volume: large-scale data sets is generally TB or so, and big data is generally from PB to EB grade level. Velocity: big data do not consume large data warehousing and its technology of data mining is different compared with traditional, thus its processing speed is very fast. Variety: There are many big data types, no longer the traditional structured data, but more unstructured, distributed and monotonous pattern, such as a log of daily use, video, pictures and so on. Virtual: Despite the capacity of big data is very large, but once the user submits the data requirements of big data can automatically and timely extract the relevant part.

Architecture Analysis & Design Language (AADL) is proposed by Society of Automotive Engineers (SAE) [4], AADL is especially effective for model based analysis and specification of complex real-time embedded systems. The main components in AADL are divided into three parts as shown in Fig. 1 [5]: software components, hardware components and composite components. Software components include data, thread, thread group, process and subprogram. Hardware components include processor, memory, bus and device. Composite components include system.

AADL provides the data component to model data types and data abstraction. We can use the data component to model the big data of cyber physical systems. The data component category supports representing data types and data abstractions in the source text at the appropriate level of abstraction for the modeling effort. The data *type* is used to type ports to specify subprogram parameter types. Data

L. Zhang (✉)
Faculty of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510090, China

Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai 200062, China
e-mail: zhanglichen1962@163.com

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_124,

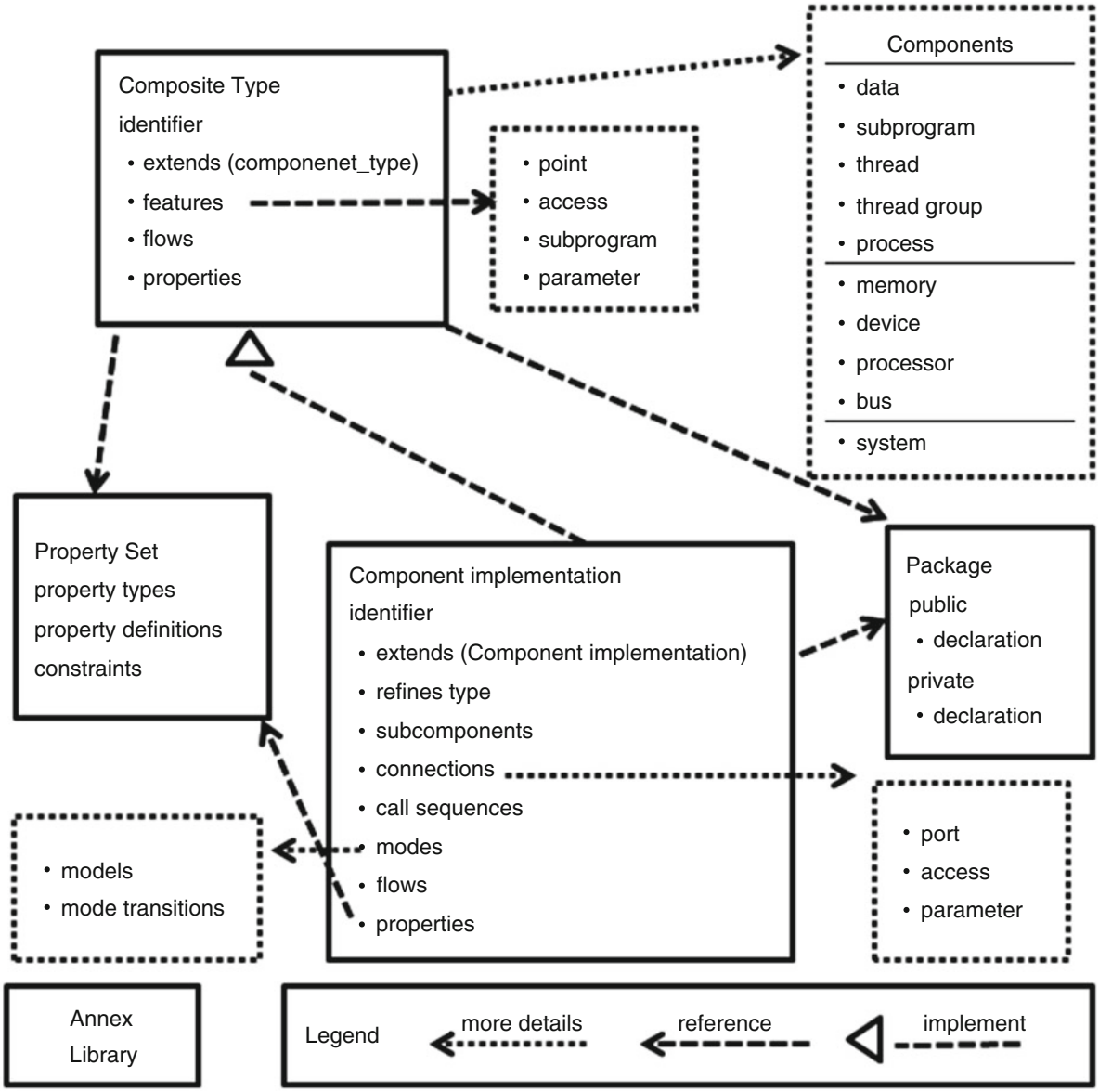


Fig. 1 AADL Elements

type inheritance can be modeled using the AADL component type extension (e.g., extends) mechanism. Class abstractions can be represented by using subprogram as a feature of a data type. Provided data features of a data component are sharable using a specified concurrency control property. A sharable data component instance is specified in a requires subclause of the component type [6]. The AADL supports modeling of three kinds of interactions between components: directional flow of data and/or control through data, event, and event data port connections; call/return interaction on subprogram entrypoints; and through access to a shared data component [7].

The data types of automatic train control system (ATC) [8] is specified as shown in Fig. 2.

The data types of automatic train control system (ATC) [9] on AADL text form is specified as follows:

```
system ATC
  features
    command_data: out data port;
    sensor_data: in data port;
    event_out: out event port;
end ATC;
system implementation ATC.instance
  subcomponents
    ATP: process ATP_process.ATP;
    ATO: process ATO_process.ATO;
    ATS: process ATS_process.ATS;
    exception: process ATC_process.general;
```

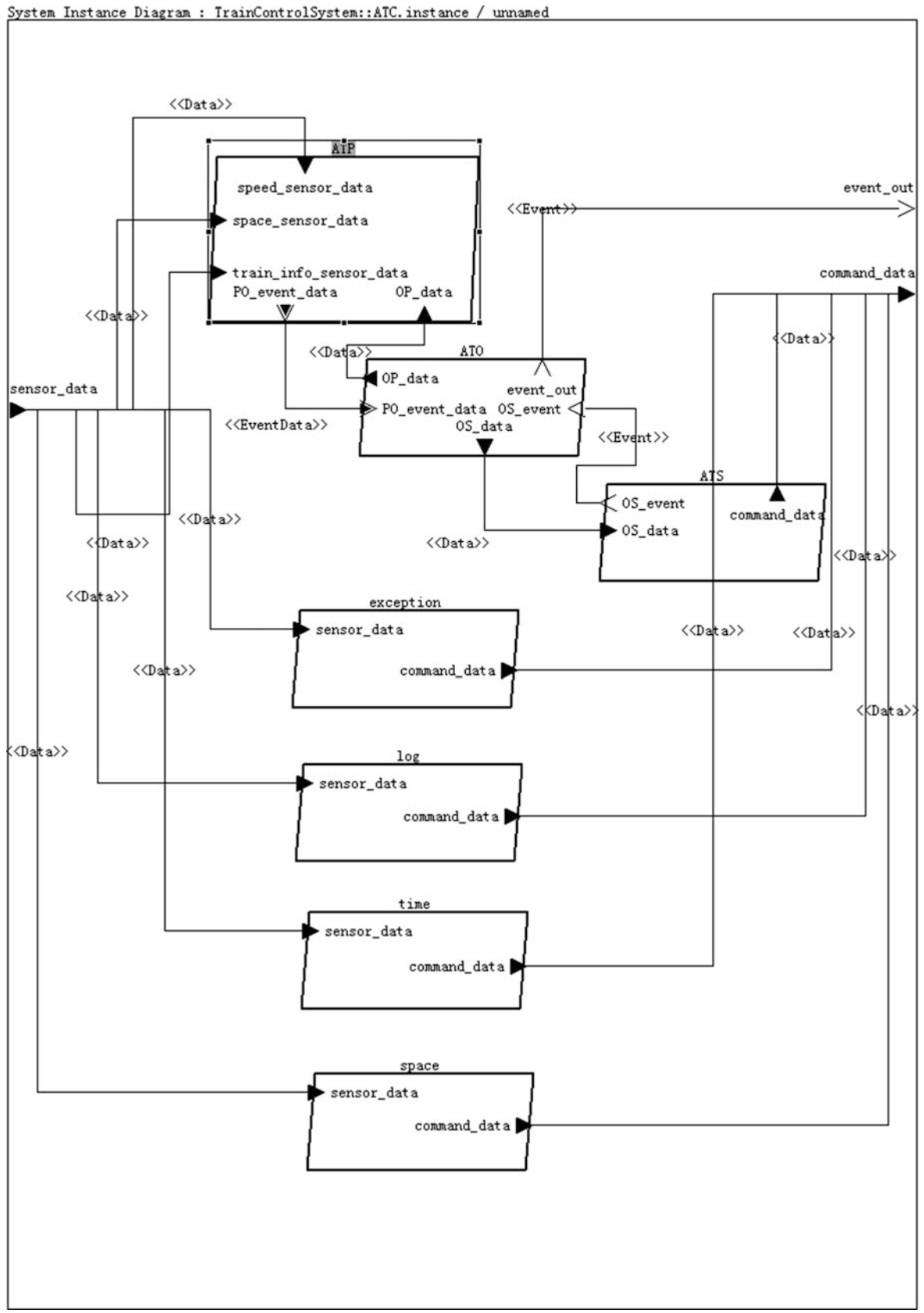


Fig. 2 The data types of automatic train control system (ATC)

log: **process** ATC_process.general;
time: **process** ATC_process.general;
space: **process** ATC_process.general;
connections
DataConnection1: **data port** sensor_data -> ATP.speed_sensor_data;
DataConnection2: **data port** ATS.command_data -> command_data;
DataConnection4: **data port** ATO.OS_data -> ATS.OS_data;
EventDataConnection3: **event data port** ATP.PO_event_data -> ATO.PO_event_data;
DataConnection5: **data port** sensor_data -> ATP.space_sensor_data;
DataConnection6: **data port** sensor_data -> ATP.train_info_sensor_data;
DataConnection7: **data port** sensor_data -> exception.sensor_data;
DataConnection8: **data port** exception.command_data -> command_data;
DataConnection9: **data port** sensor_data -> log.sensor_data;
DataConnection10: **data port** sensor_data -> time.sensor_data;
DataConnection11: **data port** sensor_data -> space.sensor_data;
DataConnection12: **data port** log.command_data -> command_data;
DataConnection13: **data port** time.command_data -> command_data;
DataConnection14: **data port** space.command_data -> command_data;
DataConnection15: **data port** ATO.OP_data -> ATP.OP_data;
EventConnection1: **event port** ATS.OS_event -> ATO.OS_event;
EventConnection2: **event port** ATO.event_out -> event_out;
end ATC.instance;

The AADL supports the concept of specifying end-to-end flows to support various forms of end to end analysis throughout a model such as end-to-end timing and latency, reliability, numerical error propagation, and processing sequences of domain objects [10]. System, process, and thread components can have flows. A flow specification declaration indicates that information logically flows from one of the incoming ports, parameters, or port groups of a component to one of its outgoing ports, parameters, or port groups. The ports can be event, event data, or data ports. The AADL standard has a set of predefined properties that can be used to model End to End data flow in big data environment. However, new properties can be added in order to represent additional information The standard properties for End to End data flows are[11]:

Expected_Latency: Time
Actual_Latency: Time
Expected_Throughput: Data_Volume
Actual_Throughput: Data_Volume

The flow specification of display interface workstation of train control systems is as follows:

device DisplayInterfaceWorkstations
features
DispDataInterface: out data port;
...
flows
flow1:flow source DispDataInterface{
Latency => 20 Ms;
};
end DisplayInterfaceWorkstations;
process DisplayProcess
features
inInterface: in data port;
outInterface: out data port;
flows
disp_flow:flow path inInterface-> outInterface{
Latency => 20 Ms;
};
end DisplayProcess;
device Displayer
features
DispDataInterface: in data port;
Power: requires bus access Tools::PowerSupply.Power;
flows
flow2:flow sink DispDataInterface{
Latency => 20 Ms;
};
end Displayer;
system implementation ATSys.Impl
subcomponents
DispWorkstations: device DisplayInterfaceWorkstations;
Disp: device Displayer;
DispProcess: process DisplayProcess;
...
connections
DataPortAccessConn1: data port DispWorkstations.DispDataInterface ->
DispProcess.inInterface;
DataPortAccessConn2: data port DispProcess.outInterface -> Disp.DispDataInterface;
...
flows

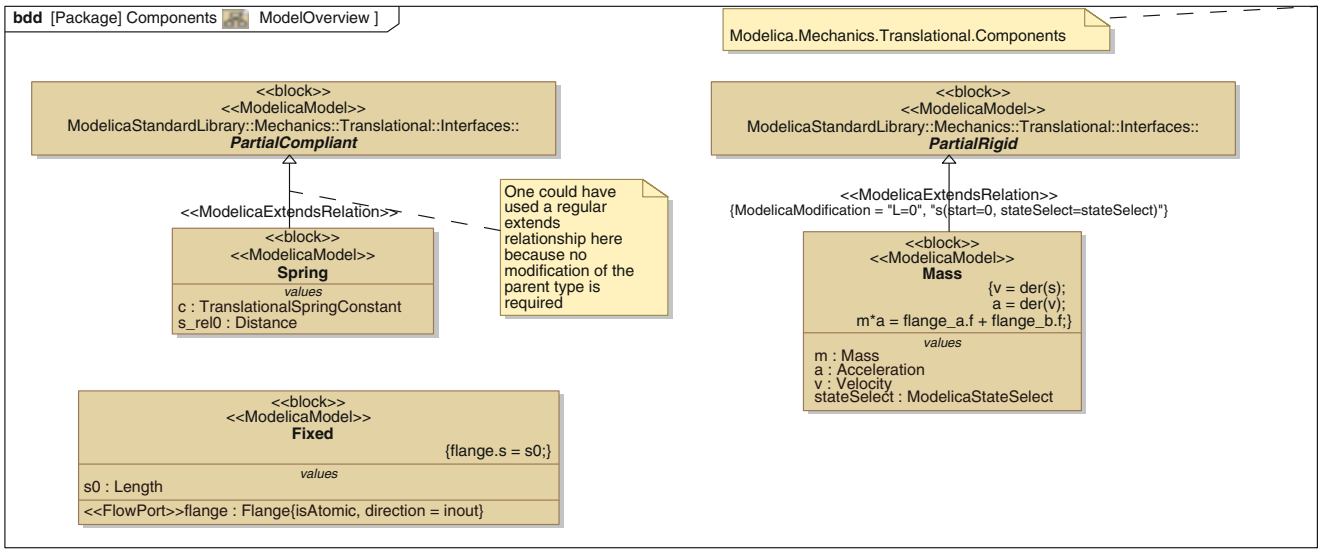


Fig. 3 the model of Sliding mass with inertia using Sysml.

e2eflow:end to end flow DispWorkstations.flow1->DataPortAccessConn1-
>DispProcess.disp_flow
->DataPortAccessConn2->Disp.flow2{
Latency => 100 Ms;
};
end ATSys.Impl;

For continuous data in big data environment, we can use one of following methods to model continuous data:

First, we can use the device component and property to model continuous data:

Device Physical entity-
Features
Continuous data variable Declaration;
Properties
Equation (
_____
);
End Physical entity;

Second, we can extend AADL data types using AADL Annex:

annex continuous data_specification { **
data variable Declaration *
equation *
** }

Third, we can use Modelica [12] to model continuous data, and then transform Modelica models to AADL Models:

Property Modelica_property is
Equation : aadltring applies to (device/system/memory/processor);
Const : aadltring applies to (device/system/memory/processor);
Const_value : aadltring applies to (device/system/memory/processor);
end Modelica_property;

Finally we can use SysML [13] to model continuous data, then transform SysML models to AADL Models. Fig. 3 represents the model of Sliding mass with inertia using Sysml[14].

Another important issue for big data driven cyber physical system design is that of handling temporal data and spatial data. We need extend AADL to include temporal information and spatial information as shown in Fig. 4.

3 Case Study: Big Data Driven Vehicular Cyber Physical Systems Design

Vehicle Ad-Hoc network is a special mobile ad-hoc network tag that loads on the vehicle 's electronic identification constructed by wireless technology. Vehicle Ad-Hoc network has the node characteristics, mobile features and data flow characteristics. The characteristics of the node manifests with a strong performance computing power, storage capacity, and almost no limited energy; Performance characteristics of its mobile network topology changes quickly, and the mobile node moves fast track predictable. Its data flow characteristics, the performance of real-time traffic information and communication load suddenly increases. Telematics and share information are collected through the car, car to car, car to the roadside infrastructure, automotive car to urban networks through interconnect, enabling more intelligent and safe driving. Fig. 5 [15] shows the system architecture of VANET from the perspective

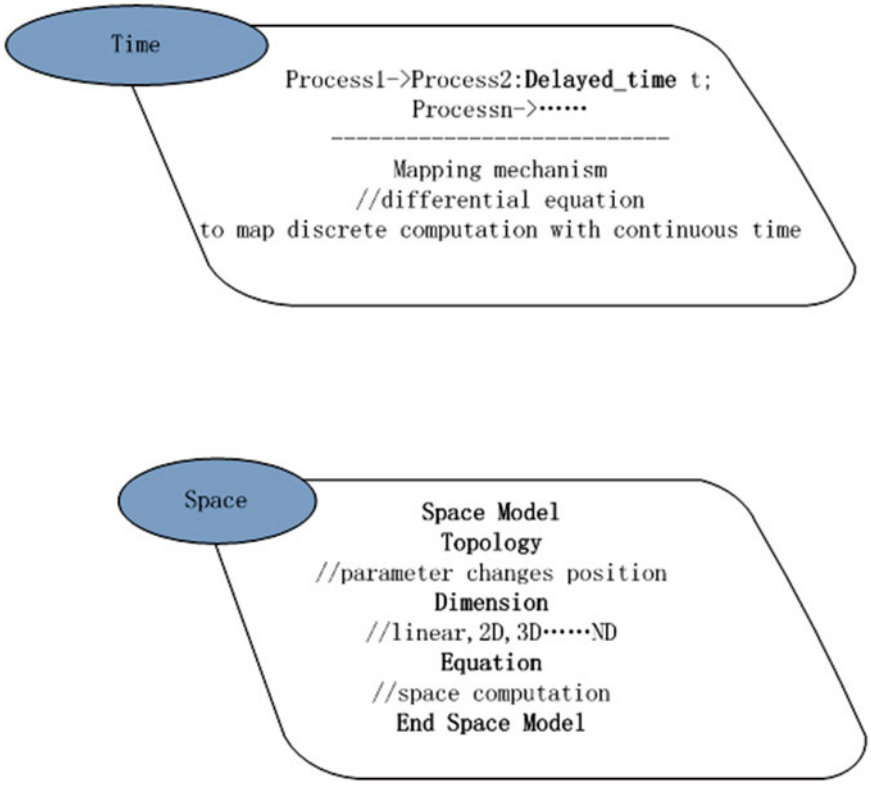


Fig. 4 temporal information and spatial information components in AADL

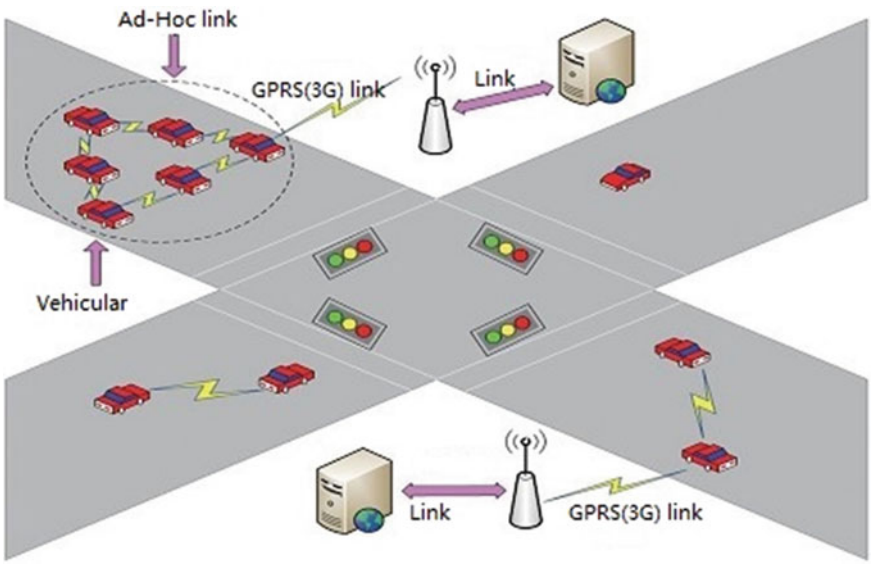


Fig. 5 The Vehicle Ad-Hoc network

of the different components and domains as well as their interactions.

Using big data manage traffic is a change in traffic management mode, while it also change the entire meaning of the management of the public transport market, hindering traffic through the bottleneck of traditional big data solutions.The big data can eliminate the limit across administrative areas, while promoting various different administrative regions from autonomous administrative region.The virtual of big data in vehicular Ad-Hoc network, enables its information management across the region, big data has the advantage of information integration and combination of efficiency. Big data helps establish a comprehensive three-dimensional traffic information system, the user may use a variety of traffic data into the system, to build an integrated public transport information use patterns, play overall traffic function, by integrating large data retrieval, using and analysis to extract relevant information to meet a variety of

transportation needs in order to solve the real-time traffic barriers.

The AADL file organization of VANET [16] is shown in Fig. 6.

Fig. 7 represents AADL Model of Road Station.

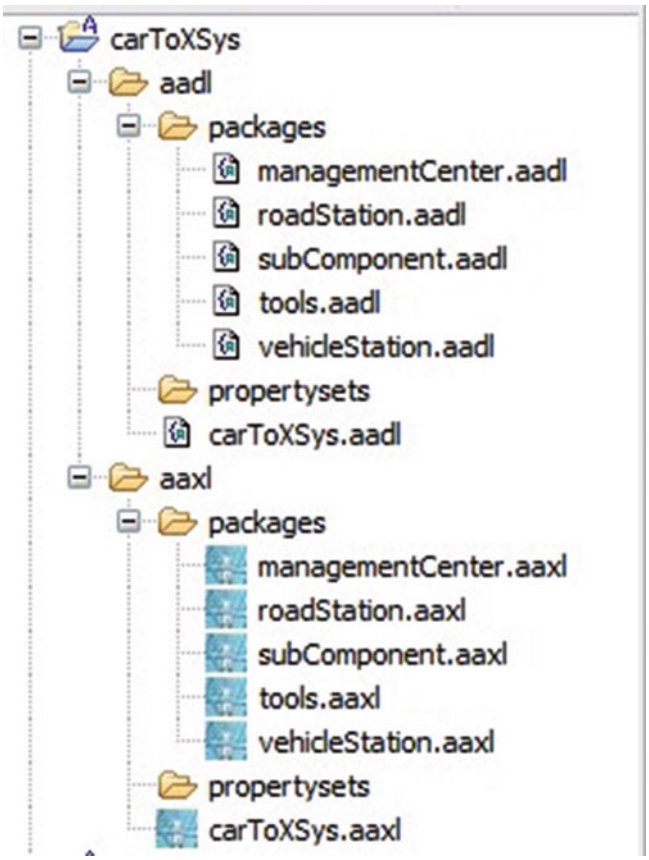


Fig. 6 Entire model of VANET in AADL

AADl data model of road station is as follows:

```
system road_station
    features
        vehicle_input : port group receptor_data_plug1;
        center_input :port group receptor_data_plug2;
        vehicle_out:port group receptor_data1;
        center_out : port group receptor_data2;
        light_out:port group light_data_plug; -
        display_out :port group display_data_plug;
end road_station;
system implementation road_station.impl
subcomponents
        the_light:device light;
        the_display :device display;
        the_vehi:device receptor1;
        the_Cen:device receptor2;
        the_tran1:device trans1;
        the_tran2:device trans2;
        PV:process process_vehicle;
        PC:process process_center;
        PL:process process_light;
        PD:process process_display;
connections
    V2P: port group the_vehi.out_Data -> PV.inData;
    P2V: port group PV.outdata ->the_tran1.lt_Data;
    P2C: port group PC.outdata ->the_tran2.lt_Data;
    C2P: port group the_Cen.out_Data ->PC.inData;
    PL2L: port group PL.outdata ->the_light.in_Data;
    PD2D: port group PD.outdata ->the_display.in_Data;
end road_station.impl;
```

The traffic light is model as follow:

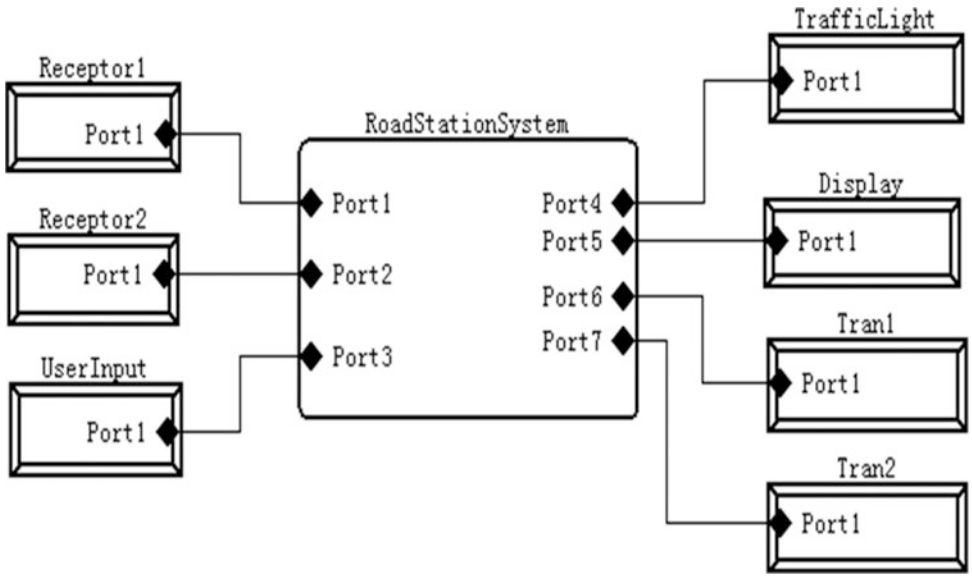


Fig. 7 AADL Model of Road Station

```
Device traffic_light
    Features
        In_data :in data port;
    Properties
            Equation => {
            Xt=I+L
            F=m*a=W*a/g
            A=μ*g
            T1=(V0-O)/a
            T=T1+X1/V0
            Const =>{ len, G, v, g', leni };
            Const_value =>{L,W,v0,g,I};
        Var=>{x1,f,a,t1,t};
        };
End traffic_light;
```

The flow specification of road station is as follow:

```
system implementation road_station.impl
subcomponents
    the_light:device light;
    the_display :device display;
    the_vehi:device receptor1;
    the_Cen:device receptor2;
    the_tran1:device trans1;
    the_tran2:device trans2;
    PV:process process_vehicle;
    PC:process process_center;
    PL:process process_light;
    PD:process process_display;
connections
    V2P: port group the_vehi.out_Data -> PV.inData;
    P2V: port group PV.outdata ->the_tran1.lt_Data;
    P2C: port group PC.outdata ->the_tran2.lt_Data;
    C2P: port group the_Cen.out_Data ->PC.inData;
    PL2L: port group PL.outdata ->the_light.in_Data;
    PD2D: port group PD.outdata >the_display.
in_Data;
flows -
  f_e2e: end to end flow the_vehi.Flow_R2C2R -> V2P-
>PV.FS1 -> P2V
  -> the_tran1.Flow_R{
  Latency => 100 Ms;
  };
end road_station.impl;
```

4 Conclusion

In this paper, we propose a big data driven cyber physical system design method based on AADL, which can specify and model the requirements of big data driven cyber physical systems, implement these requirements on big data platforms, guarantees QoS in a highly scalable manner, we present how to use AADL the extension mechanisms to model big data, especial spatial information and temporal information, we propose a scheduling algorithm to meet big data driven cyber physical system requirements.

