

Khaled Elleithy
Tarek Sobh
Editors

Innovations and Advances in Computer, Information, Systems Sciences, and Engineering

Part 1

Lecture Notes in Electrical Engineering

Volume 152

For further volumes:
<http://www.springer.com/series/7818>

Khaled Elleithy · Tarek Sobh
Editors

Innovations and Advances in Computer, Information, Systems Sciences, and Engineering

Editors

Khaled Elleithy
School of Engineering
University of Bridgeport
Bridgeport, CT
USA

Tarek Sobh
School of Engineering
University of Bridgeport
Bridgeport, CT
USA

ISSN 1876-1100

ISBN 978-1-4614-3534-1

DOI 10.1007/978-1-4614-3535-8

Springer New York Heidelberg Dordrecht London

ISSN 1876-1119 (electronic)

ISBN 978-1-4614-3535-8 (eBook)

Library of Congress Control Number: 2012940241

© Springer Science+Business Media New York 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

Preface

This book includes the proceedings of the International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 2011). The proceedings are a set of rigorously reviewed world-class manuscripts presenting the state of international practice in Innovative Algorithms and Techniques in Automation, Industrial Electronics, and Telecommunications.

CISSE 2011 is a high-caliber research for research conferences that were conducted online. CISSE 2011 received 260 paper submissions and the final program included 107 accepted papers from more than 80 countries, representing the six continents. Each paper received at least two reviews, and authors were required to address review comments prior to presentation and publication.

Conducting CISSE 2011 online presented a number of unique advantages, as follows:

- All communications among the authors, reviewers, and conference organizing committee were done online, which permitted a short 6-week period from the paper submission deadline to the beginning of the conference.
- PowerPoint presentations, final paper manuscripts were available to registrants for 3 weeks prior to the start of the conference.
- The conference platform allowed live presentations by several presenters from different locations, with the audio, and PowerPoint transmitted to attendees throughout the Internet, even on dial-up connections. Attendees were able to ask both audio and written questions in a chat room format, and presenters could mark up their slides as they deemed fit.
- The live audio presentations were also recorded and distributed to participants along with the powerpoint presentations and paper manuscripts within the conference DVD.

The conference organizers and we are confident that you will find the papers included in this volume interesting and useful. We believe that technology will continue to infuse education thus enriching the educational experience of both students and teachers.

Bridgeport, CT, December 2011

Khaled Elleithy
Tarek Sobh

Acknowledgments

The 2011 International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 2011) and the resulting proceedings could not have been organized without the assistance of a large number of individuals. CISSE was founded by Professors Tarek Sobh and Khaled Elleithy in 2005, and they set up mechanisms that put it into action. Andrew Rosca wrote the software that allowed conference management, and interaction between the authors and reviewers online. Mr. Tudor Rosca managed the online conference presentation system and was instrumental in ensuring that the event met the highest professional standards. We also want to acknowledge the roles played by Sarosh Patel and Ms. Susan Kristie, our technical, and administrative support team.

The technical co-sponsorship provided by the Institute of Electrical and Electronics Engineers (IEEE) and the University of Bridgeport is gratefully appreciated. We would like to express our thanks to Prof. Toshio Fukuda, Chair of the International Advisory Committee and the members of Technical Program Committees.

The excellent contributions of the authors made this world-class document possible. Each paper received two to four reviews. The reviewers worked tirelessly under a tight schedule and their important work is gratefully appreciated. In particular, we want to acknowledge the contributions of the following individuals: Ashraf Abdelwahed, Khalid Aboalayon, Munther Abualkibash, Tamer Abu-Khalil, Sumaya Abusaleh, Ahmad Abushakra, Mohannad Abuzneid, Naser Alajmi, Ibrahim Alkore Alshalabi, Muder Almi'ani, Anas Al-okaily, Aziz Alotaibi, Amer Al-Rahayfeh, Mohammad Rauji, Tariq Alshugran, Fahad Alswaina, Aladdin Alzubi, Mohamed Ben Haj Frej, Ying-ju Chen, Richard Colon, Khaled Elleithy, Ali El-Rashidi, Ahmed ElSayed, Mohammed Ali Eltaher, Eugene Gerety, Manan Joshi, Zakareya Lasefr, Youming Li, Ramadhan Mstafa, Ammar Odeh, Abdul Razaque, Andriy Shpylychyn, and Ajay Shrestha.