In future work, we extend MapReduce in real time aspects to of enable the scheduling of mixed hard and soft real-time MapReduce applications, we extend our method to model big data driven cyber physical systems on cloud platforms and SOA.

Acknowledgement This work is supported by the national natural science foundation of China under grant (No.61370082, No.61173046), natural science foundation of Guangdong province under grant (No.S2011010004905). This work is also supported by Shanghai Knowledge Service Platform Project (No.ZF1213)

References

1. Abhishek B. Sharma, Franjo Ivanˇci´c, Alexandru Niculescu-Mizil, Haifeng Chen, Guofei Jiang. Modeling and Analytics for Cyber-Physical Systems in the Age of Big Data. www.sigmetrics.org/.../bigdataanalytics/.../bdaw2013_submission_8.pdf
2. SAE AS5506. Architecture Analysis & Design Language (AADL) v1. SAE Aerospace Standard, 2004.
3. Big Data: the Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute. 5,2011
4. Hudak J J, Feiler P H. Developing aadl models for control systems: A practitioner's guide[J]. 2007.
5. P. H. Feiler, D. P. Gluch, J. J. Hudak. The Architecture Analysis & Design Language (AADL): An Introduction. CMU/SEI-2006-TN-001, Carnegie Mellon University & Software Engineering Institute, 2006.
6. Peter H. Feiler, Bruce Lewis, The SAE Avionics Architecture Description Language (AADL) Standard: A Basis for Model-Based Architecture-Driven Embedded Systems Engineering, Software Engineering Institute, 2003
7. SAE/AS2-C. Data Modeling Annex document for the Architecture Analysis & Design Language v2.0 (AS5506A), October 2009.
8. Huawei Zhou. Design and Implementation of ATS Based on CBTC [D]. Chengdu China:Southwest Jiaotong University.(2010)
9. Chan-Ho Cho,Dong-Hyuk Choi,Zhong-Hua Quan,Sun-Ah Choi, Gie-Soo Park and Myung-Seon Ryou. Modeling of CBTC Carborne ATO Functions using SCADE[J].11th International Conference on Control, Automation and Systems.(2011).1089-1094.
10. Feiler, P.H.; J. Hansson, "Flow Latency Analysis with the Architecture Analysis and Design Language (AADL),"Technical Note CMU/SEI-2007-TN-010, Software Engineering Institute, 2007
11. Naeem Muhammad, Yves Vandewoude, Yolande Berbers, Sjir van Loo. Modelling Composite End-to-End Flows with AADL. adams-project.org/. . ./Modelling%20Composite%20End-to- End% 20Flows%20with%20AADL.pdf
12. Modelica - a unified object-oriented language for physical systems modelling. Language specification. Technical report, Modelica Association, 2002.
13. Johnson, T. A., C. J. J. Paredis and R. M. Burkhart (2008). "Integrating Models and Simulations of Continuous Dynamics into SysML." *6th International Modelica Conference*, Bielefeld, Germany, March 3-4, Modelica Association, 135-145.

14. SysML-Modelica Integration, Preliminary Draft, 2009-01-29 www.omgwiki.org/OMGSysML/.../fetch.php?...sysml-modelica%3Asysml...
15. Yousefi, S.;, Mousavi, M.S. and Fathy, M. Vehicular Ad Hoc Networks (VANETs): Challenges and Perspectives, Proceedings of 6th International Conference on ITS Telecommunications, 2006, p761 – 766
16. Hagen Stubing, Adam Opel Gmbh Marc Bechler.simTD: A Car-to-X System Architecture for Field Operational Tests[J]. IEEE Communications Magazine. 2010,48(5):148-154.

Simulating Active Interference Cancellation in Cognitive Radio

Zenon Chaczko, Grzegorz Borowik, and Philip Hsieh

1 Introduction

Due to the dramatic growth of the wireless communication systems in recent years, the most valuable resource is the radio electromagnetic spectrum. As most of the wireless systems coexists under the similar environment and uses the same bandwidth, the interference might occur among those systems.

An approach known as Cognitive Radio (CR) is considered as a most suitable methodology [3, 11, 8] for sensing coexisting systems and classifying the interference type, then negotiating between systems so the radio spectrum can be allocated more efficiently. It allows sharing the radio spectrum with high priority subscribers and not interfering with their authority. In 1999, Joseph Mitola has proposed the concept of CR based on his earlier work related to Software Defined Radio [14, 13].

CR successfully applies in wireless communication systems. It has the ability to detect the environment and surround wireless channels, and then automatically adjust sending and receiving parameters. The system repeats the steps of detecting, perceiving and self-sensing by retrieving the status of the use of spectrum.

In this paper, the Active Interference Cancellation (AIC) [9] with Direction of Arrival estimation is used to develop an efficient Multi-input Multi-output (MIMO) system [19, 20]. By utilizing MIMO, it allows the system to use one AIC antenna to create a protection for transmitting antenna and thus enables the data to be transmitted in the interference band. The approach can achieve the goal of decreasing improvidence and increase the efficiency of bandwidth use. The method of Direction of Arrival (DOA) estimation [1, 17, 21] seems to be able to help in evaluation the channel and enhance the MIMO Active Interference Cancelling model.

Z. Chaczko (✉) • P. Hsieh
Faculty of Engineering and IT, University of Technology, Sydney, Sydney, Australia
e-mail: Zenon.Chaczko@uts.edu.au; Tsang.Hsieh@uts.edu.au

G. Borowik
Warsaw University of Technology, Institute of Telecommunications, Warsaw, Poland
e-mail: G.Borowik@tele.pw.edu.pl

2 Significance

Federal Communications Commission (FCC) has allocated the channel bandwidth to a wide range radio spectrum for different services [5]. Those include analog television, AM/FM radio, mobile phones, Global Positioning System (GPS), Ultra-Wide Band (UWB) [12, 16, 10], and several other facilities. As more wireless services are emerging to the public, there is not that much radio frequency spectrum that can be still allocated. In fact, not all the channel is occupied for most of the bands. For instance, FCC has allocated the 7.5 GHz spectrum in the frequency band between 3.1 to 10.6 GHz for the users accessing UWB [23] or for unlicensed devices. In result, the spectrum is still not utilized economically [4].

It can be found there are still many frequency bands available for licensed or unlicensed transmissions. Most of the systems use frequencies between $0 \sim 2$ GHz. In the urban area ($3 \sim 4$ GHz), the spectrum usage efficiency is measured around 0.5%. For the frequency $4 \sim 5$ GHz the efficiency drops further to 0.3%.

On the one hand, more and more wireless systems coexist in the same space and the interferences between those systems occur. Especially, for wide-band wireless systems such as UWB that occupies the frequency range from 3.1 GHz $\sim$ 10.6 GHz. Although UWB was designed as a low power and for small usage scope to avoid jamming, the interferences still exist. On the other hand, FCC has published that at certain time of day or in some locations, approx. 70% of the allocated spectrum might be in the idle state.

Cognitive Radio technique [11, 8] may prove to be a suitable solution to resolve such issues. It could be a useful standard for future wireless communication systems that has

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_125,

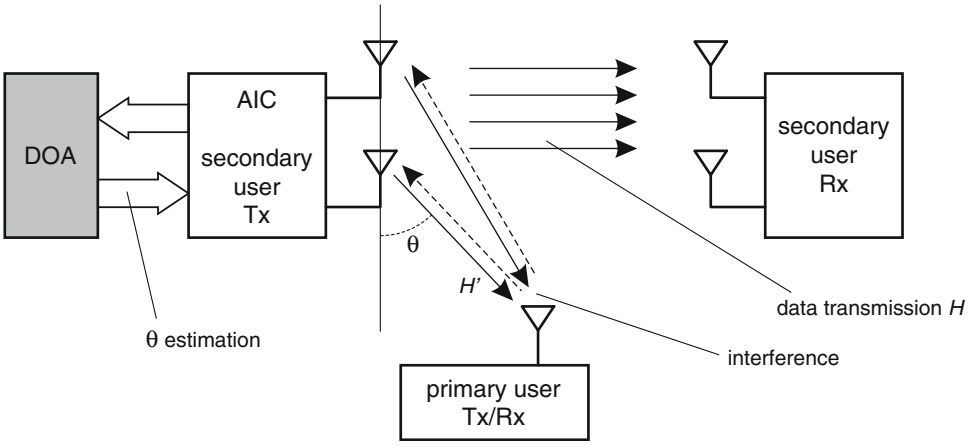


Fig. 1 Proposed MIMO AIC System model with DOA estimation

the ability to change the transmitting and receiving parameters to improve the efficiency of wireless communication and avoid interferences between different types of systems.

3 Methodology

A fundamental Active Interference Cancellation (AIC) [22] can be considered for Single-input Single-output OFDM system, where user can perform one type of transmission at a time only, i.e. either transmitting or receiving. While the system has primary and secondary users under the same environment, the interference can occur. Thus, exploiting one antenna only, the information can be easily distorted. However, by the application of the secondary user Tx, the interference can be avoided. Nevertheless, if the secondary user Tx is performing AIC to remove interference in SISO model [2, 24], data transmission is inefficient. Therefore, the Multiple-input Multiple-output AIC is proposed (Fig. 1). As shown, the interference can still affect both antennas at the secondary user Tx. However, if one antenna is used as AIC, the other antenna(s) can still transmit common data being not interfered by the primary user.

To make the MIMO AIC even more efficient, the Direction of Arrival (DOA) estimation is adopted on top of the MIMO AIC model. The reason to adopt DOA is that it can help to detect the incoming signal's direction, thus enhances the performance. In Fig. 1 the DOA component receives the data of the primary user, then it runs the algorithm to estimate the direction of the signal sent from primary user [17, 18].

3.1 Direction of arrival estimation

Direction of arrival (DOA) estimation is a method that can be used to determine the directions for both incoming and interference signal [21]. To perform the estimation the data received from the Rx antenna sensor array is used. DOA has been applied in many technologies, for example, radar or wireless communication: "Recent applications include array processing for wireless mobile communications at the base station for increasing the capacity and quality of the systems" [7].

DOA estimation algorithm is a popular approach for the research and development tasks [1, 21]. For example, the Beam-former method uses DOA by scanning the array antenna's main beam. Among the other techniques are: ESPRIT (Estimation of Signal Parameters via Rotational Invariance Technique) and MUSIC (Multiple Signal Classification). ESPRIT is parameter estimation technique that uses eigenvalues to decompose the array of input correlation matrix [15]. MUSIC is known of its super-resolution capability and simplicity, as well as, low computational complexity [7].

3.2 Application of DOA to MIMO AIC

The channel model of $\widehat{H}'$ (Fig.1) can be defined as:

$$\widehat{H}' = a \exp\left(\frac{-j2\pi f}{\lambda_c}\right)\left[\exp\left(\begin{matrix}1\\-j2\pi\Delta_r\widehat{\Omega}\end{matrix}\right)\right], \text{ where } \widehat{\Omega} \triangleq \cos\widehat{\theta}. \tag{1}$$

Assuming that the channel between the primary and the secondary user adheres to LOS (Line of Sight) model, the experiment should be carried out in the anechoic chamber. The model includes one antenna for the primary user and two antennas for the secondary user. Since the notch is supposed to occur at the primary user, the channel condition should be controlled, and the secondary user's Tx should have sufficient information about the channel.

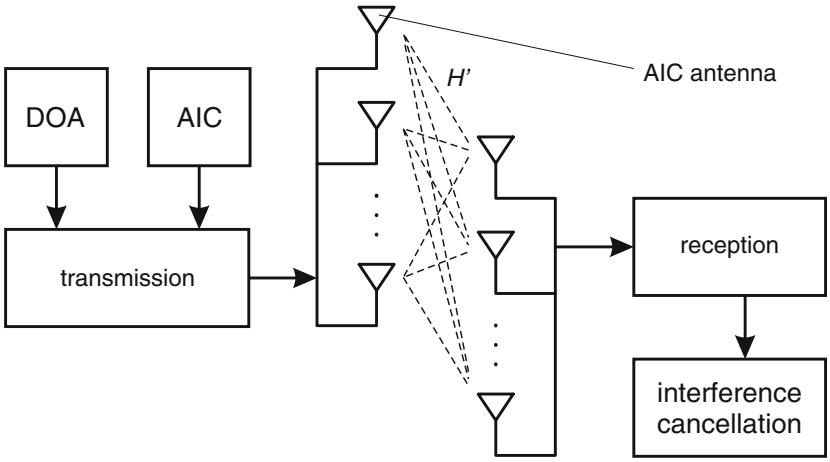


Fig. 2 System model for MIMO AIC

The estimated information $\widehat{H}'$ can be applied to the MIMO AIC system. Let define the signal received at the primary user as:

$$\widehat{Y} = \widehat{H}' \cdot X = \left[\widehat{H}'_1 \cdot \widehat{H}'_2\right] \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \left[\widehat{Y}_1 + \widehat{Y}_2\right] \tag{2}$$

where $\widehat{H}'$ is the estimated channel, X_1 is the signal transmitted by first Tx antenna of the secondary user, and X_2 is the signal transmitted by the second antenna. First antenna is used for AIC transmission, and the second to transmit the data. The transmitted data of the first antenna is:

$$X_1 = [X_1(0), \ldots, X_1(k-1), w, X_1(k+q), \ldots, X_1(127)]. \tag{3}$$

In victim band there are q number of tones being defined. The transmitted data of the second antenna is:

$$X_2 = [X_2(0), \ldots, X_2(k), X_2(k+1), \ldots, X_2(127)] \tag{4}$$

Signal Y is up-sampled with the rate r, which yields $\widehat{d}$. The interference is defined as d_1:

$$\widehat{d} = P \cdot (\widehat{Y}_1 + \widehat{Y}_2) \tag{5}$$

$$Y_1 = H'_1 \cdot X = [Y_1(0), \ldots, Y_1(k-1), 0, \ldots, 0, Y_1(k+q), \ldots, Y_1(127)] \tag{6}$$

$$d_1 = [d(k \cdot r), \ldots, d((k+q-1) \cdot r)] \tag{7}$$

Value d_1 is derived from d, and in order to cancel the interference, the Equation 8 needs to be fulfilled:

$$p_1 w_1 = -\widehat{d}_1 \tag{8}$$

P_1 is the kernel derived from P and w_y represents the AIC tones:

$$\widehat{w}_{y,op} t = -(P_1^H \cdot P_1)^{-1} P_1^H \cdot \widehat{d} \tag{9}$$

As the result, the transmitted data of the first antenna can be expressed as follows:

$$X_1 = [X_1(0), \ldots, X_1(k-1), \widehat{w}, X_1(k+q), \ldots, X_1(127)]^{\mathrm{T}} \tag{10}$$

Note that $\widehat{w} = [X_1(K), \ldots, X_1(k+q+1)]$ is equal to $\widehat{w}_y$, since the channel factor has been eliminated. In other words, firstly we obtain $\widehat{w}_y$ and secondly, by removing the channel effect, we obtain $\widehat{w}$.

The estimated channel allows evaluating the signal transmitted by the first antenna (AIC antenna). By applying the direction estimation method to the MIMO AIC system the notch is consequently developed. The function of AIC tones releases the interference between the primary and secondary user. The secondary transmiter and receiver is considered as a normal wireless system where AIC tones are generated in the transmitting data. Transmission of AIC tones may cause some noise, however, it is necessary for jamming cancelling. With MIMO model, the SNR could be improved since only one AIC tone is needed regardless of the existence of multiple antennas at the secondary user's transmitter (Fig. 1). In general case of channel effect (Fig. 2), one antenna is used to transmit data including AIC, whereas the rest of antennas transmit common data.

4 Simulation results

The main goal of the simulation was to check the impact of DOA estimation on MIMO AIC when integrating into the system.

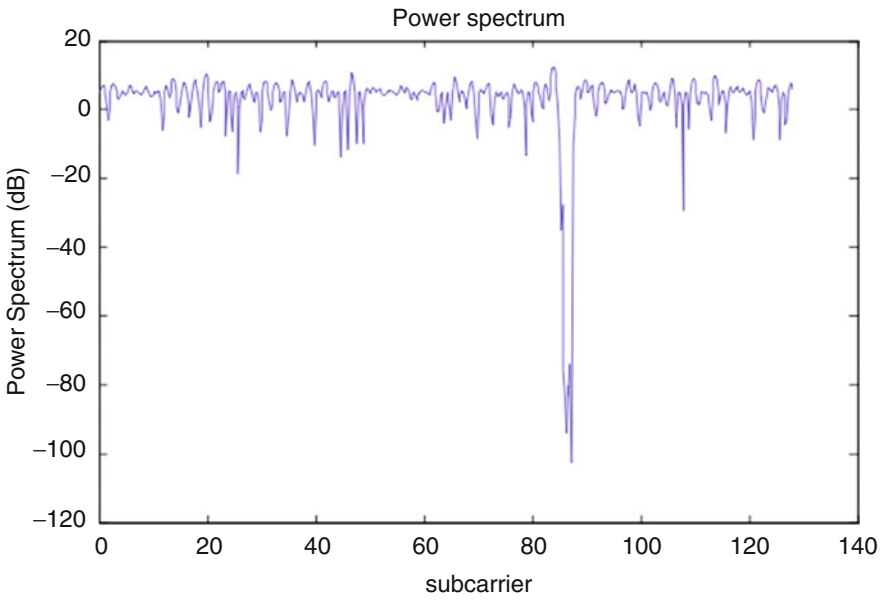


Fig. 3 Power spectrum of the transmitted signal

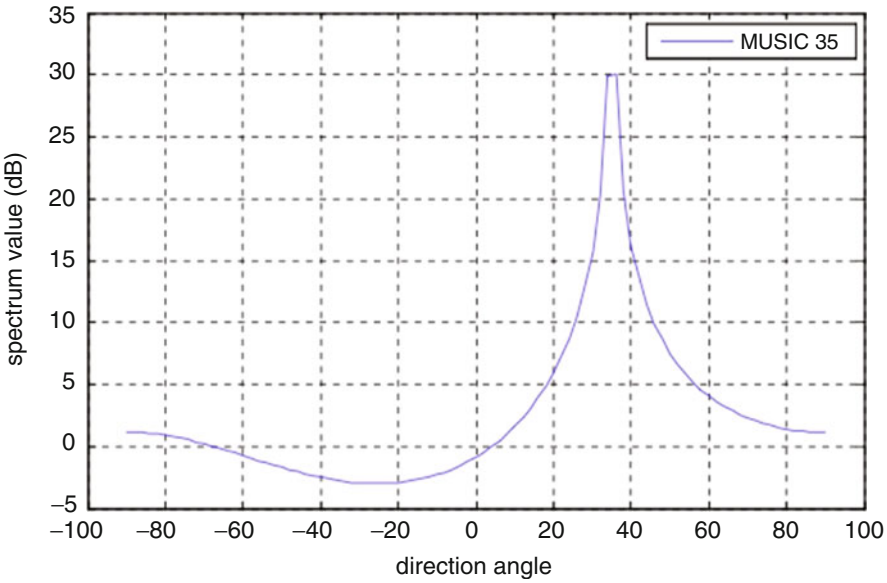


Fig. 4 Result of DOA estimation

First, the MIMO AIC simulation was considered in isolation. The setup contained two secondary user transmit antennas. The channel was flat fading and known at secondary user's Tx. The scheme provided in Section 3 has been applied to observe the power spectrum of one OFDM frame.

As shown in Fig. 3, the notch is around 80 dB. This simulation has been done for MIMO AIC without DOA estimation assuming that there were no channel effects. It means the secondary user's transmitter has a perfect knowledge about H'.

The second simulation has been done for DOA estimation integrated into MIMO AIC. The angle between the primary user and the secondary user was set at 35^0. The distance between users was greater than the adjacent antennas spacing (Fig. 1). From the figure, we can observe that the primary user send a signal to the secondary user first (dotted line). When the secondary user receives the information, the system performs the DOA estimation to detect the signal direction and the number of incoming signals. DOA was estimated as one coming channel and the direction angle was around 35^0 (Fig. 4).

The power spectrum for MIMO AIC with integrated DOA estimation that takes into account the channel effect is shown in Fig. 5. In comparison, the results of the experiment without channel effect (Fig. 3), are not as good. Noticeably, the notch is approx. 80 dB in the case without channel

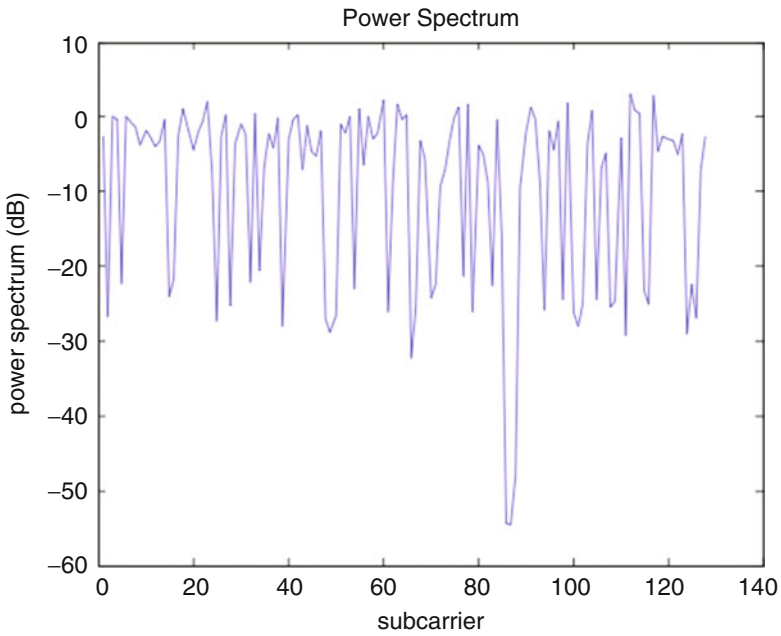


Fig. 5 Power spectrum of the transmitted signal

effect, and 50 dB with channel effect. However, the last – more complicated – scheme takes into consideration more factors and is closer to the real practical approaches. Moreover, the DOA estimation implemented in MIMO AIC helps to improve the radio performance yielding the result that is very close to the result of the perfect system.

5 Conclusions

This study introduces the concept of CR based AIC for UWB system. The simulation results indicate that AIC approach offers a better performance than the Turn-off Tones method. The proposed model of MIMO AIC system defines one antenna as an AIC antenna that covers all other antennas. The AIC antenna is used to perform interference cancellation thus other antennas can transmit and receive information data. For MIMO AIC system, the channel effect between the primary and secondary user represents an important vector of cancelling interference. In this work, the channel is represented as a LOS model. In such model, it is necessary to find out the angle of the incoming signals. The information retrieval is used when performing MIMO AIC. The sub-space techniques to estimate the direction of incoming signals are introduced as these are able to provide useful mechanisms that can be adopted into the proposed MIMO AIC method.

In this work, the proposed model utilises the LOS approach, however, in the real environment; only in some cases systems can have a clear sight of its target(s). This is due to the fact that between the Tx and Rx there could be other variances (i.e. physical or environmental) that cause the interference in the system.

This study covers only some selected issues related to the performance of the wireless communication, and how to make it more efficient and better utilized. There is still much work to be done in the future to improve the efficiency of interference cancellation.

References

1. Adve, R.: Direction of Arrival Estimation. Master's thesis, Toronto University, Canada (2003)
2. Alian, E., Saffar, H., Mitran, P.: Cross-Band Interference Reduction Trade-Offs in SISO and MISO OFDM-Based Cognitive Radios. IEEE Transactions on Wireless Communications 11(7), 2436–2445 (Jul 2012), DOI: 10.1109/TWC.2012.051512.110507
3. Brodersen, R.W., Wolisz, A., Cabric, D., Mishra, S.M., Willkomm, D.: Corvus: A Cognitive Radio Approach For Usage Of Virtual Unlicensed Spectrum. White Paper (2004)
4. FCC: Revision of Part 15 of the Commission's Rules Regarding Ultra-Wideband Transmission Systems (Apr 2002), ET Docket 98–153
5. FCC, NITA: United States Frequency Allocations – The Radio Spectrum (2003), http://www.ntia.doc.gov/files/ntia/publications/2003-allochrt.pdf, viewed 8 October 2012
6. Granelli, F., Zhang, H.: Cognitive ultra wide band radio: a research vision and its open challenges. In: 2nd International Workshop Networking with Ultra Wide Band and Workshop on Ultra Wide Band for Sensor Networks, 2005. Networking with UWB. pp. 55–59 (Jul 2005), DOI: 10.1109/NETUWB.2005.1470002
7. Harabi, F., Changuel, H., Gharsallah, A.: Estimation of 2-D Direction of Arrival with an Extended Correlation Matrix. In: 4th Workshop on Positioning, Navigation and Communication. WPNC '07. pp. 255–260 (Mar 2007), DOI: 10.1109/WPNC.2007.353642
8. Haykin, S.: Cognitive radio: brain-empowered wireless communications. IEEE Journal on Selected Areas in

Communications 23(2), 201–220 (Feb 2005), DOI: 10.1109/JSAC.2004.839380
9. Huang, S.G., Hwang, C.H.: Improvement of active interference cancellation: avoidance technique for OFDM cognitive radio. IEEE Transactions on Wireless Communications 8(12), 5928–5937 (Dec 2009), DOI: 10.1109/TWC.2009.12.081277
10. Lansford, J.: UWB coexistence and cognitive radio. In: International Workshop on Ultra Wideband Systems. Joint with Conference on Ultrawideband Systems and Technologies. Joint UWBST & IWUWBS. pp. 35–39 (May 2004), DOI: 10.1109/UWBST.2004.1320898
11. Lin, K.: Cognitive Radio – Application and Development. Chinese Taipei Electronic Components Certification Board Report 52, 31–38 (2008)
12. Liu, H.: Error performance of a pulse amplitude and position modulated ultra-wideband system over lognormal fading channels. IEEE Communications Letters 7(11), 531–533 (Nov 2003), DOI: 10.1109/LCOMM.2003.820079
13. Mitola, J., Maguire, G.Q., J.: Cognitive radio: making software radios more personal. IEEE Personal Communications 6(4), 13–18 (Aug 1999), DOI: 10.1109/98.788210
14. Mitola, J.: Cognitive Radio for Flexible Mobile Multimedia Communications. Mobile Networks and Applications 6(5), 435–441 (2001), DOI: 10.1023/A:1011426600077
15. Roy, R., Kailath, T.: ESPRIT-estimation of signal parameters via rotational invariance techniques. IEEE Transactions on Acoustics, Speech and Signal Processing 37(7), 984–995 (Jul 1989), DOI: 10.1109/29.32276
16. Saeed, R., Khatun, S., Ali, B., Abdullah, M.: An Adaptive UWB Waveform with Spectral Sharing Capability. In: Information and Communication Technologies, ICTTA '06. vol. 2, pp. 2309–2313 (2006), DOI: 10.1109/ICTTA.2006.1684766
17. Schmidt, R.: A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation. Ph.D. thesis, Stanford University, Stanford (Nov 1981)
18. Schmidt, R.: Multiple emitter location and signal parameter estimation. IEEE Transactions on Antennas and Propagation 34(3), 276–280 (Mar 1986), DOI: 10.1109/TAP.1986.1143830
19. Siriwongpairat, W., Su, W., Olfat, M., Liu, K.: Multiband-OFDM MIMO coding framework for UWB communication systems. IEEE Transactions on Signal Processing 54(1), 214–224 (Jan 2006), DOI: 10.1109/TSP.2005.861092
20. Tse, D., Viswanath, P.: MIMO I: spatial multiplexing and channel modeling. Fundamentals of Wireless Communication, pp. 290–331. Cambridge University Press (2005)
21. Wei, Z., Xiaoli, X.: Analysis and Simulation of the Direction of Arrival Estimation Algorithm of Spatial Signal. In: 8th International Conference on Electronic Measurement and Instruments. ICEMI '07. pp. 2–576–2–579 (Aug 2007), DOI: 10.1109/ICEMI.2007.4350746
22. Yamaguchi, H.: Active interference cancellation technique for MB-OFDM cognitive radio. In: 34th European Microwave Conference. vol. 2, pp. 1105–1108 (Oct 2004)
23. Yazdandoost, K., Kohno, R.: Ultra wideband antenna. IEEE Communications Magazine 42(6), S29–S32 (Jun 2004), DOI: 10.1109/MCOM.2004.1304230
24. Zhang, H., Zhou, X., Yazdandoost, K., Chlamtac, I.: Multiple signal waveforms adaptation in cognitive ultra-wideband radio evolution. IEEE Journal on Selected Areas in Communications 24 (4), 878–884 (Apr 2006), DOI: 10.1109/JSAC.2005.863876

A Development Study on Performance of a Real-Time Interface Device

Anıl Güçlü, Yağmur Atay, and Yasin Genç

1 Introduction

Hardware-in-the-Loop (HWIL) is the process of simulating engineering systems with realistic signals that is close to the real world. It is a combination of synthetic environment, models of subparts of the engineering system, software and engineering system [1, 2]. HWIL tests are common methods for designing, testing and evaluation of systems. In a HWIL test setup, whole/part of the physical system replaced with their numerical models and linked with the remaining system subparts such as sensors, controllers etc. via appropriate interfaces. The purpose of HWIL test is to combine the flexibility, cost-effectiveness and repeatability of numerical simulation with the accuracy and confidence of full hardware testing [3, 4]. It is a cost effective and efficient test method for developing engineering systems [5]. By means of HWIL tests, reliability and quality of the system increases. In addition, the system which will be tested is prevented from dangerous failures. It also reduces to development cycle of the engineering systems; therefore it is also a time effective method for development processes [6].

A HWIL system consists of at least three main subsystems which are Simulation Computer, Interface Devices (ID) and Unit Under Test (UUT). Simulation Computer includes the mathematical model of the whole/part of the engineering system, which to be tested, environment models etc. Appropriate signals are generated by the simulation computer according to predefined environment models for UUT. Responses according to the environmental model are created by the mathematical model of the subsystems. UUT is the whole/part of the engineering system which is a real hardware. ID is like a bridge between the two subsystems of the HWIL test. It receives calculated and simulated data from Simulation PC and transmits them to the UUT in the appropriate format. This action is also valid from UUT to Simulation PC.

In this study, it is aimed that to realize how the different algorithms, consisting of differing parameter types, affect the loop completion time and to increase the performance of the ID. Since there are lots of components in the HWIL test system, it is very important to know how much time passes on each component. Despite the processing speed of the development software tools, different algorithms to use those tools in ID are designed. Consequently, different tool combinations are developed. At the end, the effect of those combinations on ID performance and HWIL running time is measured. Then the best solution with highest performance is chosen as ID software.

The rest of this paper is organized as follows: Section 2 addresses system architecture. Section 3 presents the software architecture according to different combinations of the tools. Our analysis and results are presented in Section 4. Finally, conclusions and discussions are provided in Section 5.

2 Test System Hardware Architecture

Hardware-in-the-loop tests composed of three main subsystems such as; simulation computers, interface devices, and unit under tests (UUT). At first, overall system is modeled and simulated on the simulation computers in MATLAB environment. The simulation also contains the model of the UUT. If the UUT will be added to the test loop, the simulation PC must communicate with the UUT in real-time. In this case, all signals, which are sent to the UUT model in the simulation PC, has to be sent to the UUT hardware, and then the feedback data from the UUT has to be acquired and fed to the simulation as a hardware model output. During the HWIL tests, the simulation PC has to communicate with the UUT so some communication cards are needed to send and acquire messages from six

A. Güçlü (✉) • Y. Atay • Y. Genç
System Test and Evaluation Department, ROKETSAN Missile Inc., Ankara, Turkey
e-mail: aguclu@roketsan.com.tr; yatay@roketsan.com.tr; yagenc@roketsan.com.tr

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_126,

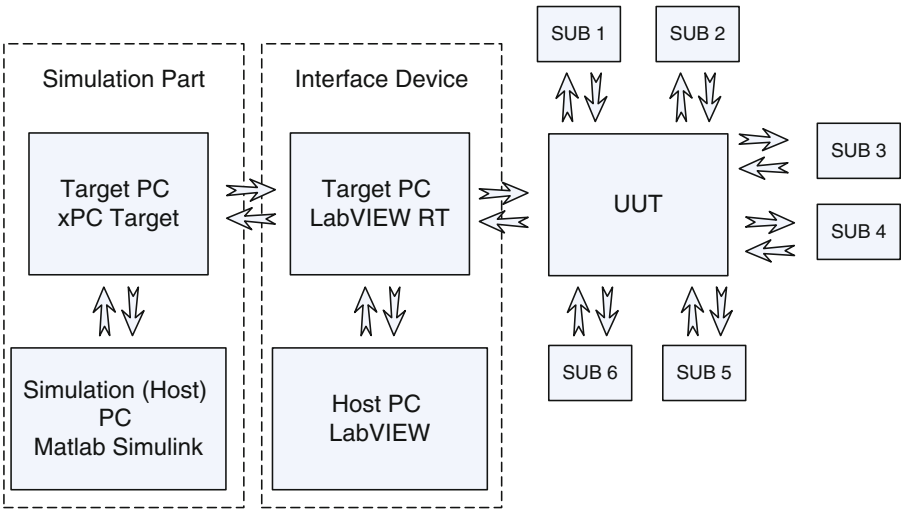


Fig. 1 Detailed System Architecture

subsystems. This communication can be hold in two different ways.

One of them is that integrating some communication cards in the simulation PC and connecting the UUT to the simulation PC directly. In software environment, all systems and subsystems are modeled in software environment including the messages and data flow among these models. However, if some hardware is included in the test, the messages have to be modified to a convenient form for the hardware. This causes extra process load to the simulation PC and may affect the test performance.

Another one is that adding an interface device between the simulation PC and the UUT. By means of the interface device (ID), messages are sent from the simulation PC to the UUT. Required modifications to the messages are done in the ID. After the messages are sent to the UUT, they are distributed to six different subsystems. Feedback data from the subsystems are collected by the UUT and sent to the ID.

The simulation PC loop completion time plays an important role during the tests. The performance of the simulation PC is affected negatively with using some communication cards on the simulation PC and modifying the simulation messages for the UUT. That's why, an ID between the simulation PC and the UUT is added. Data flow of overall system can be seen in Fig. 1.

2.1 Simulation PC

The simulation PC is composed of two computers named as Host and Target PCs. There is a master-slave relation between this two computer systems. Host PC works as the master by modifying the Target according to test scenarios and Target PC is used as slave by running convenient to Host PC decisions.

As a Host PC, a standard PC with a Windows Operating System (OS) is preferred. Modeling and offline operations are done in this PC. Matlab and Simulink are chosen as the numerical modeling and model development software platform in the Host PC.

Once the modeling is completed in the Host PC, it is compiled and deployed to the Target PC. It is again a standard industrial PC where the RT OS called xPC Target runs. This PC does the real time operations in HWIL tests. Data are transferred via VMIC interface which is General Electronics (GE) shared memory card. In our HWIL system, there is no direct connection between Target PC (xPC Target) and UUT. All data transmission among those systems is done via ID.

2.2 Interface Device (ID)

The ID is also composed of two computers, which are named as again a Host PC and a Target PCs. The ID software algorithm is developed in LabVIEW development environment. Then the developed software is compiled and deployed to target PC which has an operating system of LabVIEW RealTime. The target PC is a PXI chassis which is composed of some cards which are procured from the National Instruments (NI) Company. Data and messages are sent to the UUT by means of these cards. Communication interface between the ID and the UUT is RS485 and the data rate is 4Mbit/sec. Since the data rate is greater than 3Mbit/sec, FPGA based RS485 cards are used.

The HWIL test system is driven by the ID to synchronize the Simulation PC, the UUT, and subsystems. A graphical user interface (GUI) is designed on the host pc of the ID. After the all components are ready to test, a user has to start the simulation via Matlab-Simulink interface. Once the simulation starts, it runs just one time step. Some initial commands are generated with the first step run and written to a VMIC switch via target PC of the simulation part. Then the user starts the test via host PC of the ID. First, the ID

takes the data, which is written to the VMIC switch. The data is not convenient for the UUT so it has to be modified. After the modification process is completed, modified data are sent to the UUT via RS 485 communication interface. The UUT sends the data to the six subsystems and then takes the responses from them. After that, subsystem responses are acquired from the UUT via the ID. The acquired messages are written to the VMIC switch by the ID. Then, the ID sends a command (a VMIC interrupt) to run the simulation for one step. The target PC of the simulation part reads data from the VMIC switch and runs one step. Simulation outputs after one step run are written to the VMIC switch. These steps continue during the HWIL tests.

3 Test System Software Architecture

LabVIEW RT is a reliable stand-alone hardware control processor and Windows independent software 1. The design process of ID is focused on three main constraints as processing type, loop type, and programming method.

3.1 Processing Type

In HWIL tests, the data generated in simulation PC is not in the appropriate format for UUT. Therefore it needs some modifications which can be done at either in the ID or in the simulation PC.

3.2 Loop Type

In the development process of ID, it possible to use "for", "while" and "timed_while" loops. Additionally if "timed_while" structure is used, it is possible to appoint certain loops to certain processors of the ID.

Hence, "timed_loop" is used as processor "assigned timed_loop" or "unassigned timed_loop". As a result, four different loop structures is obtained.

3.3 Programming Method

Several sub-functions can be coded in LabVIEW software, according to use in ID, so the ID can call the sub-functions when it needs them. The other way is to make the programming of the ID as a whole function without using sub-functions. As a result, there are two alternatives to develop ID software as multi-layer or single layer. According to defined constraint, it is obtained sixteen different software designs which have the same functionality. Software combinations can be seen in Table 1 (SC: Simulation Computer, ID: Interface Device).

Table 1 Software Types

	Processing Type	Loop Type	Programming Method
1	On SC	While	Multi-Layer
2	On SC	For	Multi-Layer
3	On SC	Assigned Timed Loop	Multi-Layer
4	On SC	Unassigned Timed Loop	Multi-Layer
5	On SC	While	Single Layer
6	On SC	For	Single Layer
7	On SC	Assigned Timed Loop	Single Layer
8	On SC	Unassigned Timed Loop	Single Layer
9	On ID	While	Multi-Layer
10	On ID	For	Multi-Layer
11	On ID	Assigned Timed Loop	Multi-Layer
12	On ID	Unassigned Timed Loop	Multi-Layer
13	On ID	While	Single Layer
14	On ID	For	Single Layer
15	On ID	Assigned Timed Loop	Single Layer
16	On ID	Unassigned Timed Loop	Single Layer

Table 2 Loop times

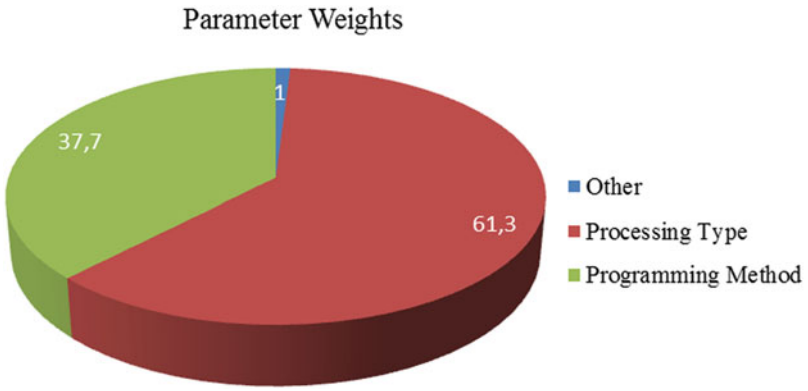
Combination Number	Loop Time (usec)
1	355
2	386
3	320
4	330
5	350
6	380
7	315
8	328
9	406
10	440
11	380
12	390
13	420
14	450
15	380
16	400

As stated before, the aim of this study is to increase the performance of the ID. At first, the ID software is developed "On ID, While, Multi-Layer" software type (combination 9). Loop time duration of the ID is measured with using the processor timer value since the operating system of the ID is real time and deterministic.

4 Tests and Results

Measured loop times according to the combination numbers are given in Table 2.

Fig. 2 Parameter weights



To analyze the effects of the parameters, which are processing type, loop type, and programming method, to the loop time in more detailed way, Minitab software is used. All the test parameters and loop time results are inserted to the software. Parameter weight distribution on the loop time is given in Fig. 2.

5 Discussions and Conclusions

At the beginning of the study, an ID code is developed with the combination number 9. After the tests are completed for the all combinations, it is obviously seen that, combination number 7, which consists of processing type on simulation computer, assigned timed loop, and single layer programming method, gives the best result among them. The time which passes on the ID device is shorten 23 %. There may be some other parameters which affect the time on the ID such as processor power, random-access-memory etc. However by using the same hardware for all the combinations, the effects of the hardware on the algorithms can be minimized, which help us to get generic results.

References

1. Mrad, F., Hassan, N., Mahmoud, S., Alawiech, B. and F. Adlouni, "Real-time Control of Free-Standing Cart-mounted Inverted Pendulum using LabVIEW RT", Industry Applications Conference, 8-12 Oct. 2000, pp. 1291–1298 (Vol2)
2. N. Syed, B. Mahesh, "An Overview of Hardware-In-the-Loop-Testing Systems at Visteon", SAE World Congress, March 8–11,2004
3. Monte MacDiarmid, Marko Bacic, "Quantifying the Accuracy of Hardware-in-the-Loop Simulations", Proceedings of the 2007 American Control Conference, July 11–13, 2007
4. B. A. James, "Using Hardware-in-the-loop (HWIL) Simulation to Provide Low Cost Testing of TMD IR Missile Systems", Part of the SPIE Conference on Technologies for Synthetic Environments: Hardware-in-the-Loop Testing III, Orlando, Florida, April 1988, pp. 432–440
5. Simon, D. Christian, B. Jean, "Real-Time PC-Based Simulator of Electric Systems and Drives", APEC 2002, pp. 433–438
6. Mihai Iacob, Gheorghe-Daniel Andreescu, "Real-Time Hardware-in-the-Loop Test Platform for Thermal Power Plant Control Systems", IEEE 9th International Symposium on Intelligent Systems and Informatics, September 8–10, 2011, Subotica, Serbia
7. LabVIEW RT manual
8. Matlab and Simulink Manual

Task Allocation within Mesh Networks: Influence of Architecture and Algorithms

Aleksandra Postawka and Iwona Poźniak Koszałka

1 Introduction

More and more computational power is needed to solve various problems. Supercomputers constantly advance in processing speed, but expensive CPU time may be wasted by inefficient Processor Allocator (PA). In order to increase the efficiency of task allocation process - many ideas, approaches, and methods in constructing task allocation algorithms are proposed in literature, including the algorithms based on First Fit [1–4] and its modifications based on Busy List [1, 2], or based on stack ideas so-called Stack Based Algorithm - family [1, 2, 5–8]. These algorithms are dedicated to the allocation process within 2D rectangle meshes. We focus not only on rectangle but also on the two other mesh processor structures (architectures) - cylinder and torus. For such scenarios a few algorithms are proposed in literature [3, 4, 9–11]. In the paper the own allocation algorithms (being some modifications of the known ideas) called Generalized Recognition Complete First Fit (GRCFF) and Generalized Stack-Based Allocation (GSBA) are proposed. The properties of these algorithms have been checked on the basis of the results of simulation experiments which were carried out with the designed own experimentation system. The efficiency of the algorithms was tested for three mesh structures consisting of the same number of the processors (nodes in mesh). To measure the algorithms' efficiency we introduced four indices of performance. The research was undertaken on task allocation algorithms with static task queue (the whole set of tasks is known before starting the allocation process). In such a case the tasks may be sorted without worrying about the starvation problem. Thus, the influence of sorting has been also taken into consideration.

The rest of the paper is organized as follows. The considered allocation problem is formulated in Section 2. In Section 3 the proposed algorithms for solving the problem are presented. Section 4 contains the description of the experimentation system and the results of investigations. The conclusions appear in Section 5.

2 Problem Formulation

Terminology. *Node* is a single *processor* and it is identified by (x, y) coordinates. Rows and columns are counted starting from the bottom-left corner - (0, 0) coordinate. A *mesh* M(W, H) is a composition of nodes. The three considered 2D structures of the mesh are defined in the following way:

- a rectangular mesh $M(W,H) = \{(x,y) : x,y \in \mathbb{Z} \wedge x \in [0,W) \wedge y \in [0,H)\}$;
- a cylinder mesh $M(W,H) = \{(x,y) : x,y \in \mathbb{Z} \wedge x \in (-W, W) \wedge y \in [0,H)\}$, where $x = (x+W) \bmod W$;
- a torus mesh $M(W,H) = \{(x,y) : x,y \in \mathbb{Z} \wedge x \in (-W,W) \wedge y \in (-H,H)\}$, where $x = (x+W) \bmod W \wedge y = (y+H) \bmod H$.

A $subgrid S_M(< i,j >, < w,h >) = \{(x,y) : x,y \in \mathbb{Z} \wedge x \in [i, i+w) \wedge y \in [j, j+h)\}$, where $i+w \in (0,W] \wedge j+h \in (0,H]$ is a rectangular grid which completely belongs to $M(W, H)$. A node $< i, j >$ indicates bottom-left corner and $< w, h >$ denotes width and height of the subgrid. A *free subgrid* is a subgrid, where all the processors are currently free and an *allocated subgrid* is a subgrid, where all the processors are currently allocated by the tasks. A rectangular *task* $J(w, h, t)$, where w denotes width, h denotes height, t denotes time needed to accomplish the task, can be allocated at the node-point (i, j),

A. Postawka (✉)
Faculty of Electronics Computer Architecture Group, Wroclaw University of Technology, Wroclaw, Poland
e-mail: aleksandra.postawka@pwr.wroc.pl

I.P. Koszałka
Faculty of Electronics, Department of Systems and Computer Networks, Wroclaw University of Technology, Wroclaw, Poland
e-mail: iwona.pozniak-koszalka@pwr.wroc.pl

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_127,

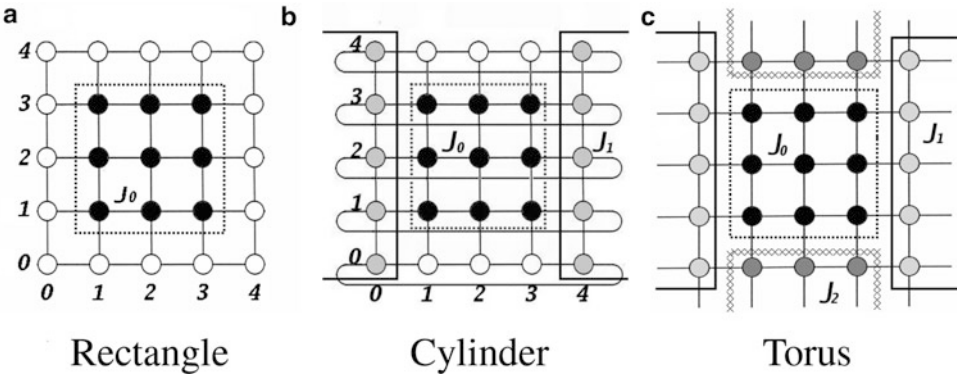


Fig. 1 Task allocation for different structures

causing that a submesh $S_M(< i, j >, < w, h >)$ is allocated. The task occupies a set of busy nodes completely until it is finished. Before allocation tasks are waiting in a queue. A *coverage* c_J of a busy subgrid with respect to waiting in queue task J is such a set of processors that allocation of J at any of these nodes would cause the task J to be overlapped with another already working task. The *coverage set* related to J is denoted by C_J and it is a set of the coverages of all busy subgrids. A *base subgrid* B_J for the task J is such a set of processors that J can be allocated at any of its nodes (without overlapping tasks already being executed). A *candidate area* CA_J for the task J is such a set of nodes which is considered to become the *base subgrid*. A *reject area* R_J for the task J is such a set of nodes that use of any of them as a *base subgrid* causes crossing the boundary of the mesh.