Bridgeport, CT, January 2012

Khaled Elleithy
Tarek Sobh

Contents

| | | |
|----------|---|-----------|
| 1 | Change Rate Concepts and their Realization in the MM&S: A Computer Program for Modeling and Simulation of Dynamic Systems | 1 |
| | Nguyen Van Sinh | |
| 2 | A Software Architecture for Inventory Management System . . . | 15 |
| | Taner Arsan, Emrah Başkan, Emrah Ar and Zeki Bozkuş | |
| 3 | Libraries Opt for More Online Sources | 29 |
| | Zeenath Reza Khan and Sreejith Balasubramanian | |
| 4 | Emerging Threats, Risk and Attacks in Distributed Systems: Cloud Computing | 37 |
| | Isabel Del C. Leguías Ayala, Manuel Vega and Miguel Vargas-Lombardo | |
| 5 | Cognitive Antenna System for Sustainable Adaptive Radio Interfaces | 53 |
| | Ligia Cremene and Nicolae Crişan | |
| 6 | Introducing the Concept of Information Pixels and the Storing Information Pixels Addresses Method as an Efficient Model for Document Storage | 63 |
| | Mohammad A. ALGhalayini | |
| 7 | Introducing the Concept of Back-Inking as an Efficient Model for Document Retrieval (Image Reconstruction). | 89 |
| | Mohammad A. ALGhalayini | |

| | | |
|-----------|--|------------|
| 8 | Automating the Transformation From a Prototype to a Method of Assembly | 99 |
| | Yuval Cohen, Gonen Singer, Maya Golan and Dina Goren-Bar | |
| 9 | Collaborative and Non-Collaborative Dynamic Path Prediction Algorithm for Mobile Agents Collision Detection with Dynamic Obstacles in 3D Space. | 107 |
| | Elmir Babovic | |
| 10 | Website Analysis of Top 100 Most Valuable Companies in Romania. | 121 |
| | Lavinia D. Rusu and Liciniu A. Kovács | |
| 11 | Comparison of PI and Fractional PI Controllers on a Hydraulic Canal Using Pareto Fronts | 135 |
| | Y. Chang | |
| 12 | Remote Sensing Investigation of Red Mud Catastrophe and Results of Image Processing Assessment | 149 |
| | J. Berke, V. Kozma-Bognár, P. Burai, L. D. Kovács, T. Tomor and T. Németh | |
| 13 | 802.11e QoS Performance Evaluation | 157 |
| | Yunus Simsek and Hetal Jasani | |
| 14 | Evaluation of Different Designs to Represent Missing Information in SQL Databases | 173 |
| | Erki Eessaar and Elari Saal | |
| 15 | Mobile English Learning System: A Conceptual Framework for Malaysian Primary School. | 189 |
| | Saipunidzam Mahamad, Fatimah Annor Ahmad Rashid, Mohammad Noor Ibrahim and Rozana Kasbon | |
| 16 | Dynamic Cache Miss-Rate Reduction | 199 |
| | Mazen AbuZaher, Bayan Alayoubi, Basma Alefeshat and Abdelwadood Mesleh | |
| 17 | Agent Simulation Group on the Robocup 3D Realization of Basic Motions. | 205 |
| | Min Zhou, Jia Wu, Hao Zheng, Xiaoming Liu and Renhao Zhou | |
| 18 | Key Generations Model for Mobile Cryptosystems. | 215 |
| | Rushdi Hamamreh | |

| | | |
|-----------|--|------------|
| 19 | Development of Stakeholder Oriented Corporate Information Security Objectives | 227 |
| | Margareth Stoll | |
| 20 | Stakeholder Oriented Information Security Reporting | 241 |
| | Margareth Stoll | |
| 21 | Experimenting with Watchdog Implementation on a Real-Life Ad hoc Network: Monitoring Selfish Behavior | 255 |
| | Tirthankar Ghosh and Tian Hou | |
| 22 | Power Consumption Evaluation for Cooperative Localization Services | 267 |
| | Patrick Seeling | |
| 23 | A Modified Banker's Algorithm | 277 |
| | Youming Li | |
| 24 | Courses Enrollment Pattern Analysis | 283 |
| | Nur Fatihah Abdul Rahim, Shakirah Mohd Taib and Saipunidzam Mahamad | |
| 25 | Integration of Safety and Smartness Using Cloud Services: An Insight to Future | 293 |
| | Neha Tekriwal, Madhumita and P. Venkata Krishna | |
| 26 | A Versioning Subsystem of Metamodeling System | 305 |
| | Rünno Sgirka | |
| 27 | Difficulties in Understanding Object Oriented Programming Concepts. | 319 |
| | Soly Mathew Biju | |
| 28 | Real-Time System for Monitoring and Analyzing Electrocardiogram on Cell Phone | 327 |
| | O. Muñoz-Ramos, O. Starostenko, V. Alarcon-Aquino and C. Cruz-Perez | |
| 29 | Research of Camera Track Based on Image Matching | 339 |
| | Yuan Wang | |