Example. Let us consider a queue of tasks $J_0(3, 3, 5)$, $J_1(2, 5, 7)$, $J_2(2, 3, 5)$. For the rectangle mesh, there may occur the situation shown in Fig. 1(a). Free nodes are located at the edges of the grid, but $J_1(2, 5, 7)$ cannot be allocated. For the cylinder structure, the possible allocation is shown in Fig. 1(b) - the next J_1 task can be allocated, but the third task in a queue not. A possible allocation for torus structure is shown in Fig. 1(c). The task $J_0(3, 3, 5)$ occupies the subgrid $S_M(< 1, 1 >, < 3, 3 >)$. The J_1 task is allocated on the subgrid $S_M(< 4, 0 >, < 2, 5 >)$. During the allocation the pointer for OY axis exceeds its maximum value of 4, thus it is changed to the value within the range by performing a modulo operation. The last task to allocate $J_2(2, 3, 5)$ can be placed on the submesh $S_M(< 1, 4 >, < 3, 2 >)$.

Problem statement. The considered allocation problem can be formulated as follows:

For given: (i) a set of tasks $J_0, J_1, J_2 \ldots, J_N$ with known sizes and execution times (w'_k, h'_k, t'_k), $k = 0,1 \ldots N$; (ii) a number of $(W \times H)$ nodes arranged in a way defined by the chosen structure $M(W, H)$.
To find: an allocation of tasks within the mesh.
Such that: the considered indices of performance reach the best value.
Subject to the constraint: the tasks should not overlap each other.

The indices of performance. To evaluate the quality of the algorithms the following indices of performance are introduced:

Allocation time t_a is the time needed to find a free subgrid and allocate the given task J

$$t_a = t_s - t_q,$$

where t_s - the time of successful allocation, t_q - the time of taking the task J from the queue.

Processing time t_p is the total time needed to execute all the given tasks.

External fragmentation F_e is the ratio of the number of free processors (N_f) to the total number of processors in the system (N_a):

$$F_e = \frac{N_f}{N_a} \cdot 100\%.$$

Latency L is the number of mesh time ticks between adding the task to the queue (for the static queue $t_0 = 0$) and its successful allocation (t_s):

$$L = t_s - t_0 = t_s.$$

3 Algorithms

Generalized Recognition Complete First Fit (GRCFF). The algorithm GRCFF is based on FF algorithm, but it is recognition complete in that both node by node verification and two job orientations are taken into consideration [1, 9]. The algorithm works in the following way: firstly C_J and R_J for the new task J are prepared. Next, all the nodes are sequentially searched in order to find the first one which does not belong to C_J or R_J (thus it can be used for J allocation). If this operation fails and J's width is different than its height, then J is rotated and algorithm attempts to allocate it again. If allocation is unsuccessful again, the algorithm waits (until one of the tasks being executed will

Table 1 GRCFF algorithm

GRCFF

```
set nodes from reject area as busy
rotated = false
do
    for each task w from the working tasks vector w_J
      calculate bottom-left coordinates for c_J (x_c = x_w − h_J + 1, y_c
= y_w − w_J + 1)
      calculate width and hight for c_J (w_c = w_J + w_w − 1, h_c = h_J
+h_w − 1)
      for x_c ≤ i < x_c + w_c
        for y_c ≤ j < y_c + h_c
          x = (i + W)modW
          y = (j + H)modH
          set node (x, y) as busy
    for 0 ≤ i < W
      for 0 ≤ j < H
        if node (i, j) is not busy
          allocate task J in subgrid S_M(< i,j >< w_J, h_J >)
          exit(success)
    if rotated
      exit(fail)
    else
      rotate task J
      rotated = true
while rotated
```

be completed and new free space on mesh occurs). The modification refers to wrapping coverage areas and defining R_J. The abbreviated description of GRCFF algorithm is given in Table 1. In order to explain how the algorithm works - let us consider the task $J(w, h, t)$ and $M(W, H)$ for different architectures. In the case of rectangle, to allocate J, there are the coordinates $0 \leqslant i \leqslant W - w$ and $0 \leqslant j \leqslant H - h$ considered (not belonging to R_J). The R_J example for task $J(1, 2, 5)$ (Fig. 2(a)) and rectangle mesh is shown in Fig. 2 (b). In the case of cylinder, tasks can be allocated at one of the edges as well. The tasks are wrapped and placed on the opposite edge, thus R_J is calculated only for OY axis (Fig. 2 (c)). In order to allocate J within cylinder mesh, nodes with the coordinates $0 \leqslant i \leqslant W - 1$ and $0 \leqslant j \leqslant H - h$ are considered. For torus there are no borders and the first task can be placed at any node. The special feature of torus is the lack of R_J (Fig. 2(d)). To allocate J the nodes with coordinates $0 \leqslant i \leqslant W - 1$ and $0 \leqslant j \leqslant H - 1$ are considered. Despite the fact that in torus there is no R_J, additional c_J's have to be considered, which are inessential in the rectangle mesh (Fig. 2(e) and 2(f)). Not only the tasks are wrapped around the edges, but C_J as well.

Generalized Stack-Based Allocation (GSBA). This algorithm uses elementary areas subtraction operation. In the first part of its performance, it follows SBA in action (which is recognition complete [1]). Firstly, C_J for the current task J is calculated. Together with CA_J (at the beginning it is the entire mesh without R_J) it is placed on the top of the stack. The c_J's are subtracted from CA_J's and the results (new CA_J's with reduced C_J's) are placed on the stack again. The algorithm works until the stack is empty or CA_J without c_J's to subtract is found. If the task J cannot be allocated, then it is rotated and the algorithm attempts to allocate it again. If allocation fails, algorithm waits. GSBA is the alteration of Improved Stack Based Algorithm (ISBA) [1]. The modifications refer to (i) C_J and initial CA_J determination, (ii) area subtraction operation and (iii) test for areas overlapping. GSBA in action is shown in Table 2. The tasks can be allocated on the borders as well, so they are wrapped around two edges of the grid. It causes that C_J wraps together with them. Wrapping C_J causes that some coordinates are negative (see definition of S_M), thus it is complicated to determine how the two areas overlap. In the worst case CA_J inside the mesh can be overlapped with c_J in four places (see Fig. 3(a)). The problem was solved by changing the coordinate system. The bottom-left corner of CA_J is moved to the point (0,0) as in Fig. 3(b) and the coordinates are recalculated properly for c_J too. The attention has to be paid to the condition of wrapped task's coordinates signs. Since CA_J is anchored in the center of coordinate system it is easy to subtract areas once overlapped (now it can be done in the same way as for rectangle). If the x_c coordinate of c_J is negative and it fulfills the condition $x_c + W < w_c$ (where w_c is the CA_J's width), it is clear that for OX axis the areas are overlapped twice. The same should be checked for the OY axis.

4 Investigations

Experimentation system. The system consists of the following elements:

Controlled input: A - allocation algorithm {GRCFF, GSBA}; S - queue sorting {unsorted, time descending, time ascending, task size descending, task size ascending}.

Problem's parameters: P_1 - the number of tasks in the queue; P_2 - the range of values from which the task's dimensions will be chosen by random; P_3 - the range of execution time for each task in the queue; P_4 - the size ($W \times H$) of structured mesh.

Outputs: t_a-allocation time; t_p-processing time; F_e-external fragmentation; L-latency.

Experimentation system was written in C++ and compiled by *g++ 4.1.2* with *-O3* optimization flag enabled. Simulations were performed on Intel(R) Xeon(R) 2.67 GHz processor. This research was supported in part by PL-Grid Infrastructure. For each simulation (single experiment) the tasks are randomly generated due to given parameters and inserted into the queue. The queue can be sorted in ascending or descending order by such properties as task execution time or task dimension. After all the tasks are executed, the calculated statistics are returned.

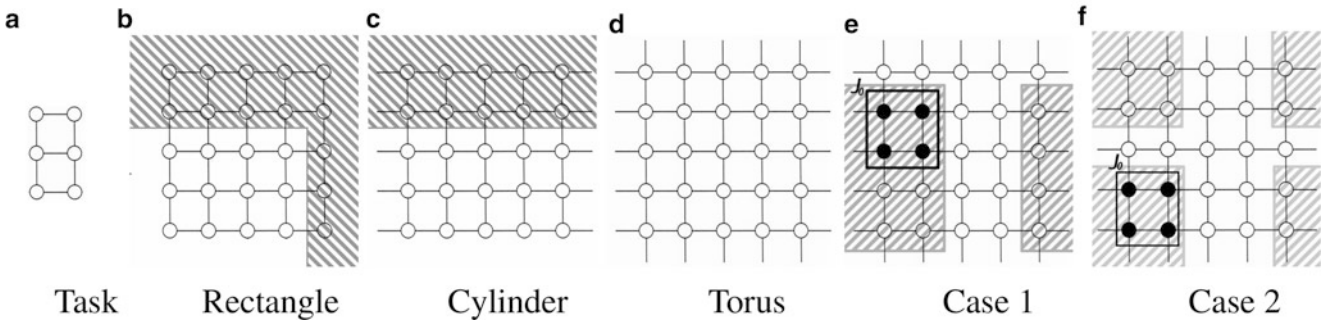


Fig. 2 Reject areas (b, c, d) for different architectures and torus coverage (e, f)

Table 2 GSBA algorithm

GSBA

```
rotated = false
do
   create initial candidate area CA_J
   for each task w from the working tasks vector w_J
      calculate bottom-left coordinates for c_J (x_c = x_w − h_J + 1, y_c =
y_w − w_J + 1)
      calculate width and hight for c_J (w_c = w_J + w_w − 1, h_c = h_J + h_w
−1)
      add c_J to vector vec_cJ
   push onto the stack pair (CA_J, vec_cJ)
   while stack is not empty
      pop from the stack pair (CA_J, vec_cJ)
      for each c_J in vec_cJ
         remove c_J from vec_cJ
         if c_J overlaps CA_J
            vec_CA = (area subtraction) CA_J − c_J
            for each new CA_J in vec_CA =
               push onto the stack pair (CA_J, vec_cJ)
      if vec_cJ is empty
         allocate task J in bottom-left node of CA_J
         exit(success)
      if rotated
         exit(fail)
      else
         rotate task J
         rotated = true
while rotated
```

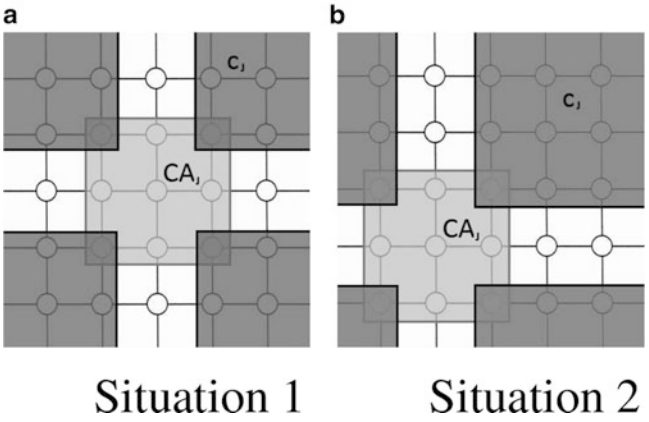


Fig. 3 Changing the coordinate system for torus architecture

Experiment 1. Influence of used architecture on average allocation time. *Experiment design:* $W = H = X$ and $P_1 = X^2$ P_2: $1 \div \lceil 0.4 \cdot X \rceil$

P_3: $1 \div 1000[ticks]$ P_4: X^2 for $X \in \{$ 100, 200, 300, 400, 500, 600, 700, 800$\}$

Results: Figures 4(a), 4(b) show the dependency of average t_a and grid size for each of the three investigated mesh architectures, for GRCFF and GSBA algorithms. GRCFF algorithm is $O(W \cdot H)$ what can be noticed in Fig. 4(a). Cylinder and torus architectures are more complex than rectangular one, but t_a are not significantly greater (for the same dimensions), e.g., for grid size of (800 × 800) t_a for cylinder and torus architecture is greater respectively by 17% and 27% than for rectangle structure. GSBA has almost the same t_a for all grid dimensions (Fig. 4(b)). When the number of tasks and their sizes increase proportionally to the growth of network size, t_a for GSBA is constant. The complexity of mesh architecture has an impact on t_a, but for each structure the dependencies are very similar to others. In line with expectations the average time needed to allocate the task in the cylinder architecture is longer than in the rectangular mesh and for torus allocation takes the most time. Removing each of the borders makes the subtraction operation more complex, because more special cases have to be taken into consideration. It can be observed (see Fig. 4(b)) that t_a in torus is nearly twice longer than in rectangular mesh. For example, for the biggest grid sizes t_a for cylinder and torus is greater by 36% and 91% than for rectangle architecture, respectively.

Experiment 2. Influence of tasks dimensions on the processing time. *Experiment design:* $P_1 = 10000$ P_2: $(1 \div 10), (10 \div 25), (15 \div 30), (20 \div 35), (25 \div 40), (30 \div 45)$ P_3: $1 \div 1000[ticks]$ P_4: 100×100

Results: It can be observed (Fig. 4(c), 4(d)) that for almost all investigated cases torus architecture achieves the shortest t_p. Moreover, for both GRCFF and GSBA it can be noticed that the greater tasks surface area, the more profitable is using more complex architectures - e.g., in the case of GSBA (Fig. 4(d)) and task sizes in the ranges $(10 \div 25)$, $(20 \div 35)$, $(30 \div 45)$ profit from the use of the torus instead

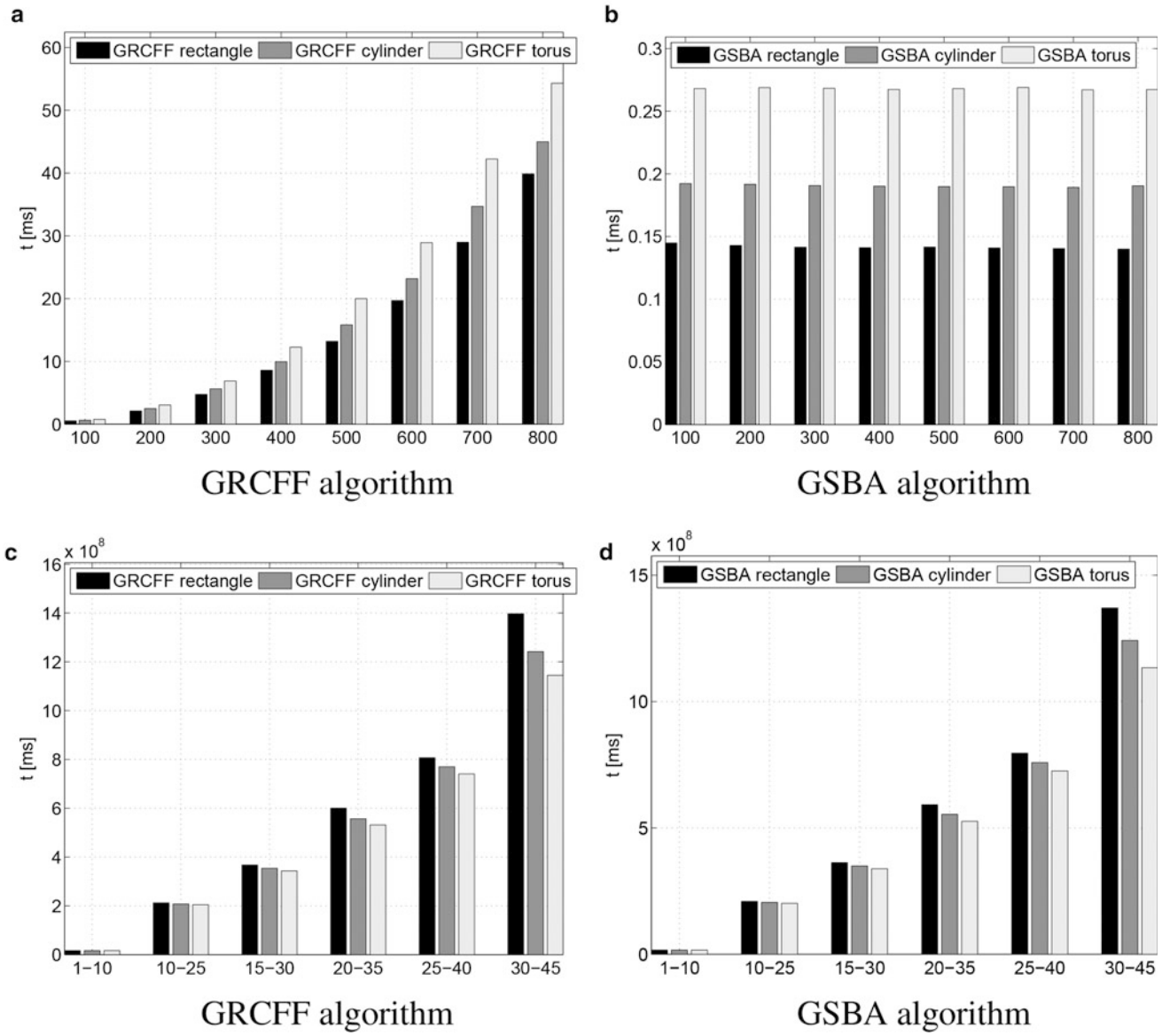


Fig. 4 Average tasks allocation time for different structures and different grid's side length (a, b) The dependency of processing time for tasks with different sizes - grid size 100 × 100 (c, d)

of a rectangular architecture is: 3.7%, 11.1% and 17.3%, respectively. For smaller tasks it is not so obvious and sometimes other architectures return shorter t_p than torus, e.g., for tasks from the range $1 \div 10$ the algorithm GRCFF achieved the shortest t_p for cylinder and the longest for rectangular structure - but when the algorithm GSBA is used (Fig. 4(c)), statistics are inverted. However, the differences are so small, that could be negligible (0.01%-0.2%).

Experiment 3. Examination of mesh architecture and sorting queue. *Experiment design:* P_1: 10000 P_2: $(5 \div 15) \times (5 \div 15)$, $(5 \div 30) \times (5 \div 30)$, $(15 \div 30) \times (15 \div 30)$ P_3: $1 \div 1000[ticks]$ P_4: 50x50

Results: For each mesh architecture and for each value of P_2 parameter five different simulations have been performed. Each of them used another sorting. The data set is either completely unsorted or sorted by task execution time and task surface area in ascending or descending order. In case 'unsorted' tasks are inserted into the queue directly from the random task generator. The obtained results are presented in Fig. 5. The characteristics of the results for F_e and t_p are very similar because of close relation between these parameters - the greater average number of unused nodes, the more time will take the execution of the same tasks. Therefore only diagrams for F_e have been presented (Fig. 5(a) and Fig. 5(b)). Tasks with the smallest

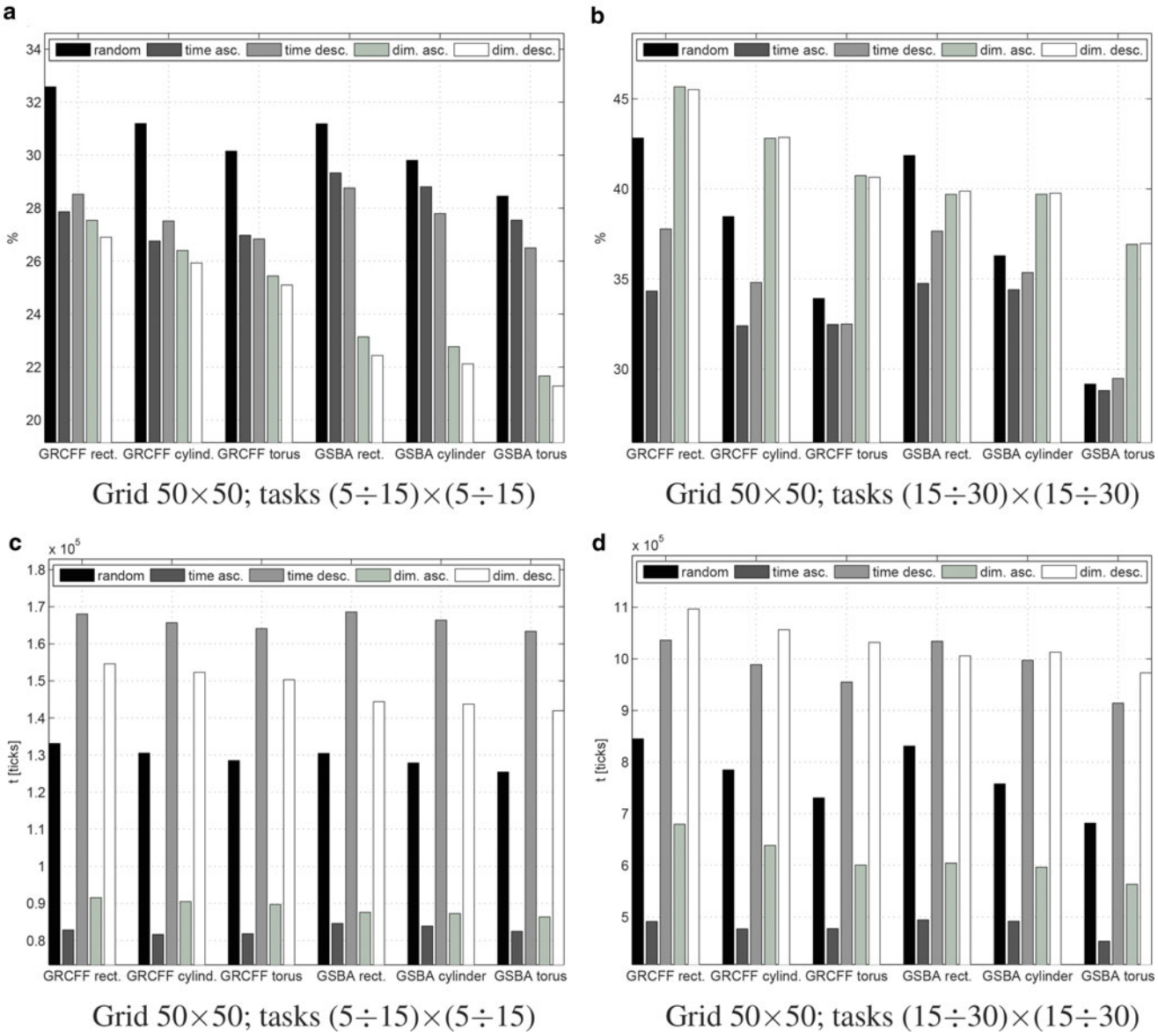


Fig. 5 External fragmentation (a, b) and latency (c, d)

surface area $(5 \div 15) \times (5 \div 15)$ are characterized by the least F_e (Fig. 5(a)). It is quite obvious because smaller tasks can fit better. As expected, torus shaped mesh usually achieves the best results, because it occurs that tasks, which would not be allocated in rectangular mesh, fit when wrapped. Looking at figures 5(a) and 5(b) it can be easily observed that sorting tasks can both improve and worsen achieved results. Sorting tasks by execution time in ascending order gives less F_e than for the random queue. Moreover, it can be noticed that for GRCFF algorithm it is better to sort tasks by time in ascending order, while for GSBA sometimes better results are obtained when sorting descending (Fig. 5 (a)). However, the greater tasks surface area, the more profitable it becomes for GSBA to sort ascending. For example, in the case of grid size (50 × 50) and torus architecture, F_e for sorting ascending in comparison to sorting descending (see Fig. 5) is higher by 1.1% for task sizes (5 ÷ 15), lower by 0.1‰ for task sizes (5 ÷ 30) and lower by 6.7‰ for task sizes (15 ÷ 30) - see Fig. 5(b). Sorting tasks by their dimensions often makes the F_e greater and consequently gives longer t_p (Fig. 5(b)). In the case of sorting ascending at the beginning of such simulation tasks fit very well, but later there are only the tasks with the biggest surface area and the gaps between them cannot be filled. In the case of sorting descending large tasks at the beginning make F_e greater. In some special cases sorting tasks queue by dimensions helps, but if there is time available for sorting tasks it is better to sort them by execution time.

Completely different conclusions can be drawn from the observation of average L factor. The best results are

achieved when tasks are sorted in ascending order, either by time or by surface area. In the first case, the average L for GSBA and tasks from $(5 \div 15)$ for torus in Fig. 5(c) is lower by 34%, and in the second case it is lower by 31% (in comparison to 'unsorted' queue). It is quite obvious that allocating the shortest tasks first and leaving the longest tasks at the end shortens the average waiting time. The same effect is gained when smaller tasks, which fit better, are placed earlier. Average is counted for many tasks with short latency and few tasks with greater value.

Sorting tasks by investigated parameters in descending order is never a good choice, because it extends the task's waiting time for its execution. In such cases the L value is greater by 30% for sorting by time and by 13% for sorting by tasks dimensions (considering GSBA in Fig. 5(c) and comparing with unsorted queue).

5 Conclusions

On the basis of the experiments it may be concluded that better performance was obtained for GSBA algorithm than in case of GRCFF. Although the allocation time was longer, better results were gained for processing time, external fragmentation and latency. As expected, in comparison to other investigated mesh architectures, in most cases torus structured mesh ensured achieving the best results.

Torus architecture achieves the best performance for the tasks with greater surface area. Because of wrapping tasks on the edges it occurs that there can be allocated task, which would not fit in other structures. Although there is no reject area in torus, more coverage areas have to be taken into consideration (these wrapped on the edges). This is the reason why the performance does not increase much. In the case of smaller tasks the processing time for torus architecture is not always shorter than for other structures, thus such network is more profitable for tasks requiring more processing nodes.

The usage of sorting algorithms can significantly improve the performance of the allocation algorithms. The values for processing time and external fragmentation are the least when tasks in the queue are sorted by time, in any order. In case of latency the lowest values are obtained for sorting by time or by task's surface area, both in ascending order. The average latency is very important because it indicates how long average system's user will have to wait for data. But the processing time is also significant. The only operation which improves all the indices of performance is sorting the task queue by time in ascending order. One of the advantages of GSBA algorithm is quite constant allocation time for different grid sizes. GRCFF algorithm can be considered only for small grids, because allocation time increases radically.

In the nearer future, the authors plan to expand the experimentation system following ideas proposed in [6] and to implement other allocation algorithms presented in [2].

Acknowledgement This work was supported by the statutory funds of the Faculty of Electronics, Wroclaw University of Technology, Wroclaw, Poland.

References

1. Zydek, D., Selvaraj, H.: Fast and efficient processor allocation algorithm for torus-based chip multiprocessors. Computers and Electrical Engineering 37, 91–105 (2011)
2. Majkowska, A., Zydek, D., Koszałka, L.: Task Allocation in Distributed Mesh-Connected Machine Learning System: Simplified Busy List Algorithm with Q-Learning Based Queuing. Advances in Intelligent Systems and Computing 226, Springer, 763–772 (2013)
3. Jakimczuk, Ł., Kmiecik, W., Poźniak-Koszałka, I.: Task Allocation Algorithms for 2D Torus Architecture. The Twelfth IARIA International Conference on Networks, 169–174 (2013)
4. Zhu, Y.: Efficient Processor Allocation Strategies for Mesh-Connected Parallel Computers. Journal of Parallel and Distributed Computing 16, 328337 (1992)
5. Yoo, B.S., Das, Ch.R.: A fast and efficient processor allocation scheme for mesh-connected multicomputers. IEEE Transactions on Computers 51, 46–60 (2002)
6. Koszałka, L., Kubiak, M., Poźniak-Koszałka, I.: Comparison of SBA - Family Task Allocation Algorithms for Mesh Structured Networks. LNCS, Springer 4331, 21–30 (2006)
7. Poźniak-Koszałka, I., Proma, W., Koszałka, L., Pol, M., Kasprzak, A.: Task Allocation in Mesh Structure: 2Side LeapFrog Algorithm and Q-learning Based Algorithm, LNCS, Springer 7336, 576–587 (2012)
8. Kaminski, R., Koszałka, L., Poźniak-Koszałka, I., Kasprzak, A.: Evaluation and Comparison of Task Allocation Algorithms for Mesh Networks. Proc. of 9th ICN, IEEE CPS, 104–108 (2010)
9. Geunmo, K., Hyunsoo, Y.: On Submesh Allocation for Mesh Multicomputers: A Best-Fit Allocation and a Virtual Submesh Allocation for Faulty Meshes. IEEE Transactions on Parallel and Distributed Systems 9, 1127–185 (1998)
10. Borowiec G., Postawka A., Koszaka L.: Static task allocation algorithms and influence of architecture on Mesh Structured Networks. Computer Systems Engineering PBW, 66–80 (2014)
11. Dally, W.J.: Performance analysis of k-ary n-cube interconnection networks. IEEE Transactions on Computers 39, 7127–785 (1990)

An Overview of Chip Multi-Processors Simulators Technology

Malik Al-Manasia and Zenon Chaczko

1 Background

A Computer System Architecture (CSA) simulator is a software based tool, that is useful in modelling various components and devices of the computer to forecast the output and performance level given a particular input. Computer simulation is very often the driving force in computer architecture researches and designs. This is due to the enormous work in the design of compound microarchitecture, which is rarely modeled in an analytical manner making the process of prototyping and all design stages excessively expensive. A comprehensive simulation is equally expensive as it needs well crafted simulators, realistic workloads and numerous machine cycles.

A mandatory level of details is achieved at a slow pace, because even the high-speed simulators may require a period of many weeks or months to complete a task. This becomes even worse when multi-core processor designs become dominant. The total throughput of the host chip increases with increasing generations; however, the single-thread performance, which is a fundamental element in simulation, does not show any improvement because of limitations on the power. It is also necessary to note that a target-architecture model gradually becomes complex since there has to an increasing number of cores. The increased performance of the host cores does not adapt to the target increasing complexity.

Researchers are aware of the problem by exploring various strategies of reducing simulation time. Some earlier works have applied the sampling method [1, 2] to get a statistical representation of a given part of the application in order to provide similar insights in an entire simulation, but in a short period. Other researchers [3–5] suggest conducting the parallel simulation in order to decrease a single-simulation period through operating it on numerous threads. This method is suitable for design phases, which needs a faster evaluation of one or more design points. However, if many parameters are considered, running multithreaded simulations simultaneously gives better simulation performance[6].

The strategies of reducing simulation period are necessary for cycle-accurate simulators [7, 8]. These simulators provide the fundamental elements of research in computer architecture to enhance understanding of the stresses relating to workload from the micro architectural structures. However, their time consuming characteristic is a limitation to their in the exploration of chips consisting of tens or more cores. Generally, simulators that are cycle accurate are not suitable for large-scale architectures[6].

2 CSA Simulators Importance

Simulations are vital and understanding of their role makes it possible to carry our various complex activities. Simulation can be important in a system migration as the old systems, which are expensive, can be simulated to keep them operational. Therefore, simulators can assist the organization in migrating to the new hardware and software easier. This allows for significant savings to business, in terms of finances and time. Simulation is also important in hardware development. According to Magnusson et. al, simulators can be useful in modelling computer systems regardless of their availability [9]. This makes the process of hardware development simpler since the hardware could be in the development process or still non-existent. In addition, it limit indirect costs as it reduces the need for costly hardware prototypes [10]. Simulation can be useful in saving costs. This is through simulating the costly hardware using a cheaper one. For instance, the standard x86-based PC, saving the organization, costs associated with expensive hardware. It is also useful in software development especially for

M. Al-Manasia (✉) • Z. Chaczko
Faculty of Engineering, University of Technology Sydney, Broadway, NSW 2007, Australia
e-mail: malik.a.al-Manasia@student.uts.edu.au; zenon@eng.uts.edu.au

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_128,

developers of operating systems and device drivers. Programs operating in kernel mode can be very difficult to debug; however, simulators provide non-intrusive debugging options that make the development of such programs simpler. Simulators also function to provide security since they run in a controlled environment, which enhances security. This allows companies to be able to try the new systems before they are used in the process of production. This is important in boosting the company's overall security system. Simulators are useful in restoring previous states of simulated computers.

3 Taxonomy of CSA Simulators

Classifications of computer architecture simulators are categorized basing on their contexts. To start with, they are classified according to their **scope**, which include Microarchitecture, Full System, Application-based, Instruction Set and Cycle Accurate simulators. Microarchitecture simulation technique is useful in modeling the design and performance of microprocessors and its elements. On the other hand, full system simulators model features as privileged modes and imitate the function of peripheral components in order to provide support to the operating system. This makes it possible for them to run complicated multithreaded workloads. Simics [11], gem5 [12], RSIM [13] and SimOS [14] use this method. Application- based or user level simulators emphasizes on the user part of applications. Their development and use are much easier as they do not require device-timing models, large disk images and booting of the operating system. However, they are only useful in supporting primary workloads. Graphite [4], Sniper [15], ZSim [16], CMP$im [17] and McSimA+ [18] use this method.

The other classification is based on ***input*** or ***workload***, which include trace-driven and execution-driven simulators. Trace-driven simulators make use of sessions, which were previously recorded when executing simulated applications. It involves recording of operations of memory and other important instructions when running an application in a trace generation environment [19]. This type of environment can be the real hardware, which is the target. However, it can also result from software generations in case the hardware is under development or non-existent. It is necessary to note that irrespective of simplification of execution, trace-driven simulation lacks a vibrant behavior present in multi-threaded programs running on several processors simultaneously. This inconsistency is evident when running this type of simulation on another Symmetric Multi-Processor (SMP) system other than the one used in the trace [19, 20]. Another limitation of trace-driven execution is the time limitation and data storage necessary to record the session(s) [19]. Goldschmidt and Hennessy [19], in their report state that trace-driven simulation should not be used in the simulation of systems depending on timing or parallel system.

Execution-driven simulators implement applications basing on a simulated processor. This type of simulation does not require traces and can be conducted on one machine [19]. This makes the issues of instructions timing and concurrency to be avoided. Accuracy of the simulator is the only affection of correctness of the result from an execution-driven [19]. This method is highly valuable when optimizing a target as it avails all the data, which is in use or being produced by the target [10].

Finally, *detail* is also major category used in classifying simulators. It includes functional (what is done) and timing (when is it done) simulators. Functioning simulators focus on the achievement of the same function as those of the modeled components. Timing simulators, however, attempt to focus on the accurate reproduction of performance or timing characteristics of the respective targets.

4 CSA simulators quality attributes and evaluation parameters

There are different perspectives of balancing various factors in an optimal manner to enhance the functions of the simulator in terms of accuracy, flexibility, performance, degree of details and functionality. The availability of many simulators with specific designs for limited tasks based on definite aspects comes because maximization all of these aspects can be complicated. However, some simulators have a more general approach.

Performance is a term used in measuring the speed of the simulator in completing a specific workload. Its measurement is in terms of slowdowns, which is also known as the host number of instructions is executed for every simulated instruction in the target. The performance of processors is measured in MIPS (Millions of Instructions Per second) which is using in benchmarking simulators.

The other criterion is **functionality**, which is what the simulator can perform or be able to perform after the modification. A functional simulator can execute a good part of the software that operates on the target, and it will function properly without issues. Functionality also includes which ISA architectures the simulator support? How many and what type of workloads it can run? Can it run multithreaded applications on homogenous and heterogeneous multicore systems?

Accuracy defines how well a simulator repeats the target behavior. And it represents to which degree the simulated target performance is matching with the real system performance. Therefore, it is necessary to note that a more accurate simulator incurs more slowdowns; hence it will simulate a

Table 1 Simulators Categorized by Features. Abbreviations: (FS/A)-Full-System (FS) vs. Application-Level (A), (Free-AC) free Academic Institution, (UIUC/NCSA) Univ. of Illinios/NCSA OS License, (BSD) Berkeley-Style Open Source License, (SS)-Simulation Speed.

	Year, Availability	FS/A	LoD	Simulated Platform	SS	Extensibility	Special Features
Simics	1994, commercial/ free AC	FS/A	OoO	UltraSparc, PowerPC, MIPS,ARM	+	API provided	Simulate "everything", heterogeneous nets
gem5	2003, free BSD	FS/A	OoO	Alpha, ARM, SPARC, MIPS, POWER and x86	+	Modular /OO design, Python config. files	Written from scratch, pervasive OO, configurability
RSIM	1997, UIUC/NCSA	FS	OoO	SPARC V9/Solaris	+	Modular design	Simulate adv. CPUs in network
SimOS	1994, freely available	FS	Variable	MIPS/SGI IRIX, Alpha/UNIX	+	Uncertain	Simulate entire system
Graphite	2010, free	A	No	x86	++ +	Python config files	Parallel simulator
Sniper	2011, free	A	OoO*	x86	++ +	Python config files	Interval simulation
CMP$im	2008, free	A	No	x86	++	Python config files	Memory system simulator
ZSim	2013, free	A	OoO*	x86	++ +	Python config	Fast interpreter design
McSimA +	2013, free	A	OoO*	x86	++ +	Python config files	Middle ground between FS/A

high level of details. Absolute accuracy shows how close the simulation is to the real world whereas relative accuracy shows how correct a model is between diverse configuration settings.

Finally, **flexibility** is also used as a criterion of the simulator as it gives a description of whether the design of the simulator is versatile, it ability to operate from dissimilar hosts and ease of porting it to other hosts among others. It also evaluates the ability of a simulator to have extendable modular design. These entire requirements validate the flexibility of a simulator.

Numerous processor and system simulators are already available as shown in Table 1. All listed simulators have their own merits, Howver, a full-system simulator might be particularly useful when the simulation involves heavy I/O activities or extensive OS kernel function support. However, these simulators are relatively slow and make it difficult to isolate the impact of architectural changes from the interaction between hardware and software stacks. Moreover, because they rely on existing OSes, they usually do not support manycore simulations well.

A pure application-level simulation is insufficient, even if I/O activity and time/space sharing are not the main areas of focus. For example, thread scheduling in a manycore processor is important for both performance accuracy and research interests. Thus, it is desirable for application-level simulators to manage threads independently from the host OS and the real hardware on which the simulators run.

For example, Graphite [4] uses less detailed models, such as the one-IPC model, to achieve better simulation speed. Sniper [15] uses better abstraction methods such as interval-based simulation to gain more accuracy with less performance overhead. While these simulators are good for early stage design space explorations, they are not sufficiently accurate for detailed microarchitecture-level studies of manycore architectures. Graphite [4] and Sniper [15] are considered faster simulators because they use parallel simulation to improve the simulation speed. Trace-driven simulations can also be used to trade simulation accuracy for speed. However, these are not suitable for multithreaded applications because the real-time synchronization information is usually lost when using traces. Thus, execution-driven simulations (i.e., simulation through actual application execution) are preferred. On the other hand, full-system simulators model both microarchitecture-level details and OSes. Thus, they sacrifice simulation speed for accuracy. Instead, it is desirable to have a simulator to model manycore microarchitecture details while remaining faster than full-system simulators, which have both hardware and software overhead. Since, simulators targets simulation of thousand-core chips, they have to be a user-level simulator for now. No current main-stream OS scales to thousands of cores, and ISAs also limit the number of cores [16].

For example, x86's xAPIC only supports up to 256 cores. The simulators have varying characteristics, strengths and weaknesses, and relevance to different research areas. While most development projects are only a few years old, a simulator such as SimOS can be considered to be too old to be used actively in research now. CMPSim and ZSim have the problem that they are not available publicly, and would be hard to obtain. Some simulators can be integrated with useful timing simulators as – TFsim and GEMS – use a timing-first approach to simulation, where a less extensive model simulates the timing of the processor, which is then

later verified by a larger and more complex execution-drive simulator, such as Simics. All the simulators can simulate Chip Multi Processors (CMP) architectures, but only a few can simulate SMT.Industry simulators typically run at a speed of 1 to 10 kHz. Academic simulators, such as gem5 [12], GEMS [7] and PTL-Sim [21] are not truly cycle-accurate compared to real hardware, and therefore they are typically faster, with simulation speeds in the tens to hundreds of Kilo Simulated Instructions Per Second (KIPS). Cycle-accurate simulators face a number of challenges in the multi-core era. First, these simulators are usually single-threaded, and their performance does not increase with increasing core counts. Second, given its slow speed, simulating with large caches becomes increasingly challenging because a slow simulation speed does not allow for simulating huge dynamic instruction counts.

5 Case Study: Sniper

Sniper is a high-speed and accurate x86, parallel simulator. Sniper integrates interval simulation method and expands the general functionality of Graphite simulation infrastructure; permitting for fast and accurate simulation. Trading off simulation speed for accuracy allows Sniper to have a variety of flexible simulation options when exploring various homogeneous and heterogeneous multi-core architectures.

This study describes the architecture of Sniper Simulator, the key components, configuration of simulation, result tools and different visualization alternatives. To show how to use the sniper simulator practically, this study demonstrates how to start various test runs, outcomes and interpretation of recorded results. The tests of simulation depend on pre-tailored test binaries accompanying the simulator. The sniper simulator is well developed and has a manual for setup configurations. The instructions contained in the manual makes it easy for users to know where to start and how to use the device during operations. The sniper has documentation of reasonable quality in comparison to other similar solutions. When using the sniper simulator a person can do timing simulations for multi-program workloads and multi-threaded, shared memory applications with tens to hundred cores. The speed of performing this operation is higher than the speed of existing simulators. The central feature of the sniper simulator is the core model based on interval simulation, a fast mechanistic core model. Simulating at intervals allows a person to raise the level of abstraction in architectural simulation. This guarantees faster development of the simulator and evaluation times. For example, this happens it "jumps" between miss events called intervals. The sniper simulator is validated against Nehalem and multi-socket Intel Core2 systems thus there are average performance estimation errors within 25 percent at the simulation speed of up to many MIPS [15].

The Sniper simulator and interval core model is vital in the uncore and system level studies requiring many more details than the traditional one-IPC models and when the cycle-accurate simulators are too slow to enable reasonable simulation workloads [22]. Another benefit is that the interval core model supports the creation of Cycles Per Instruction (CPI) stacks. These stacks portray the number of cycles that get lost due to the varying characteristics of the system (i.e., the branch predictor or cache hierarchy). This enables to use the Sniper for characterization of applications and software/ hardware co-design.

5.1 Sniper Configuration

The configuration of the sniper simulator is achievable with command line parameters and configuration files. In practice, the default configuration exists at sniper/config/base.cfg. The script of run-sniper accepts the command line parameters and the evaluation of parameters or configuration files occurs from the left to the right with respect to the command line. Also, newer values override the older values.

5.2 Simulation Results Tools

The sniper simulator creates files at the end of a simulation process; the sim.cfg which has configuration alternatives that are in use. The sim.out file shows the primary statistics and the sim.stats[.sqlite3] has the following set of tool to capture key points of the simulation.**:**

- ***CPI Stacks.*** Interval simulation is unique because allows for the creation of CPI stacks that summarize the place where time is spent. The CPI stack can be described as a stacked bar that shows different components that contribute to the overall performance as shown in fig. 1. The base CPI exists at the bottom and shows the meaningful work that is done. The CPI stack is significant when gaining insight in the performance of the application.
- ***Power and Area Stacks.*** To estimate the power consumption of a program, the sniper simulator integrates with the McPAT power and modeling framework. When the files, sniper/tools/mcpat.py and area.py, are run in the directory containing the sniper output files, power and area stacks can be generated. With respect to this data, the files, area.png and power.png are generated together with the output texts that depict power consumption of the application, broken down by the component. There are options of choosing plot dynamic, static or total power of the chip is for every component (Fig. 2).

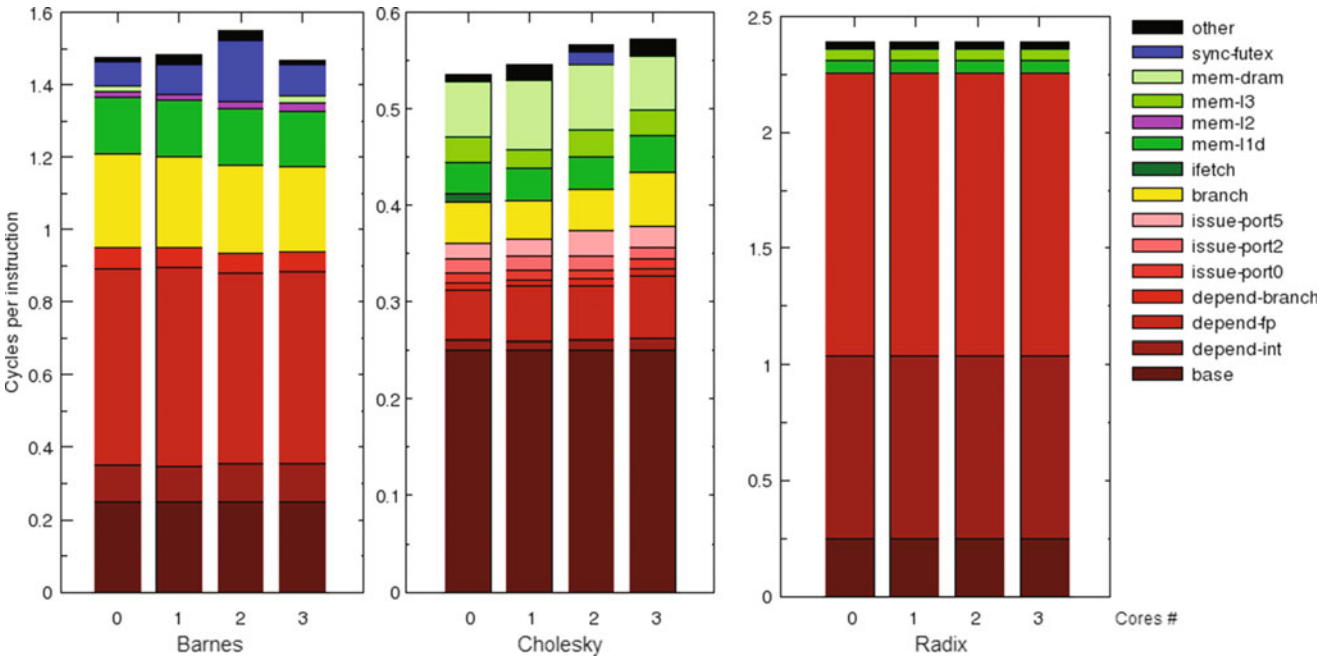


Fig. 1 Cycles Per Instruction stack

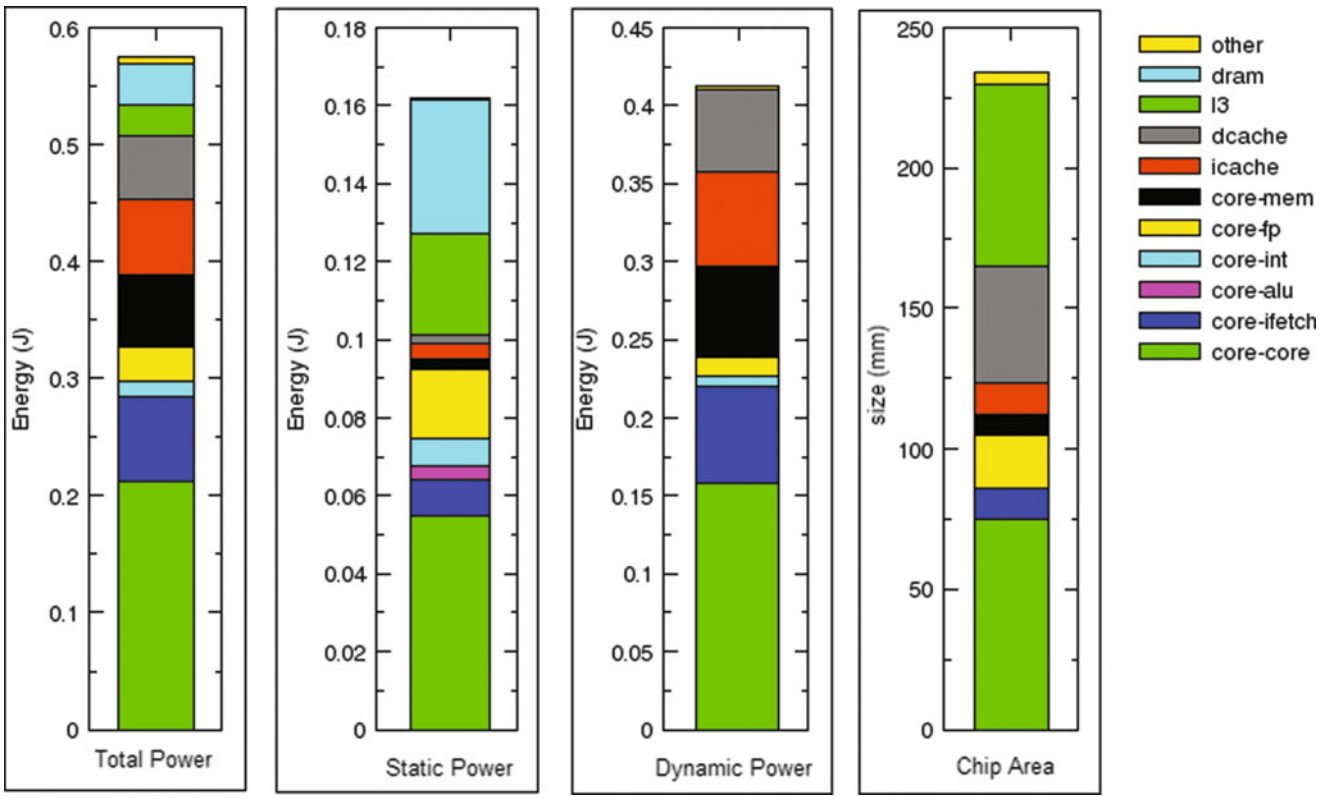


Fig. 2 Power and area stacks

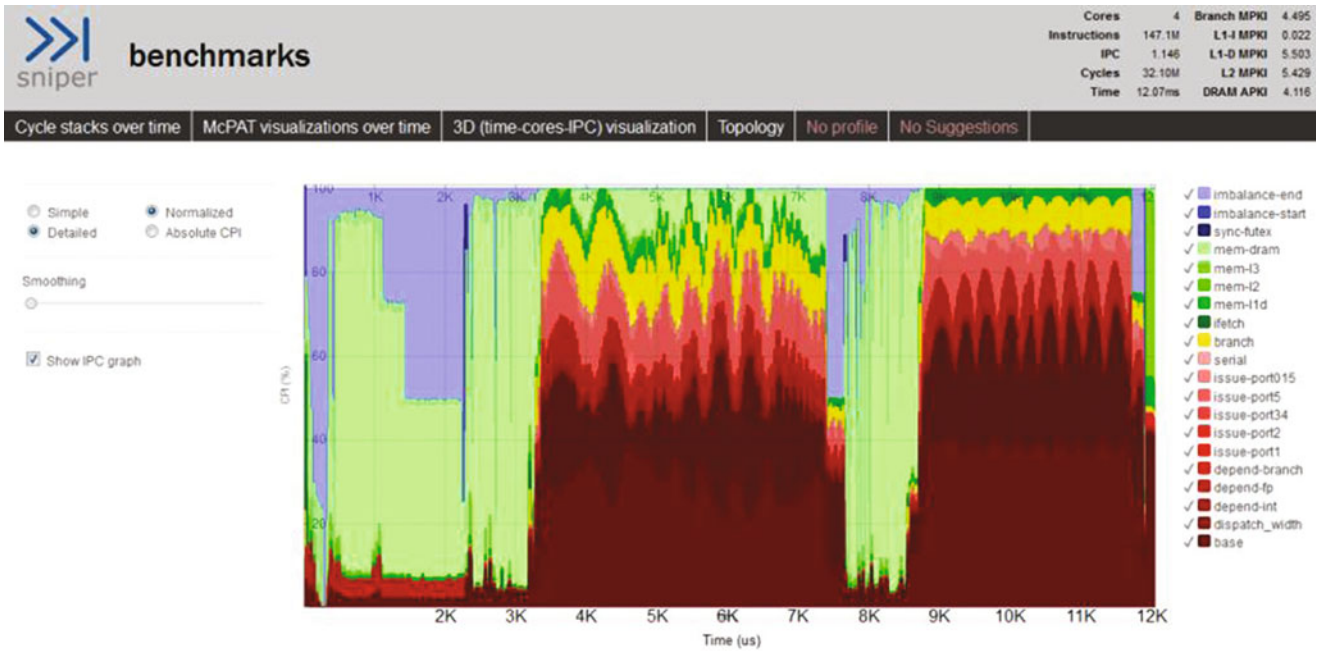


Fig. 3 Cycle stacks plotted over time

- ***Visualization.*** The options of Sniper visualizations help users comprehend the behavior of the software that works in the simulator. Since no single visualization can satisfy needs of various users, the simulator has to produce a number of visualizations. Currently, three main groups of visualizations (Fig. 3). For example, they include cycle stacks plotted against time, McPAT plotted against time and 3D Time-Cores-IPC visualizations. When a user passes the –viz option to the run-sniper command, the visualizations are generated. The Sniper also creates an extra section of system topology on the produced –viz web page.
- ***Multiple Multi-threaded workloads.*** One can run multiple, multi threaded benchmarks simultaneously by installing the integrated benchmarks suite first. Thereafter, the user can utilize the benchmarking parameters to show the benchmarks and configurations that should be run. There is no maximum number of threads that a user can run when utilizing the –benchmark parameters. However, it is worth noting that this mode does not support the ROI handling or the multi-threaded workloads [22].
- ***Scripting.*** The Sniper has a Python interpreter that is used to request for statistics and update the configuration of hardware parameters when the simulator is in use. Consequently, this enables generation of IPC traces and DVFS updates.