| | | |
|-----------|--|------------|
| 30 | Curriculum Design Change of the Industrial Engineering BA Program | 349 |
| | Eszter Bogdány, Ágnes Balogh, Gabriella Cerhádi, Tibor Csizmadia and Réka Polák-Weldon | |
| 31 | Comparing Two Methods of Sound Spatialization: Vector-Based Amplitude Panning (VBAP) Versus Linear Panning (LP) | 359 |
| | Jonathan Cofino, Armando Barreto and Malek Adjouadi | |
| 32 | Contrast Enhancement in Image Pre-Compensation for Computer Users with Visual Aberrations | 371 |
| | Jian Huang, Armando Barreto, Malek Adjouadi and Miguel Alonso | |
| 33 | Interaction with 3D Environments Using Multi-Touch Screens | 381 |
| | Francisco Ortego, Naphtali Rishe, Armando Barreto and Melek Adjouadi | |
| 34 | TCP with Extended Window Scaling | 393 |
| | Michal Olšovský and Margaréta Kotočová | |
| 35 | Offering SaaS as SOA Services | 405 |
| | Ali Bou Nassif and Miriam A. M. Capretz | |
| 36 | Using Conceptual Mini Games for Learning: The Case of “The Numbers’ Race” (TNR) Application | 415 |
| | C. T. Panagiotakopoulos and M. E. Sarris | |
| 37 | Visual Cryptography Based on Optical Image Projection | 431 |
| | Rita Palivonaite, Algimantas Aleksa and Minvydas Ragulskis | |
| 38 | A New Service Offered by Digital Radio for Vehicle Drivers . . . | 443 |
| | Cabani Adnane and Mouzna Joseph | |
| 39 | Separation of Concerns in Extensible Control Systems | 451 |
| | Martin Rytter and Bo Nørregaard Jørgensen | |
| 40 | Illicit Image Detection: An MRF Model Based Stochastic Approach | 467 |
| | Mofakharul Islam, Paul Watters, John Yearwood, Mazher Hussain and Lubaba A. Swarna | |

| | | |
|-----------|---|------------|
| 41 | Illicit Image Detection Using Erotic Pose Estimation Based on Kinematic Constraints | 481 |
| | Mofakharul Islam, Paul Watters, John Yearwood, Mazher Hussain and Lubaba A. Swarna | |
| 42 | Energy Efficient Public Key Cryptography in Wireless Sensor Networks. | 497 |
| | Vladimir Cervenka, Dan Komosny, Lukas Malina and Lubomir Mraz | |
| 43 | Implementation of VLSB Steganography Using Modular Distance Technique. | 511 |
| | Sahib Khan and Muhammad Haroon Yousaf | |
| 44 | A Graphic User Interface for H-Infinity Static Output Feedback Controller Design | 527 |
| | J. Gadewadikar, K. Horvat and O. Kuljaca | |
| 45 | Active Contour Texture Segmentation in Modulus Wavelet Feature Spaces | 537 |
| | Ashoka Jayawardena and Paul Kwan | |
| 46 | A Framework for Verification of Fuzzy Rule Bases Representing Clinical Guidelines. | 545 |
| | M. Esposito and D. Maisto | |
| 47 | Communication Impact on Project Oriented Teaching in Technology Supported Education | 559 |
| | Martin Misut and Katarina Pribilova | |
| 48 | A Failure Modes and Effects Analysis of Mobile Health Monitoring Systems. | 569 |
| | Marcello Cinque, Antonio Coronato and Alessandro Testa | |
| 49 | SemFus: Semantic Fusion Framework Based on JDL. | 583 |
| | Havva Alizadeh Noughabi, Mohsen Kahani and Behshid Behkamal | |
| 50 | Development of GUI Based Test and Measurement Facilities for Studying Properties of MOS Devices in Clean Room Environment. | 595 |
| | Shaibal Saha and Supratik Chakraborty | |