5.3 Comparison between Sniper and Graphite

In the interest of comprehending the distinguishing features between graphite and sniper, it is imperative to underline that graphite is the base of sniper whereby it adds, removes and reworks on its features. This work delves on the strengths and weaknesses of different approaches. By design, the default simulator of the Sniper is configured in a similar manner as Nehalem machine the Gainestown micro-architecture. Sniper produces the most significant instruction latencies such as FP and math ops. In the past, Sniper used the Graphite set up as the starting point but Sniper designers designed today's models.

There are several distinct features between the simulators. To differentiate between Sniper and graphite, graphite has been added the MESI coherence protocol while Sniper uses a shared-version of MSI. Recently, Graphite has also been integrated with ATAC [23] and DSENT [24]. Further on, Graphite supports full-mode that is not in the recent versions of Sniper. Full-mode tolerates the distribution of single application simulation across many machines. This is very handy in handy if one has a large memory footprint that would be large enough to stimulate on a single machine. However, this can make the simulation reduce in case of barrier synchronization is used instead of relaxed synchronization that graphite

uses automatically [25]. It is important that the system calls are rooted so that they are executed on behalf of the operator. Sniper uses light-mode entirely, when it passes the system calls onto the OS.

Sniper has more quality visualization features that allow better understanding individual runs with topology information for the architecture being generated. McPAT energy was also incorporated into the visualization and it relays good results when they are run with Sniper core and other models. Further on, Sniper has Python scripting support and a strong link between the SimAPI, scripting and the simulator. Sniper is also compatible with MPI applications as well as the interchange that takes place between different processes through clear virtual-to physical mappings provided by the OS [26]. Sniper had initially added true-DVFS support that allowed each machine to run its own frequency. Sniper database statistics substructures are also sophisticated to allow automatic generation of graphs and reports from the databases. Other features of the Sniper include loop tracer, fault injection, statistically distributed DRAM latencies and micro-op support [22].

6 Conclusion

It is critical to use a simulator tools that model many-core micro architecture details at a faster rate than the full-system simulators characterized by software and hardware overheads. At present, the main challenge is the simulation of thousand-core chips, however, there is no mainstream operating system that can scale to this level. Moreover, the Instruction Set Architectures (ISAs) limits the number of cores thus making it hard to achieve the desired target. In particular, x86's Advanced Programmable Interrupt Controllers (xAPICs) can support maximum 256 cores. At this level of operation, choosing a simulator has to be based on the user-level simulator type. The Sniper simulator is easy to use, and it can adequately balance performance and accuracy trade-offs. However, not all instructions supported by various systems (i.e. the SSE4 and 64-bit x86) are modeled in Sniper.

The number of processes in a system and cores of each packet continues to grow creating the challenge of simulating the growth in system sizes. A rapid growth of multi-core technology combined with much larger cache sizes, requires long and accurate simulations to test the next generation of system designs.

References

1. E. Perelman, G. Hamerly, M. Van Biesbrouck, T. Sherwood, and B. Calder, "Using SimPoint for accurate and efficient simulation," in *ACM SIGMETRICS Performance Evaluation Review*, 2003, pp. 318–319.
2. R. E. Wunderlich, T. F. Wenisch, B. Falsafi, and J. C. Hoe, "SMARTS: Accelerating microarchitecture simulation via rigorous statistical sampling," in *Comp. Architecture, 2003. Proc. of 30th Annual International Symposium*, 2003, p. 84–95.
3. J. Chen, M. Annavaram, and M. Dubois, "SlackSim: a platform for parallel simulations of CMPs on CMPs," *ACM SIGARCH Computer Architecture News*, vol. 37, pp. 20–29, 2009.
4. J. E. Miller, H. Kasture, G. Kurian, C. Gruenwald, N. Beckmann, C. Celio, *et al.*, "Graphite: A distributed parallel simulator for multicores," in *High Performance Computer Architecture (HPCA), 2010 IEEE 16th International Symposium on*, 2010, pp. 1–12.
5. S. S. Mukherjee *et al.*, "Wisconsin Wind Tunnel II: a fast, portable parallel architecture simulator," *Concurrency, IEEE*, vol. 8, pp. 12–20, 2000.
6. A.-M. M. C. Z., "A Survey of Computer System Architecture Simulators, Case Study: Sniper " in *APCASE 2014*, South Kuta, Indonesia, 2014, pp. 014–015.
7. M. M. Martin, *et al.*, "Multifacet's general execution-driven multiprocessor simulator (GEMS) toolset," *ACM SIGARCH Comp. Arch. News*, v.33, pp. 92–99, 2005.
8. T. F. Wenisch, R. E. Wunderlich, B. Falsafi, and J. C. Hoe, "TurboSMARTS: Accurate microarchitecture simulation sampling in minutes," in *ACM SIGMETRICS Performance Evaluation Review*, 2005, pp. 408–409.
9. P. S. Magnusson, F. Dahlgren, H. Grahn, M. Karlsson, F. Larsson, F. Lundholm, *et al.*, "SimICS/sun4m: A virtual workstation," in *Proceedings of Usenix Annual Technical Conference*, 1998, pp. 119–130.
10. T. Austin, E. Larson, and D. Ernst, "SimpleScalar: An infrastructure for computer system modeling," *Computer*, vol. 35, pp. 59–67, 2002.
11. P. S. Magnusson, *et al.*, "Simics: A full system simulation platform," *Computer*, vol. 35, pp. 50–58, 2002.
12. N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, *et al.*, "The gem5 simulator," *ACM SIGARCH Comp. Arch. News*, vol.9, pp.1–7, 2011.
13. V. S. Pai, P. Ranganathan, and S. V. Adve, "RSIM: An execution-driven simulator for ILP-based shared-memory multiprocessors and uniprocessors," *Proceedings of the Third Workshop on Computer Architecture Education*, 1997.
14. M. Rosenblum, S. A. Herrod, E. Witchel, and A. Gupta, "Complete computer system simulation: The SimOS approach," *Parallel & Distributed Technology: Systems & Applications, IEEE*, vol. 3, pp. 34–43, 1995.
15. T. E. Carlson, W. Heirman, and L. Eeckhout, "Sniper: exploring the level of abstraction for scalable and accurate parallel multi-core simulation," in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, 2011, p. 52.
16. D. Sanchez and C. Kozyrakis, "ZSim: Fast and Accurate Microarchitectural Simulation of Thousand-Core Systems," 2013.
17. A. Jaleel, R. S. Cohn, C.-K. Luk, B. Jacob, "CMP$im: A Pin-based on-the-fly multi-core cache simulator," *Proceedings of the 4th Annual Workshop on Modeling, Benchmarking and Simulation (MoBS), ISCA co-located*, 2008, pp.28–36.
18. J. H. Ahn, S. Li, O. Seongil, and N. P. Jouppi, "McSimA+: A Manycore Simulator with Application-level + Simulation and Detailed Microarchitecture Modeling," *ISPASS*, Apr 2013.
19. S. R. Goldschmidt and J. L. Hennessy, *The accuracy of trace-driven simulations of multiprocessors* vol. 21: ACM, 1993.
20. M. D. Dikaiakos, A. Rogers, and K. Steiglitz, "Fast: A functional algorithm simulation testbed," in *Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, MASCOTS'94., Proceedings of the Second International Workshop on*, 1994, pp. 142–146.

21. M. T. Yourst, "PTLsim: A cycle accurate full system x86–64 microarchitectural simulator," in *Performance Analysis of Systems & Software, ISPASS 2007. IEEE International Symposium on*, 2007, pp. 23–34.
22. T. E. Carlson and W. Heirman, "The Sniper User Manual," Nov 13, 2013. Available: http://groups.csail.mit.edu/carbon/?page_id=62
23. *MIT-DSENT*. Available: https://sites.google.com/site/mitdsent
24. (May 2014). *Sniper: Exploring the Level of Abstraction for Scalable and Accurate Parallel Multi-Core Simulations*. Available: *http://snipersim.org/w/Paper:Sc2011Carlson*
25. (May 2014). *The Sniper Multi-Core Simulator*. Available: http://snipersim.org/w/The_Sniper_Multi-Core_Simulator

A Survey on Design and Implementation of Floating Point Adder in FPGA

Luka Daoud, Dawid Zydek, and Henry Selvaraj

1 Introduction

Field Programmable Gate Arrays (FPGAs) are increasingly being used for high performance applications. FPGAs adopt such applications that require high numerical stability and accuracy. They are becoming more attractive solution compared to Application Specific Integrated Circuits (ASIC). FPGAs are hardware reconfigurable computing tools that are suitable to design applications that are frequently upgraded during runtime. Applications can be wiritten in high level language and synthesised in FPGAs [1] that make them popular and easier to implment. FPGAs become more and more popular in the distributed computing [2], where system nodes contain FPGAs and are able to program them according to current needs [3]. Also, they are suitable to design algorithms such [4], [5] allocation algorithms for Chip Multiprocessors [6], [7]. Most of numerical applications demand high level of accuracy in their calculations, and wide range of numbers. Floating point format satisfies such requirements. It has a wide range of numbers that can be presented with the fixed number of bits. Hence, most of applications implemented in FPGAs are represented in floating point format. In FPGAs, performance and area are the main issues to design efficient floating point units. Some compromises must be found between speed, accuracy, and the FPGA resources.

L. Daoud (✉) • D. Zydek
Department of Electrical Engineering, Idaho State University, Pocatello, ID, USA
e-mail: daouluka@isu.edu; zydedawi@isu.edu

H. Selvaraj
Department of Electrical and Computer Engineering, University of Nevada, Las Vegas, NV, USA
e-mail: henry.selvaraj@unlv.edu

1.1 Floating Point Number Format

Floating point numbers are one possible way of representing real numbers in binary format; the IEEE 754 standard [8] presents two different floating point formats, Binary interchange format and Decimal interchange format. In this review, binary floating point operations are indicated. Fig. 1 shows the IEEE 754 binary interchange floating point format. These binary formats are encoded in k bits in the following three fields ordered as shown in the figure. It consists of:

(a) 1-bit sign S.
(b) w-bits biased exponent $E = e + bias$, where $bias = (2^{w-1} - 1)$.
(c) $(p - 1)$-bits trailing significand (or mantissa) field digit string $M = d_1 d_2 ... d_{p-1}$. The leading bit of the mantissa, d_0, is implicitly encoded in the biased exponent E, where the binary number value m is represented by digit string as $(d_0.d_1 d_2 . . . \mathrm{d}_{p-1})$.

Floating point number can be encoded in 16, 32, 64, or multiple of 32 bits. Based on the number of encoding bits, the values of *k, p, w, and bias* - for binary interchange formats - can be determined [8]. The implicit digit d_0 is determined by the exponent E value. If $0 < E < (2^w - 1)$, so d_0 is '1', hence, the floating number is normalized and its value is determined as $(-1)^s \times m \times 2^e$. However, to represent smaller number, if E is 0 and M is nonzero, so d_0 is '0', hence, the floating number is subnormal (or denormalized) and its value is determined as $(-1)^s \times m \times 2^{(1-bias)}$. Table 1 shows two formats that are most-widely used in many applications, single and double floating point formats. Converting a decimal number and representing it into a binary floating point format is a simple process. Let, for example, a decimal number $(2770)_d$ is encoded in 32 bits floating point number, known as single precision floating point format. First, "101011010010" is the corresponding binary format. By moving the radix point to the left such that

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_129,

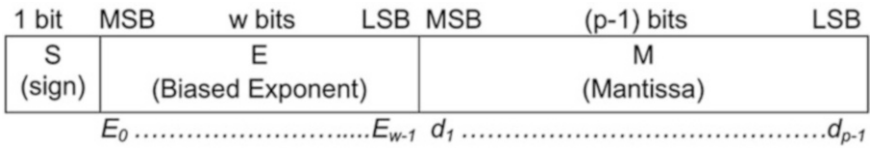


Fig. 1 Binary Interchange Floating Point Format.

Table 1 Single and Double Precision Floating Point Format Summary

Format	Format width (k)	Precision (p)	Exponent width (w)	Exponent bias
Single	32	24	8	127
Double	64	53	11	1023

Table 2 Special Numbers Represented in A Binary Floating Point Format

Sign (S)	Exponent (E)	Mantissa (M)	Object Represented
×	1111....1111	NZ	NaN
1	1111....1111	0000....0000	-∞
0	1111....1111	0000....0000	+∞
×	0000....0000	NZ	Un-normalized
1	0000....0000	0000....0000	-0
0	0000....0000	0000....0000	+0

only one bit '1' on the left of that radix point (this bit '1' is the hidden bit, d_0), the number can be represented as "1:01011010010×2^{11}". So, this number is a positive (S=0), normalized (d_0='1') with mantissa (M) = "01011010010", $e = 11$, and exponent $E = 127 + 11 = 138 = (10001010)_b$. Therefore, this number can be encoded in a single precision format as [SEM]="0 10001010 01011010010000000000000". Moreover, based on the exponent E and the mantissa M, the floating point format can represent special numbers. Table 2 presents special cases of the binary floating point format, where, NZ and NaN are not zero and not a number, respectively.

The IEEE 754 standard for floating point arithmetic is the most widely used standard for floating point computation. It specifies basic and extended floating point number formats, and six numerical operations (Addition, Subtraction, Multiplication, Division, Square-root, and remainder). Also, it provides rules for converting between integer and floating point formats, and converting between different floating point formats.

1.2 Floating Point Addition Technique

Adding two floating point numbers is executed by the following steps: (1) Compute the exponent difference [d] and set the result exponent to be the larger exponent. (2) Mantissa alignment: this is done by right-shifting the number with the smaller exponent d positions. (3) Add/Subtract the two aligned mantissas. (4) Normalization: the resulting mantissa is normalized to comply the floating point number format. (5) Rounding: the result is rounded according to the specified mode to fit the precision. In this review paper, we cover the implementation of the floating point addition algorithms in FPGA. This paper is organized as follows: the next section presents the addition algorithms; section 3 demonstrates hardware implementations of floating point in FPGA. Rounding and exceptions of floating points are presented in section 4 and 5, respectively. Finally, section 6 concludes the paper.

2 Floating Point Addition Algorithms

Addition is the most complicated process compared to multiplier. Comparison of the exponents, alignment of the mantissa, and the normalization methods are much complicated and time consuming in the addition process. For such complexity of the floating point addition, efficient implementations of addition unit in FPGA are being conducted to decrease the addition latency and area consumed in the FPGA. Since floating point addition is more complex than multiplication, this section discusses different techniques for the implementation of the floating point in FPGA. The goal of the algorithms is to optimize area, performance, or to work with higher frequency. Optimization is for only one achievement at the cost of the other. This section explores the trade-offs between size, latency and frequency for floating point addition algorithms. Let A and B are two floating numbers with signs, mantissas, and exponents, S_A, M_A, E_A, and S_B, M_B, E_B, respectively for the two numbers A, and B.

2.1 Standard Floating Point Algorithm

The architecture of the standard floating point algorithm [9], [10] is quite simple. It is a direct implementation of the basic stages required to add two floating point numbers, it checks first whether the numbers are normalized to set the hidden bit (d_0), then it follows the basic steps:

1-. Exponent difference and swap process.
2-. Mantissa alignment (Pre-shift).
3-. Mantissa addition/subtraction.

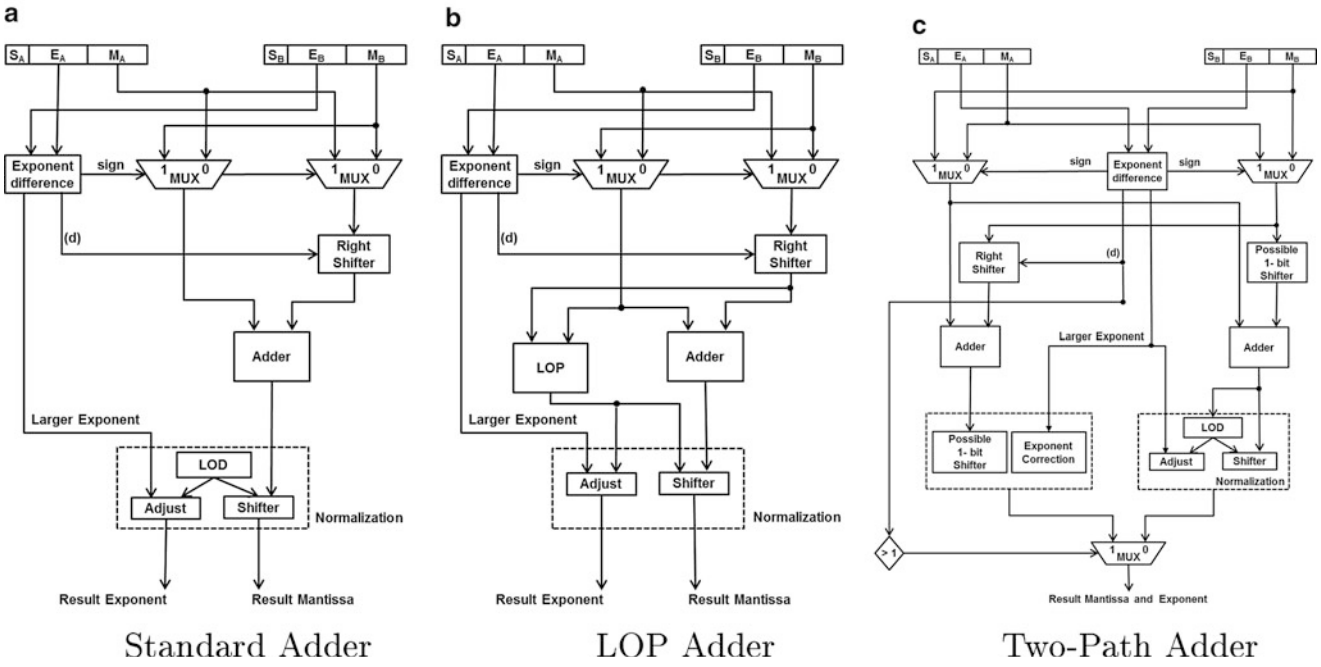


Fig. 2 Floating Point Addition Algorithms

4-. Mantissa Normalization (Post-shift).
5-. Rounding.

Fig. 2a demonstrates the architecture of the standard floating point algorithm. The mantissa normalization is performed through two processes. First, the Leading One Detector (LOD) [11] defines the position of the first leading one, then; the resulting mantissa is shifted by this position. The architecture of the standard floating point algorithm is area-efficient, but it does not provide high performance.

2.2 Leading One Predictor (LOP) Algorithm

In this algorithm, Leading One Predictor (LOP) [12], [13], [14] is implemented to reduce the latency of the normalization process. LOP replaced the LOD circuit to anticipate the position of the most significant one of the addition result in parallel with the mantissa addition. LOP algorithm exhibits higher performance compared to the standard algorithm. However, more area is used. Fig. 2b shows the LOP algorithm architecture.

2.3 Two-path Algorithm

Two-path (known as Far and Close data-path) algorithm [15] is designed to enhance the performance of the floating point addition. Since not all hardware modules are needed for some addition cases, far and close data-path algorithm dedicates two different parallel paths to reduce the latency. One of these two paths is chosen based on the exponent difference. When the exponent difference is 0 or 1, the close path is taken, where the mantissa alignment (pre-shift) is done by at most one bit shift, which is easy to implement. On the other hand, when the exponent difference is larger than one, far path is taken. In that case, a possible one bit shifter is needed in the normalization process. The two-path algorithm is shown in Fig. 2c. Over years there have been some of improvements in the two-path algorithm [9], [16], [17]. Also, a compound adder is integrated in the design to outperform the rounding process. Standard floating point algorithm is the best in terms of the area, but it exhibits less performance compared to the other algorithms. Two-path algorithm is the best one in terms of the performance, but it consumes more area for implementing two different paths. LOP algorithm, however, improves the performance of the standard floating point algorithm with extra area. It is in a moderate state between standard and LOP algorithms. It consumes more area than the standard algorithm and less than two-path algorithm, and its performance locates between those two algorithms.

3 Hardware Implementation of Floating Point Adders in FPGA

Since floating point addition is more complex than multiplication [18], research is dedicated to achieve efficient addition. So, in this section, we demonstrate the basic

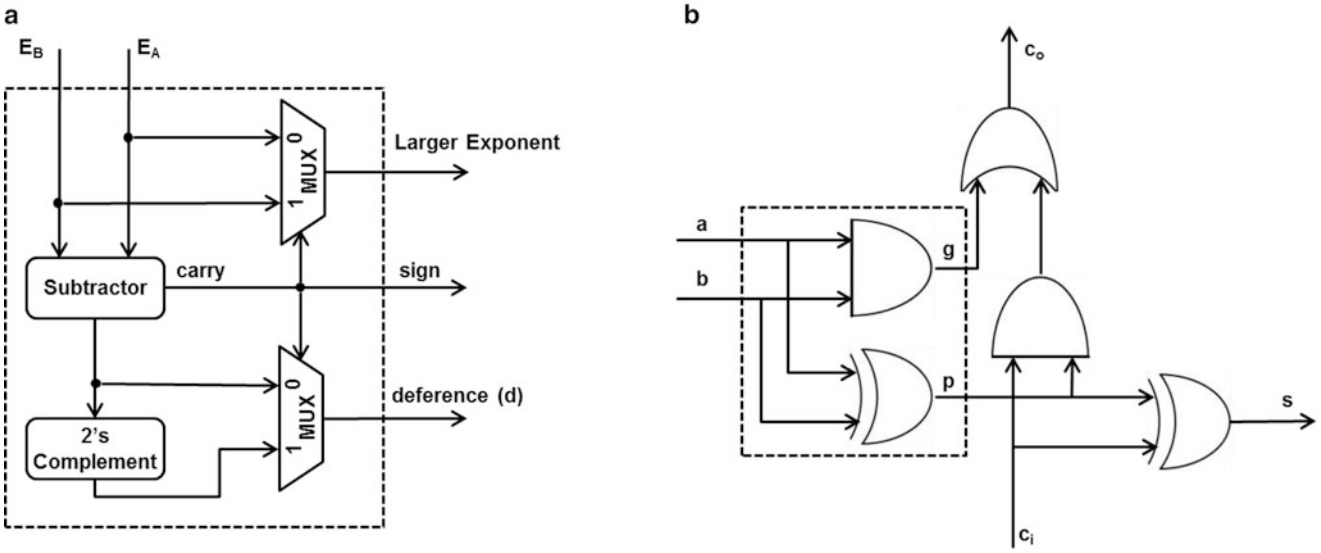


Fig. 3 Adder Modules (a) Exponent Differnce and (b) Carry Look Ahead Adder

components of a floating point adder and efficient implementations of adder in the FPGA.

3.1 Floating Point Adder Components in Hardware

The basic floating point adder consists of the following elementary modules to complete the addition process:

- The difference between the two exponents is calculated by a subtractor (adder) unit.
- A swap unit (multiplexer) to select the mantissa with the smallest exponent that needs alignment.
- The mantissa alignment is done by a right-shifter component.
- Addition/subtraction of the two mantissas is computed by an adder unit.
- The post-normalization unit that normalizes the result of the addition process. It consists of two components: a leading-one detector (LOD) that counts the number of leading-zeroes and a left/right shifter based on the effective operation.
- Rounding component to map the result into finite precision.

Efficient implementation of these components in hardware leads to an efficient floating point adder.

3.1.1 Exponent Difference Module

The task of the exponent difference is to provide us with two values, the sign difference which defines the smallest exponent, and the difference value that is needed for mantissa alignment. Fig. 3a shows a simple implementation of exponent difference circuit. The two exponents are subtracted by 2's complement adder to get the sign value to define the smallest exponent and choose the correct exponent difference value.

3.1.2 Adder Module

Adder circuit is the main component in the floating point unit. It only implements the operation (addition or subtraction). The vital issue of the adder circuit is the propagation of the carry. Adder circuits have been improved in order to gain less latency. Ripple-Carry adder, Carry-Skip adder, Carry-Select adder, and Carry Look Ahead adder [19], [20], [21] are the most well-known adders. In Ripple-Carry adder, the carry bit is calculated alongside the sum bit, and each bit waits until the previous carry has been calculated. A Carry-Skip adder improves the delay of the Ripple-Carry adder. The Carry-Select adder is a particular way to implement an adder. It consists of two Ripple-Carry adders and a multiplexer. It's simple and a quite faster with a gate level depth of $O(\sqrt{n})$, where n is the bit number width. On the other hand, a Carry Look Ahead (CLA) adder [22] is the fastest adder compared to its counterparts. It improves the addition latency. It calculates the carry bits before the sum which reduces the wait time to calculate the result of the large value bits. Fig. 3b shows a 1-bit CLA adder structures. From adder implementation analysis in [23], [24], the Xilinx built-in adder is the most efficient adder. It shows the least combinational delay and smallest area requirement. That is because the FPGA structure provides special support operations such as carry propagation. The CLA adder, however, provides the best delay among customized adders because the carry is calculated separately.

3.1.3 Shifter Module

The basic floating point adder requires two shift components, right shifter and left shifter for mantissa alignment and normalization. In [23], [25], shifters in FPGAs are implemented as a series of multiplexers. The shifting operation is done in constant time regardless the number of bits to be shifted. The advantage of these shifters is that pipeline can be applied to increase the throughput. Embedded multipliers available in Virtex II are exploited to implement such shifters, direct wires connections to implement right shifter and reverse connections for left shifters. Two types of shifter were implemented in [24], align shifter and barrel shifter. They have the same area, but barrel has lower latency. Align and barrel have the same concept, but the multiplexers in align shifter were implemented behaviorally. The shifter modules in the align shifter were designed using concatenation operation on VHDL.

3.1.4 Leading One Detector (LOD) Module

After the addition process, the next step is to normalize the result to comply the IEEE floating point format. The first step is to hit the first one of the result number, and then shifting process is applied. In order to identify the first one of the result number, a special component has to be implemented, called Leading One Detector (LOD) or Leading Zero Counter (LZC). Oklobdzija [11] designed an efficient LOD circuit with less area and efficient delay. Behavioral and Oklobdzija LODs were implemented in [24]. The behavioral model was implemented using 'case' statements in VHDL defining each possibility behaviorally. Oklobdzija LOD shows less latency and area compared to the behavioral VHDL.

3.1.5 Leading One Predictor (LOP) Module

In order to normalize the result of addition, the first bit with value one of the result is searched. The process occurred after the addition process by the LOD. However, the searching process and the addition process can be executed simultaneously to reduce the latency. Leading One Predictor (LOP) [12], [13], [14] can predict the position of the first one of the result number in parallel with the adder. The prediction might be in error by one bit, so the correction of this error result in a delay increase. Therefore, [26] presented a novel design incorporates a concurrent position correction logic, operating in parallel with the LOP, to detect the presence of that error and produce the correct shift amount. An LOP has three major modules: the pre-encoder, an LOD, and an error-detection tree. Although LOP reduces the latency of the floating point adder unit, it consumes a lot of resources to implement in FPGA. In [24], LOP, LOD, and adder were implemented on a Virtex-II Pro FPGA showing that LOP has 11.8 % enhancement in performance but at cost of large area.

3.2 Efficient Implementation of Floating Point Adder Algorithms in FPGA

Floating point addition are the most frequent floating point operations. Therefore, a lot of research has been conducted to design efficient floating point adders in order to reduce the latency and improve the performance [9], [15], [16], [19], [27], [28], [29]. Some of works have been carried out implementing floating point units in FPGA [10], [19], [25], [30], [31], [32] in order to attain efficient latency and area designs. In [10], authors developed a tool that gives the user the option to create numerous collection of floating point units with different throughput, latency, and area characteristics. One of the drawbacks of their work is that they lack implementation of overflow and underflow exceptions. In [31], authors analyzed the floating point multiplier and adder/subtractor units following the IEEE 754 format for single and double precision. They considered the number of pipeline stages as a parameter to measure the throughput and the area consumed in the FPGA. In that work, *frequency/area* was used as a new metric. However, implementations of denormalized and *NaN* (not a number) have not been provided in their evaluation. [32] presented efficient implementation of floating point adder using 5 pipeline stages for two-path algorithm with LOP in FPGAs. On the other hand, [33] achieved good floating point performance in Virtex-IV. The specific architecture of the target FPGA and parallel optimization approach have been considered to attain high performance. [25] took the advantages of the special function elements of the FPGAs. That optimization got the benefit of the embedded multipliers and shift registers of Xilinx Virtex-II. For further performance, [34] proposed a novel approach of the floating point addition by efficiently exploiting both paths from the two path adder. Thus, two floating point additions can be executed simultaneously, each one on a different path. [24] provided a design trade-off analysis of floating point adders in FPGAs. In this work, standard, leading-one predictor (LOP); and two-path floating point addition algorithms have been implemented in FPGAs. Their components have been implemented and synthesized onto Xilinx Virtex-II Pro FPGA device. Standard and LOP algorithms were implemented into 5 stages and compared with Xilinx IP. Research on efficient implementations of floating point units on FPGAs is still being conducted. [35] demonstrated a novel technique to implement a double precision IEEE floating point adder that can complete the addition operation within two clock cycles. In addition, [36] optimized the individual complex components of the adder module. In order to reduce latency, [36] suggested to perform the rounding process in parallel with the addition process before normalization. Moreover, [37] proposed a novel architecture to optimize the floating point computation units in hybrid

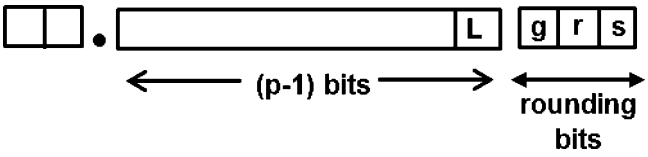


Fig. 4 Mantissa Format of the Result Number before Rounding.

FPGAs. It optimized the implementation of addition, subtraction, multiplication and division units in order to gain a better reduction in area and low power consumption. On the other hand, [22] implemented a single floating point adder in a way that takes advantages of the operations with the special numbers that don't need a lot of processes, for example, addition to zero, infinity, or *NaN*. Furthermore, it supports addition of two normalized, denormalized, or mixed numbers.

4 Rounding

Floating point number is represented by a limited number of bits. Therefore, rounding is required to modify an infinite precise number in order to fit the destination's format. Rounding is a many-to-one mapping that maps an unrepresentable number into a representable one. Thus, rounding mode affects the results of the most arithmetic operations. In IEEE 754 floating point [8] representation, four rounding modes are defined: round towards nearest even (REN), round towards $+\infty$ (RP), round towards $-\infty$ (RM), and round towards 0 (RZ). The mantissa before rounding has the format as shown in Fig. 4. In this figure, L is the Least Significant Bit (LSB), g is the guard bit, r is the round bit, and s is the sticky bit. The guard bit is one bit less significant than that is restricted by the format. The sticky bit is a flag indicating the existence of any bits in the result less significant than the guard bit. It's the logical OR of all of the less significant bits in the result. Based on these bits and the rounding mode, the mantissa number can be rounded by neglecting these bits or adding '1' when necessary to the LSB of the mantissa. In this case, overflow may happen, requiring a right shift and increment of the exponent by one. In infinity rounding mode, adding "10" may be required. Let's define that *sum* is the result of adding/ subtracting two numbers, A and B, i.e, $sum = A \pm B$.

4.1 Rounding towards Nearest Even (REN)

REN rounding gives the closest possible exact value[21]. Hence, it is the default rounding mode of the IEEE standard and mostly used in all the arithmetic implementations in software and hardware. REN rounding [38] relies on the effective operation and the g, r, s bits. If the effective operation is addition, a normalized result is obtained or overflow can occur requesting a one bit right shift and exponent increment by one. In case of the effective operation is subtraction, one of these four results is obtained: positive and normalized, positive and unnormalized requiring 1-bit left shift, positive and unnormalized requiring left shift more than 1-bit, or a negative result. In REN rounding [38], it selects between *sum* and *sum* + 1, based on the effective operation and the bits g, r, s. In order to enhance the performance of the floating point a compound adder is used [38], [39] that computes *sum* and *sum* + 1 simultaneously.

4.2 Rounding to Infinity

Two types of rounding to infinity: rounding towards Positive Infinity (RP), and rounding towards Negative Infinity (RM) [38], [39]. In this rounding the sign of the floating point result is considered. In rounding to RP the result is rounded up if the sign of the result is positive and any bits to the right of the result LSB is '1'. However, in rounding to RM a 1 is added to the result if the sign is negative and any bits to the right of the result LSB is '1'. In rounding to infinity, *sum* + 2 is required to correct rounding in case of overflow requiring a 1-bit normalizing right shift. *sum* + 2 can be calculated from a compound adder by using a row of half adders above the compound adder as suggested in [40].

4.3 Rounding towards Zero (RZ)

Rounding towards zero is the simplest one. The rounded mantissa is obtained by discarding bits that are less significant than LSB. It requires the least amount of hardware.

5 Exceptions

There are some cases in which a floating point operation produces a value that is not representable in the floating-point number system. The IEEE 754 defines five types of exceptions [8]: Overflow, Underflow, Invalid Operation, Inexact Result, and Division-by-Zero. In such cases, exceptions are signaled by setting a flag or setting a trap. Overflow and underflow are the most frequent exceptions that occur during addition. Invalid Operation exception flag is set when the given operation cannot be performed on the operands, for instance, subtraction of infinity or *NaN* inputs. On the other hand, Inexact Result exception is set when the rounded result is not exact or it overflows without an overflow trap. Division-by-Zero flag is set when the divisor is zero; the result is set to signed infinity. Overflow exception is

defined by the rounding mode used. In REN, Overflow flag is raised if the exponent of the rounded result has exceeded the format's largest finite number. On the other hand, Underflow exception occurs when there is a loss of accuracy. The underflow flag is set when the number is too small to be represented fully in the floating point format.

6 Final Remarks

Floating point numbers are widely adopted in numerous applications due to their representation of wide range of numbers. For such applications, FPGAs are extensively used to gain high performance, where conventional processors implementations do not satisfy the real-time computations. Implementing Floating point arithmetic unit on FPGA rises to the challenge of balancing between the consumed area and the performance. The design of floating point addition is relatively complex than other flotation point arithmetic operations. In this review paper, floating point addition algorithms have been presented and previous works for efficient hardware implementation of such addition algorithms on FPGAs have been reviewed. The main components of the floating point adder and their hardware design were explained. Some of these components still need enhancement to achieve less area or/and much performance. As a future work, we plan to improve hardware implementation of such modules to match with the FPGAs' architectures.

References

1. L. Daoud, D. Zydek, and H. Selvaraj: A Survey of High Level Synthesis Languages, Tools, and Compilers for Reconfigurable High Performance Computing. In: Advances in Systems Science. Springer (2014) 483–492, DOI: 10.1007/978-3-319-01857-7_47.
2. G.Chmaj, K. Walkowiak, M. Tarnawski, and M. Kucharzak: Heuristic Algorithms for Optimization of Task Allocation and Result Distribution in Peer-to-Peer Computing Systems. International Journal of Applied Mathematics and Computer Science **22**(3) (2012) 733–748, DOI: 10.2478/v10006-012-0055-0.
3. G. Chmaj, H. Selvaraj, and L. Gewali: Tracker-Node Model for Energy Consumption in Reconfigurable Processing Systems. In: Advances in Systems Science. Springer (2014) 503–512, DOI: 10.1007/978-3-319-01857-7_49.
4. L. Daoud, and V. Goulart: High Performance Bitwise OR Based Submesh Allocation for 2D Mesh-Connected CMPs. In: Proceedings of the Euromicro Conference on Digital System Design (DSD), IEEE (2013) 73–77, DOI: 10.1109/DSD.2013.134.
5. L. Daoud, M. E. Ragab,and V. Goulart: Faster Processor Allocation Algorithms for Mesh-Connected CMPs. In: Proceedings of the Euromicro Conference on Digital System Design (DSD), IEEE (2011) 805–808, DOI: 10.1109/DSD.2011.107.
6. D. Zydek, G. Chmaj, and S. Chiu: Modeling Computational Limitations in H-Phy and Overlay-NoC Architectures. The Journal of Supercomputing (2013) 1–20, DOI: 10.1007/s11227-013-0932-9.
7. D. Zydek: Processor Allocator for Chip Multiprocessors. PhD thesis, University of Nevada, Las Vegas, USA (2010)
8. IEEE Computer Society: IEEE Standard for Floating-Point Arithmetic. (Aug. 29, 2008)
9. S. F. Oberman, H. Al-Twaijry, and M. J. Flynn: The SNAP Project: Design of Floating Point Arithmetic Units. In: Proceedings of the 13th IEEE Symposium on Computer Arithmetic, IEEE (1997) 156–165
10. J. Liang, R. Tessier, and O. Mencer: Floating Point Unit Generation and Evaluation for FPGAs. In: Proceedings of the 11th Annual IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM), IEEE (2003) 185–194
11. Vojin G. Oklobdzija: An Algorithmic and Novel Design of a Leading Zero Detector Circuit: Comparison with Logic Synthesis. IEEE Transactions on Very Large Scale Integration (VLSI) Systems **2**(1) (1994) 124–128
12. E. Hokenek, and K. Montoye: Leading-Zero Anticipator (LZA) in the IBM RISC System/6000 Floating-Point Execution Unit. IBM Journal of Research and Development **34** (1990)
13. N. Quach, and M. J. Flynn: Leading One Prediction–Implementation, Generalization, and Application. Computer Systems Laboratory, Stanford University (1991)
14. H. Suzuki, H. Morinaka, H. Makino, Y. Nakase, K. Mashiko, and T. Sumi: Leading-zero Anticipatory Logic for High-Speed Floating Point Addition. IEEE Journal of Solid-State Circuits **31** (1996)
15. P. M. Farmwald: On the Design of High Performance Digital Arithmetic Units. Technical report, Lawrence Livermore National Lab, CA (USA) (1981)
16. A. Beaumont-Smith, N. Burgess, S. Lefrere, and C-C. Lim: Reduced Latency IEEE Floating-Point Standard Adder Architectures. In: Proceedings of the 14th IEEE Symposium on Computer Arithmetic, IEEE (1999) 35–42
17. P-M. Seidel, and G. Even: On the Design of Fast IEEE Floating-Point Adders. In: Proceedings of the 15th IEEE Symposium on Computer Arithmetic, IEEE (2001) 184–194
18. M. M. Ozbilen, and M. Gok: A Multi-Precision Floating-Point Adder. In: Porceedings of Research in Microelectronics and Electronics, IEEE (2008) 117–120
19. R. N. Giri, and M. K. Pandit: Pipelined Floating-Point Arithmetic Unit (FPU) for Advanced Computing Systems using FPGA. International Journal of Engineering and Advanced Technology (IJEAT) (2012) 2249–8958
20. S. Xing, and W. Yu: FPGA Adders: Performance Evaluation and Optimal Design. Design & Test of Computers, IEEE **15**(1) (1998) 24–29
21. Milos Ercegovac, and Tomas Lang: Digital Arithmetic. Access Online via Elsevier (2003)
22. P. S. Gollamudi, M. Kamaraju: Design of High Performance IEEE-754 Single Precision (32 bit) Floating Point Adder Using VHDL. International Journal of Engineering Research & Technology **2** (2013)
23. I. O. Flores, M. Jimenez, and D. Rodriguez: Optimizing the Implementation of Floating Point Units for FPGA Synthesis. In: Proceedings of Computing Research Conference CRC2002. (2002)
24. A. Malik, D. Chen, DY. Choi, M. H. Lee, and S-B. ko: Design Tradeoff Analysis of Floating-Point Adders in FPGAs. Canadian Journal of Electrical and Computer Engineering **33**(3/4) (2008) 169–175
25. E. Roesler, and B. Nelson: Novel Optimizations for Hardware Floating-Point Units in a Modern FPGA Architecture. In: Field-Programmable Logic and Applications: Reconfigurable Computing Is Going Mainstream. Springer (2002) 637–646
26. J. D. Bruguera, and T. Lang: Leading-One Prediction with Concurrent Position Correction. IEEE Transactions on Computers **48**(10) (1999) 1083–1097

27. A. M. Nielsen, D. W. Matula, C. N. Lyu, and G. Even: An IEEE Compliant Floating-Point Adder that Conforms with the Pipelined Packet-Forwarding Paradigm. IEEE Transactions on Computers **49** (1) (2000) 33–47 007.
28. W-C. Park, S-W. Lee, O-Y. Kwon,T-D. HAN and S-D. Kim: Floating Point Adder/Subtractor Performing IEEE Rounding and Addition/Subtraction in Parallel. IEICE transactions on Information and Systems **79**(4) (1996) 297–305
29. N. Quach, N. Takagi, and M. Flynn: On Fast IEEE Rounding. Computer Systems Laboratory, Stanford University (1991)
30. W. B. Ligon III, S. McMillan, G. Monn, K. Schoonover, F. Stivers, and K. D. Underwood: A Re-evaluation of the Practicality of Floating-Point Operations on FPGAs. In: Proceedings of the IEEE Symposium on FPGAs for Custom Computing Machines, IEEE (1998) 206–215
31. G. Govindu, L. Zhuo, S. Choi, and V. Prasanna: Analysis of High-Performance Floating-Point Arithmetic on FPGAs. In: Proceedings of the 18th International Symposium on Parallel and Distributed Processing, IEEE (2004) 149–156
32. A. Malik, and S-B. Ko: Effective Implementation of Floating-Point Adder Using Pipelined LOP in FPGAs. In: Proceedings of the Canadian Conference on Electrical and Computer Engineering, IEEE (2005) 706–709
33. P. Karlstrom, A. Ehliar, and D. Liu: High performance, Low Latency FPGA Based Floating Point Adder and Multiplier Units in a Virtex 4. In: Proceedings of the 24th Norchip Conference, IEEE (2006) 31–34
34. A. Amaricai, M. Vladutiu, L. Prodan, M. Udrescu, and O. Boncalo: Exploiting Parallelism in Double Path Adders' Structure for Increased Throughput of Floating Point Addition. In: Proceedings of the 10th Euromicro Conference on Digital System Design Architectures, Methods and Tools, IEEE (2007) 132–137
35. S. Ghosh, P. Bhattacharyya, and A. Dutta: FPGA Based Implementation of a Double Precision IEEE Floating-Point Adder. In: Proceedings of the 7th International Conference on Intelligent Systems and Control (ISCO), IEEE (2013) 271–275
36. Y. Huijing, Y. Fan, and H. Dandan: High Performance FPGA Implementation of Floating Point Addition. Applied Mechanics and Materials **380** (2013) 3316–3319
37. H. Anand, D. Vaithiyanathan, R. Seshasayanan: Optimized Architecture for Floating Point Computation Unit. In: Proceedings of the 2013 International Conference on Emerging Trends in VLSI, Embedded System, Nano Electronics and Telecommunication System (ICEVENT), IEEE (2013) 1–5
38. J. D. Bruguera, and T. Lang: Rounding in Floating-Point Addition Using a Compound Adder. University of Santiago de Compostela, Spain Internal Report (2000)
39. S. F. Oberman: Design Issues in High Performance Floating Point Arithmetic Units. PhD thesis, Stanford University (1996)
40. N. T. Quach, and M. J. Flynn: An Improved Algorithm for High-Speed Floating-Point Addition. Computer Systems Laboratory, Stanford University (1990)

Hybrid GPU/CPU Approach to Multiphysics Simulation

Dawid Krol, Jason Harris, and Dawid Zydek

1 Introduction

System identification and mathematical models are integral parts of various areas of business, science, and engineering. Modern simulation and Computer Aided Design (CAD) allows engineers from different fields to increase the effectiveness of their work and research. Most engineers are familiar with various software packages and some basic programming knowledge [1]. This results in various ideas like the software development approach for discrete simulations presented in [2]. Computer science knowledge allows engineers using custom libraries to create applications that suit their requirements and expectations. This trend is not surprising since multiphysics simulation is crucial for almost every civil engineer project, automobile design facility, chemical laboratory, and nuclear power plant. Simulation helps to improve the understanding of functionality and behavior of a model. Using models allows prediction of future conditions and foreseeing possible issues. Very often simulating a model is the only potion of a problem that can be solved. Furthermore, multiphysics simulation, saves a lot of time, energy, and subsequently, money [1].

Multiphysics simulation is a complex process that requires significant time and computational resources. It involves a number of physical phenomena usually described by Partial Differential Equations (PDEs). Currently one of the most popular approaches to solve PDEs is the Finite Element method (FE method). Originally developed in 1943 A. Hrennikoff and R. Courant it gained popularity in 1960 when applied to the problem of electromagnetic wave propagation [3]. The commercial market offers a wider variety of multiphysics simulation software like COMSOL or ANSYS Multiphysics, but also open source software like the, MOOSE Framework (released on March 21st, 2014) [4]. Although these aforementioned applications can be easily used on a typical modern personal computer, more advanced simulations that incorporate a significant number of physics phenomena requires very precise results. These simulations are performed on meshes assembled from millions of shapes and require an enormous amount of computational resources. Engineers and scientists are eager to use new technologies that may offer promising possibilities for performing simulations faster and easier. An example of this is the Hardware-Physical (H-Phy) and Overlay-Network-on-Chip (Overlay-NoC) approach to manage computational resources in multi-processor supercomputers [5]. The multi-node distributed processing systems are structures that are able to deliver high performance to many applications that can be processed in the distributed manner. Such systems consist of many devices having computational capabilities and connected into one logical structure that works on the given task. The distributed processing system based on the overlay network was proposed in [6], where authors achieved optimized operation in terms of the OPEX (operational expenditure) through the use of peer-to-peer networking mechanisms. This work shows the importance of distributed-processing related algorithm decision strategies and their influence on the computation/processing efficiency.

One of the most popular technologies that lie underneath high computational capabilities of many supercomputers is the solution offered by NVidia Company. The idea assumes using modern Graphical Processing Units (GPUs) to perform general purpose computing (GPGPU). GPU is composed of a number of independent multiprocessor units called Streaming Multiprocessors (SMs). SMs execute in parallel thousands of instances of code called kernels. A kernel is a code written in CUDA, which is the acronym for Compute

D. Krol (✉) • D. Zydek
Department of Electrical Engineering, Idaho State University, Pocatello, ID, USA
e-mail: kroldawi@isu.edu; zydedawi@isu.edu

J. Harris
Department of Nuclear Engineering and Health Physics, Idaho State University, Pocatello, ID, USA
e-mail: harrjaso@isu.edu

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 366,
DOI 10.1007/978-3-319-08422-0_130,

Unified Device Architecture. CUDA is a scalable parallel programming model and software environment that originates from the C language and was developed by NVidia [7]. Although a single processor within a SM does not provide high performance and the single thread is not executed as fast as it would be on a modern Central Processing Unit (CPU), the ability to execute a massive number of threads in parallel gives GPU exceptional performance.

GPU receives praise and has been applied to a number of projects. The most significant example is the supercomputer Titan, which consist of over 18 000 NVidia Tesla K20 GPUs and almost 3000 000 AMD Opteron CPU cores. This combination of processing devices accumulates to the peak performance of over 20 petaflops. According to the TOP 500 ranking, Titan is currently ranked 2nd in the world [8]. The past few years also brought other notable applications of GPU and CUDA software in various field of research. In [9], the author uses the CUDA application to simulate lava flow. The code, originally designed and implemented as an HTML base web application, was ported to CUDA and executed on GTX580 (Fermi architecture) and GTX680 (Kepler architecture). Experiments show that in both cases the GPU based application was faster than CPU based software by 7% to 29% (depending on used memory). In [10], GPU was used to accelerate the training of the Locally Connected Neural Pyramid (LCNP). The LCNP training consist of, among others, frequent conversion from RGB color to YIQ color, which is a highly parallel process. Experiments were carried out on GTX480 (Fermi architecture). Porting the code responsible for conversion to GPU allowed authors to accelerate the training process over 22 times. The GPU based approach was also used in [11], where authors intended to improve the performance of the IPEG2000 encoding process. Implementing the code using CUDA allowed them to decrease the coding process duration. Experiments performed on GTX 580 (Fermi architecture) show that the algorithm proposed by authors was nearly 20 times faster than the reference algorithm. Significant speed-up of 11.5 was also obtained in [12].

Authors have also used 7800GTX (G70) to solve the Raleigh-Taylor instability problem. In [13], a 20 times increase in performance was reported. Authors applied GPU to an Euler computation on a full hypersonic vehicle with complex geometry. The same problem was investigated in [14] and a speed-up of 29 times was obtained. In [15], authors performed two-dimensional Euler computations on a computer equipped with 32 GPUs and achieved an enhanced performance by 496 times. Recent research also brought interesting concepts of frameworks and libraries that extends and simplifies usage of CUDA. In [16] and [17], a framework that encapsulates CUDA kernels using expression templates and operator overloading was presented. An approach that uses semi-automatic mechanisms that ports code to GPU were proposed in [18]. Another framework, called CU++, which encapsulates CUDA kernels, was considered in [19]. CU++ focuses on parallel array arithmetic by using simplified data structures and intelligent indexing to speed-up operations.

In the following paper a hybrid GPU/CPU approach to perform a multiphysics simulation is proposed. The idea behind the approach is to move classic Central Processing Unit (CPU) heavy computational parts of a common multiphysics simulation system to GPU. Specified parts of code have to be redesigned and implemented in the form of kernels. The concept is applied to the FE library called libMesh, which is a module in the MOOSE Framework. As a result, all existing applications that use the MOOSE Framework will work without modification with the GPU enabled version of MOOSE.

The remainder of the paper is organized as follows. In Section 2, the motivation of the considered approach is presented. Motivation includes a brief description of the MOOSE Framework, expectations and observations regarding the porting part of the framework to GPU, and description of the presented hybrid approach. Section 3 contains implementation of the considered approach, together with a description of modification that has to be applied to the existing FE library. The description of an experiment, together with results, are presented and discussed in Section 4. Section 5 contains the summary of the paper and proposition of further research directions.

2 Motivation

The name of the MOOSE Framework comes from Multiphysics Object Oriented Simulation Environment. The software was developed at Idaho National Laboratory (INL) and originally was designed to run on a multiprocessor supercomputer. MOOSE is the multiphysics parallel computational framework. The framework allows solving of computational engineering problems in an organized and coordinated manner. Although the application was designed for a supercomputer, it can also be used on standard machines with a few or even just one CPU. Since March 21st 2014, MOOSE has become an open source application.

The MOOSE Framework consists of four independent but cooperating layers. Among these layers are well-known and open source components, like FE solvers PETSc and Trilinos, and FE library libMesh. The diagram presented in Fig. 1 illustrates the structure of applications that uses the MOOSE framework [1]. The Multiphysics Module layer consists of information about physical phenomena (like the variables, boundary conditions, etc.), physic kernel that describes the PDEs, and the mesh file that represents the

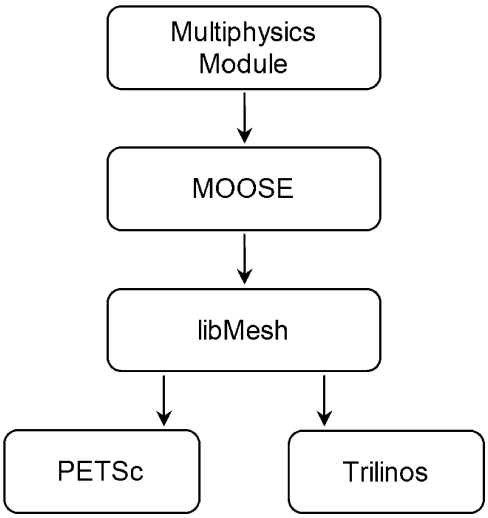


Fig. 1 MOOSE Framework diagram

object. Underneath, the Multiphysics Module framework is located. The main goal of the framework is to provide the most flexible and convenient access to the FE library. Functions implemented in MOOSE allow initializing various components used in simulation, assigning a memory to the application, parsing the input file, etc. MOOSE also provides the framework to create custom physics kernels and variables associated with methods implemented in libMesh. The framework interface in the FE library is called libMesh. FE library is responsible for discretizing the weak form of a PDE, creating matrices uses to solve them, applying shape functions to cells in a mesh, and invoking methods provided by solvers. LibMesh also provides support for adaptive mesh refinement. The last layer in the framework is a FE solver that solves the PDE for specified parameters, points, and cells in the mesh.

The simulation process starts with a single thread that extracts data from the input file and initializes all required components like MPI solvers and MOOSE. After the environment for an application is created and memory is allocated, PDEs are passed and discretized. Finally, the framework runs the simulation by giving control to libMesh. In the FE library a simulation is carried out in parallel on a number of processors specified in previous steps. Each thread is responsible for solving the equation for a single cell in a mesh. After the simulation results are gathered, a log is again created by a single thread [4].

Although the simulation is already performed in parallel with MPI on multiprocessor supercomputers, large scale simulations of models represented by a multimillion cell mesh still calls for more computational power. Since the frequency boundary of the CPU was almost reached, the only solution is to increase the number of computational nodes. However, upgrading existing supercomputers by adding new CPUs is very pricey and the increase of performance may be relatively small compared to the cost of modernization. The application of GPUs that can run hundreds of thousands of threads in parallel seems to be a very good alternative. However, there are some issues with porting the whole framework to GPU. The biggest of them is the architecture of libMesh. LibMesh is a highly object oriented application written in C++. Investigation of the code reveals that the library strongly relies on abstract classes, polymorphism, and Standard Template Library (STL) containers. These kinds of custom objects are hard to port directly to the CUDA code. The main problem is that STL containers use pointers extensively. For example, when the STL vector is copied to GPU using standard CUDA methods the values of elements stored in vector pointers to the elements are copied. Actual data is still stored in the computer memory (for GPU the device computer is a host) and cannot be accessed directly by GPU. When the CUDA kernel refers to the GPU memory, it points to the address that does not exist. This causes a fatal error. The solution of this problem would be to completely redesign the architecture of libMesh, however it would be an extremely time consuming and expensive project, not to mention that results still may be unsatisfying. Because of this, a hybrid approach that uses GPU and CPU is proposed in this paper.

It has been observed that the beginning of simulation is identical to the one in the standard approach. One thread initializes all required components, parses input, and spawns a number of threads that solve the equations in parallel. At the stage when threads are executed in parallel by CPUs another level of parallelism is added. If the threaded approach of the part of the code is highly parallel (like matrix multiplication) then it invokes a CUDA kernel that is executed by the underlying GPU [7]. The more computational demanding is the part of the code ported to GPU the bigger increase of performance would be. After the GPU finishes execution of the heavy computational part of the code, results are passed back to the host and the CPU can continue execution of rest of the code. At the end, as in standard simulation flow, results are gathered and the output is prepared.

The proposed method is a less expensive and easier alternative to implement porting the whole FE library to GPU. The only difficulty is to discover parts of the code that are appropriate to move to the GPU. However, it is still less time consuming than porting the whole library. Ideally when most of the heavy computational sections are reimplemented in CUDA, the performance of simulation would be significantly enhanced. An increase of effectiveness is accompanied with relatively (when compared to modern multiprocessor nodes) small cost. Furthermore, since a node with a CPU is a host for a GPU, the single GPU can be shared by all CPUs that resides in a node. It is

also important to notice that the proposed approach does not affect existing applications built of the MOOSE Framework. To make the application work all that needs to be done is recompilation.

3 Implementation

The original CUDA Software Development Kit was released in February 2007. Since that time the CUDA enabled GPU architecture has changed many times. A typical part of a device lifecycle is a stage at which it is compared to its predecessors. In this process a software benchmark was required. Matrix multiplication is a highly parallel problem that is very time consuming when executed by a single thread and very easy to solve by well-organized parallel threads. Matrix multiplication is also an integral part of every FE method based multiphysics simulation. At the same time it is one of the most time-consuming operation simulations [7]. Therefore, matrix multiplication was a natural candidate to test the hybrid approach presented in this paper.

The code used for matrix multiplication in libMesh can be found in various classes that implement vectors and matrices. The examples are method *left_multiply_transpose* and *right_multiply_transpose* in *DenseMatrix* class or *multiply* in *DenseMatrixBase* class. The matrix multiplication algorithm used in libMesh is a standard CPU algorithm presented in [7]. The pseudo code in presented in Algorithm 1.

```
1 for i = 0 to M do
2 |   for j = 0 to N do
3 |   |   for k = 0 to L do
4 |   |   |   C[i * L + j] += A[i * L + k] * B[j + k * L]
5 |   |   end
6 |   end
7 end
```

Algorithm 1 CPU Multiplication Algorithm

Where M, N, and L are the number of rows in matrix A, number of columns in matrix B, and number of rows and columns in matrix B and A, respectively. All matrices are presented as a vector. To adjust libMesh to the hybrid approach considered in this paper above, the code was replaced with the following one:

```
1 if CUDA_ENABLED == 1 then
2 |   CUDA_matrixMultiplication
3 else
4 |   ...
5 end
```

Algorithm 2 Hybrid GPU/CPU approach

When *CUDA_ENABLED* value is 1, which is the default value, matrix multiplication is performed by GPU. Otherwise the original algorithm, represented by ... is executed. The general function *CUDA_matrixMultiplication* is responsible for setting the environment for the CUDA kernel, invoking it, and gathering results. The first stage is converting libMesh matrices to standard pointer based arrays. The process of conversion consists simply from assigning memory to the pointer and rewriting data from the libMesh matrix to standard array. Before *CUDA_matrixMultiplication* is terminated, reverse translation has to be performed. The process is very similar, however this time values are read from standard arrays and assigned to appropriate elements in the libMesh matrix. A very important part of *CUDA_matrixMultiplication* is allocating the memory on GPU, populating data from the host memory to GPU memory and from GPU to host memory, and releasing the memory on GPU. This can be accomplished using standard CUDA functions called *cudaMalloc*, *cudaMemcpy*, and *cudaFree*, respectively. The most important part of the proposed function is invoking the CUDA kernel that is responsible for performing matrix multiplication. The modified version of Tiled CUDA Multiplication Algorithm presented in [7] was used in this paper. Algorithm 3 presents the pseudo code of the algorithm.

Similar to the previous algorithm, matrices A, B, and C are given by vectors. Value N is the number of columns in matrix B; and L is the number of rows in matrix A and columns in matrix B. Base on thread and block indices, coordinates of elements of matrices A and B in the GPU global memory are calculated. Calculated coordinates are coordinates of a tile in matrices A and B. Data within a tile forms submatrices *tA* and *tB*. After that each thread copies a single element of matrices A and B to local shared memory. As a result small submatrices *tA* and *tB* are created. Both submatrices are located in fast local memory that is shared only by threads within one block. The algorithm uses *tA* and *tB* to accumulate values in *tC*. After that the tile is moved to the right along the row and down along the column in matrices A and B respectively. The process is repeated with a new location of a tile.

4 Experiments

4.1 Experiment Description

To examine the effectiveness of the proposed hybrid approach, an experiment was performed. Performance of the proposed method was compared to the performance of the standard CPU approach. In the experiment two multiphysics heat conduction problems, diffusion and diffusion with convection were considered. A weak form of the PDE that describes the convection diffusion problem (and

Algorithm 3 Tiled CUDA Multiplication Algorithm

```
idA = (blockId.y * blockSize + threadId.y) * L + threadId.x
idB = blockId.x * blockSize + threadId.x + threadId.y * N
result = 0
Create shared tA and tB of size blockSize * blockSize
for i = 0 to L/blockSize do
    tA[threadId.x + threadId.y * blockSize] = A[idA]
    tB[threadId.x + threadId.y * blockSize] = B[idB]
    idA+ = blockSize
    idB+ = blockSize * N
    synchronize
    for j = 0 to blockSize do
        result+ =
        tA[threadId.x * blockSize + j] * tB[threadId.x + j * blockSize]
    end
    synchronize
end
idC = (blockId.y * blockSize + threadId.y) * L
idC+ = blockId.x * blockSize + threadId.x
C[idC] = result
```

therefore diffusion only also) is given by the equations below.

$$-\nabla \cdot k\nabla u + \beta \cdot \nabla u = f \tag{1}$$

Where k is diffusivity, β is the vector field of velocity, $-\nabla \cdot k\nabla u$ is diffusion, and $\beta \cdot \nabla u$ is convection. After moving f to the left side of equation, multiplying both sides by the shape function ψ, and integrating the equation over the domain Ω, Equation 2 is obtained [4].

$$-\int_{\Omega} \psi(\nabla \cdot k\nabla u) + \int_{\Omega} \psi(\beta \cdot \nabla u) - \int_{\Omega} \psi f = 0 \tag{2}$$

Applying the divergence theorem to Equation 2 transforms it to Equation 3.

$$\int_{\Omega} \nabla\psi \cdot k\nabla u - \int_{\partial\Omega} \psi(k\nabla u \cdot \widehat{u}) + \int_{\Omega} \psi(\beta \cdot \nabla u) - \int_{\Omega} \psi f = 0 \tag{3}$$

When represented in terms of multiphysics kernels and boundary conditions, Equation 3 has the following form [4].

$$(\nabla\psi, k\nabla u) - \langle \psi, k\nabla n \cdot \widehat{n} \rangle + (\psi, \beta \cdot \nabla u) - (\psi, f) = 0 \tag{4}$$

Where $(\nabla\psi, k\nabla u)$, $(\psi, \beta \cdot \nabla u)$, and (ψ, f) are multiphysics kernels and $\langle \psi, k\nabla n \cdot \widehat{n} \rangle$ is a boundary condition [4].

These multiphysics phenomena were applied to four empty copper cylinders sealed from the top and bottom. The cylinders meshes consist of cubic cells and are given by 250 000 points, 500 000 points, and 750 000 points respectively.

The experiments were performed on a standard desktop running the Ubuntu 12.04 64-bit Linux operating system. The computer is equipped with quad core Intel Xenon processors. Each core works at a 2.8 GHz frequency. The computer has 3.9 GB of memory and is a host for the NVidia Quadro 5000 GPU. According to the device datasheet, the GPU provides 352 CUDA cores clocked with 513 MHz each. The device is equipped with 2.5 GB of GDDR5 memory.

4.2 Results

Table 1 contains results obtained from the performed experiments. The first column indicates the number of points the object consists of and the second column specifies the phenomena (Diff for diffusion and Con-Diff for convection diffusion). Each simulation was repeated 100 times to reduce the influence of random factors (e.g. responding for requests from other applications) that may affect performance. The speed-up factors obtained were averaged and the standard deviation was calculated.

As it can be seen in cases in which the number of points was relatively small, CPU approach outperforms the proposed Hybrid approach. This is caused by an additional steps that are related to the GPU/CPU approach. The libMesh data structure has to be translated to standard arrays, memory on GPU has to be allocated, and data has to be copied.

Table 1 Comparison of standard and hybrid approach performance

Points	Physics	Speed-up factor (t_{CPU}/t_{Hybrid})			
		min	max	avg	std. dev.
250 000	Diff	0.828	0.862	0.853	0.010
500 000	Diff	1.068	1.074	1.071	0.009
750 000	Diff	1.149	1.170	1.154	0.009
250 000	Con-Diff	0.932	0.942	0.938	0.011
500 000	Con-Diff	1.029	1.039	1.034	0.008
750 000	Con-Diff	1.128	1.151	1.132	0.009

Therefore, although execution of the kernel may be faster than the multiplication process performed by CPU, together with the mentioned overhead, the overall performance is worse. Nevertheless, when the number of points grows, the proposed approach is more efficient. For both meshes with 500 000 points and 750 000 points, the hybrid approach reduces the total simulation time by over 1.03 (3%). It can be also observed that with the increase in number of points the speed-up factor grows. Therefore it can be assumed that the performance increase would be even greater for a mesh that consists of 1 000 000 or more points. A reasonable idea to test this hypothesis would be to implement a mechanism that tracks simulation and record profiling. These outcomes could then be used to decide whether to use the hybrid or CPU approach.

The obtained results also show that the complexity of multiphysics phenomena has an influence on performance. For the same model the speed-up factor obtained for a simple diffusion phenomena is definitely larger than for the convection diffusion phenomena. Therefore, it may be assumed that for complex multiphysics problems, the difference in time consumption between matrix multiplication and other parts of the code is getting smaller. As a result, in the worst-case scenario, the speed-up factor will be very close to one. However it will never drop below one so the hybrid approach would be at worst as good as the CPU approach. This observation draws the conclusion that other heavy computational and parallel parts of the code should be moved to GPU.

5 Conclusions

In this paper a hybrid method to perform multiphysics simulations was presented. The approach keeps the standard method of executing simulation in parallel on high performance computers (HPC). However parts of the code, which are beneficial to execute by a larger number of threads, are moved to GPU. GPU is acting as an underlying computational unit that is used by a host CPU. The proposed method can be applied as a modification of any multiphysics simulation library and does not enforce alterations of existing applications that are using the library.

The presented approach was applied to the FE library called libMesh which is a module in the open source multiphysics framework MOOSE. In this portion of the research, the part of code that was ported to GPU is a matrix multiplication. Every time threads executed by CPU reach the matrix multiplication function, matrices are copied to GPU memory, a kernel is launched, and the results are copied back from the GPU memory. To test this approach one of the most computational demanding operations in multiphysics simulation, which is matrix multiplication, was ported to GPU. The CUDA algorithm selected to perform the multiplication is a Tiled CUDA Multiplication Algorithm. Modification was evaluated by performing an experiment. The experiment consisted of two multiphysics simulations - diffusion and convection diffusion. Physics phenomena were applied to four empty cooper cylinders sealed from top and bottom. Each cylinder is represented by a mesh that consists of a different number of points. Results shows that the hybrid approach decreases the time required by the simulation. The larger point mesh is composed from the bigger speed-up is obtained by executing matrix multiplication by GPU. Nevertheless it can be also observed that the speed-up factor decreases with growth in complexity of the physics problem. When more complex physics is simulated then other parts of code become more and more time demanding whereas time required for matrix operations remains almost unchanged.

Further research may include extraction and porting other heavy computational and highly parallel parts of the code. A good idea would be to design and implement a mechanism that tracks the parameters of simulation and record the used method and performance. Results would help to determine which approach, Hybrid or Standard, offers better performance.

References

1. Krol, D., Zydek, D.: Solving PDEs in Modern Multiphysics Simulation Software. In: 2013 IEEE International Conference on Electro/Information Technology (EIT 2013), pp. 1–6., IEEE Computer Society Press, 2013, doi: 10.1109/EIT.2013.6632675
2. Chmaj, G., Zydek, D.: Software Development Approach for Discrete Simulators. In: 21st International Conference on Systems Engineering (ICSEng 2011), pp. 273-278, IEEE Computer Society Press, 2011, doi: 10.1109/ICSEng.2011.56
3. Zimmerman, W. B. J.: Multiphysics Modeling With Finite Element Method. World Scientific, Series on Stability, Vibration, and Control of Systems, Series A, Vol 18, 2008
4. Idaho National Laboratory: MOOSE Workshop. 2014
5. Chmaj, G., Walkowiak, K.: Decision Strategies for a P2P Computing System. Journal of Universal Computer Science, Vol. 18, N. 5, pp. 599–622, 2012, doi: 10.3217/jucs-018-05-0599

6. Zydek, D., Chmaj, G., Chiu, S.: Modeling Computational Limitations in H-Phy and Overlay-NoC Architectures. The Journal of Supercomputing, 2013, doi: 10.1007/s11227-013-0932-9
7. Krol D., Zydek D., Selvaraj H.: Matrix Multiplication in Multiphysics Systems Using CUDA. In: International Conference on Systems Science 2013 (ICSS 2013), pp. 493–502, 2013, doi: 10.1007/978-3-319-01857-7_48
8. TOP 500 Supercomputer Ranking webpage, www.top500.com
9. DAmbrosio, D., Spataro, W., Parise, R., Rongo, R., Filippone, G., Spataro, D., Iovine, G., Marocco, D.: Lava ow modeling by the Sciara-fv3 parallel numerical code. In: Parallel, Distributed and Network-Based Processing (PDP), 22nd Euromicro International, pp. 330–338, 2014, doi: 10.1109/PDP.2014.68
10. Kurhade, A., Thakare, A., Phadke, A.: CUDA Accelerated Fast Training of Locally Connected Neural Pyramid Using YIQ Color Coding. In: Advance Computing Conference (IACC) 2014 IEEE International, pp. 1–6, 2014, doi: 10.1109/IAdCC.2014.6779482
11. Lee, J.-W., Kim, B., Yoon, K.-S.: CUDA-based JPEG2000 Encoding Scheme. In: Advanced Communication Technology (ICACT), 2014 16th International, pp. 671–674, 2014, doi: 10.1109/ICACT.2014.6779047
12. Hagen, T. R., Lie, K.-A., Natvig, J. R.: Solving the Euler Equations on Graphics Processing Units. Lecture Notes in Computer Science, Vol 3994, pp. 220–227, 2006
13. Elsen, E., LeGresley, P., Darve, E.: Large calculation of the ow over a hypersonic vehicle using a GPU. Journal of Computational Physics, Vol 227, N 24, pp. 10148–10161 2008
14. Brandvik, T., Pullan, G.: Acceleration of a 3D Euler solver using commodity graphics hardware. In: 46th AIAA Aerospace Sciences Meeting and Exhibit, 2008
15. Phillips, E. H., Zhang, Y., Davis, R. L., Owens, J. D.: Rapid aerodynamic performance prediction on a cluster of graphics processing units. In: 47th aerospace sciences meeting and exhibit, 2009
16. Chen, J., Joo, B., Watson, W., Edwards, R.: Automatic ofoading C++ expression tem-plates to CUDA enabled GPUs. In: Parallel and distributed processing symposium workshops and PhD forum, pp. 2359–2368, 2012, doi:10.1109/IPDPSW.2012.293
17. Enmyren, J., Kessler, C. W.: SkePU: A multi-backend skeleton programming library for multi-GPU systems. In: Proc 4th int workshop on high-level parallel programming and applications, 2010
18. Corrigan, A., Camelli, F., Lohner, R., Mut, F.: Semi-automatic porting of a large-scale Fortran CFD code to GPUs. International Journal for Numerical Methods in Fluids, Vol 69, N 6, pp. 314–331, 2011
19. Chandar, D. D. J., Sitaraman, J., Mavriplis, D.: CU++: an object oriented framework for computational uid dynamics applications using graphics processing units. The Journal of Supercomputing, Vol 67, N 1, pp. 47–68, 2014, doi: 10.1007/s11227-013-0985-9

ERRATUM 1

Chapter 25
Multi-Agent Reinforcement Learning Control for Ramp Metering

Ahmed Fares and Walid Gomaa

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 330,
DOI 10.1007/978-3-319-08422-0_25, © Springer International Publishing Switzerland 2015

DOI 10.1007/978-3-319-08422-0_131

The affiliation of the second author W. Gomaa was incorrect. The correct information is given below:

Computer Science and Engineering Department
Egypt-Japan University for Science and Technology (E-JUST)
Currently on leave from the faculty of Engineering, Alexandria University
Alexandria, Egypt

A. Fares (*)
Computer Science and Engineering Department, Egypt-Japan
University for Science and Technology (E-JUST), Alexandria, Egypt
e-mail: ahmed.fares@ejust.edu.eg

W. Gomaa
Currently on leave from the faculty of Engineering, Alexandria
University, Alexandria, Egypt
e-mail: walid.gomaa@ejust.edu.eg

The online version of the original chapter can be found under
http://dx.doi.org/10.1007/978-3-319-08422-0_25

ERRATUM 2

Progress in Systems Engineering

Henry Selvaraj, Dawid Zydek and Grzegorz Chmaj

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 330,
DOI 10.1007/978-3-319-08422-0, © Springer International Publishing Switzerland 2015

DOI 10.1007/978-3-319-08422-0_132

The volume numbers were captured incorrectly as 330. The correct volume number should be 366 and this has been updated in the print edition (on the cover) and in the electronic renditions.

The online version of the original book can be found under
http://dx.doi.org/10.1007/978-3-319-08422-0