| | | |
|-----------|--|------------|
| 51 | Prediction of Failure Risk Through Logical Decision Trees in Web Service Compositions | 609 |
| | Byron Portilla-Rosero, Jaime A. Guzmán and Giner Alor-Hernández | |
| 52 | SEC-TEEN: A Secure Routing Protocol for Enhanced Efficiency in Wireless Sensor Networks | 621 |
| | Alkore Alshalabi Ibrahim, Abu Khalil Tamer and Abuzneid Abdelshakour | |
| 53 | An Integration of UML-B and Object-Z in Software Development Process | 633 |
| | Mehrnaz Najafi and Hassan Haghighi | |
| 54 | Algorithm for Dynamic Traffic Rerouting and Congestion Prevention in IP Networks | 649 |
| | Martin Hrubý, Margaréta Kotočová and Michal Olšovský | |
| 55 | Fovea Window for Wavelet-Based Compression. | 661 |
| | J. C. Galan-Hernandez, V. Alarcon-Aquino, O. Starostenko and J. M. Ramirez-Cortes | |
| 56 | Energy Aware Data Compression in WSN. | 673 |
| | Roshanak Izadian and Mohammad Taghi Manzuri | |
| 57 | Energy Consumption Text and Image Data Compression in WSNs | 683 |
| | Roshanak Izadian and Mohammad Taghi Manzuri | |
| 58 | New QoS Framework for Mobile Ad hoc Networks Based on the Extension of Existing QoS Models. | 697 |
| | Peter Magula and Margaréta Kotočová | |
| 59 | Method for Data Collection and Integration into 3D Architectural Model | 707 |
| | L. Kurik, V. Sinivee, M. Lints and U. Kallavus | |
| 60 | Statistical Analysis to Export an Equation in Order to Determine Heat of Combustion in Blends of Diesel Fuel with Biodiesel | 719 |
| | C. G. Tsanaktsidis, V. M. Basileiadis, K. G. Spinthoropoulos, S. G. Christidis and A. E. Garefalakis | |
| 61 | The Retail Banking Adverse Selection: RCBS Calculator Solution. | 729 |
| | M. Hedvicakova, I. Soukal and J. Nemecek | |

| | | |
|-----------|---|------------|
| 62 | Project Management in Public Administration Sector | 741 |
| | M. Hedvicakova | |
| 63 | Access Point Checking to Improve Security in Wireless Infrastructure Networks | 751 |
| | Ammar Odeh and Miad Faezipour | |
| 64 | Comparison of Fractional PI Controller with Classical PI using Pareto Optimal Fronts | 763 |
| | O. J. Moraka | |
| 65 | A Pattern-Based Approach for Representing Condition-Action Clinical Rules into DSSs | 777 |
| | A. Minutolo, M. Esposito and G. De Pietro | |
| 66 | Authorization of Proxy Digital Signature in Workflow Systems | 791 |
| | Samir Fazlagic and Narcis Behlilovic | |
| 67 | Semi-Agile Approach to Software Development Process | 801 |
| | Deniss Kumlander | |
| 68 | The Influence of Student Body-Talk Reaction in Formulating Effective Teaching Strategy | 811 |
| | Ahmad Sofian Shminan and Runhe Huang | |
| 69 | Interactive Mind Map Desktop Widget: A Proposed Concept. . . | 829 |
| | Tan Wei Xuan, Shakirah Mohd Taib and Saipunidzam Mahamad | |
| 70 | An Algorithm for Replication in Distributed Databases | 839 |
| | Adrian Runceanu and Marian Popescu | |
| 71 | General Dispatching of Lignite Mining Pit. | 849 |
| | Constantin Cercel and Florin Grofu | |
| 72 | Towards Improving the StatscanTM X-Ray Image Quality through Sliding-Mode Control of the C-Arm | 857 |
| | M. Esmail, M. Tsoeu and L. John | |
| 73 | Mechanical Energy Conversion to Electromagnetic Energy for Magnetic Fluids: Theoretical Fundaments and Applications. | 871 |
| | Aurel-George Popescu and Adrian Runceanu | |

| | | |
|-----------|---|-------------|
| 74 | Initial Steps Towards Distributed Implementation of M-Urgency | 883 |
| | Shivsubramani Krishnamoorthy, Arun Balasubramanian and Ashok K. Agrawala | |
| 75 | Optimal Selection of Components in Fault Detection Based on Principal Component Analysis | 901 |
| | Patricia Helen Khwambala | |
| 76 | E-Learning Environment Identification System: Error Injection and Patterns Dynamics | 917 |
| | Deniss Kumlander | |
| 77 | Energy Consumption by Deploying a Reactive Multi-Agent System Inside Wireless Sensor Networks | 925 |
| | Alcides Montoya and Demetrio Ovalle | |
| 78 | Network Intrusion Detection System Based on SOA (NIDS-SOA): Enhancing Interoperability Between IDS | 935 |
| | Wagner Elvio de Loiola Costa, Denivaldo Lopes, Zair Abdelouahab and Bruno Froz | |
| 79 | Lyrebird: A Learning Object Repository Based on a Domain Taxonomy Model | 949 |
| | Ingrid Durley Torres, Jaime Alberto Guzman Luna and Jovani Alberto Jimenez Builes | |
| 80 | Design Process and Building Simulation | 961 |
| | Heitor da Costa Silva, Clarissa SartoriZiebell, Lennart Bertram Pöhls and Mariana Moura Bagnati | |
| 81 | Behavioral Models with Alternative Alphabets. | 975 |
| | Mohammed Lafi and Jackson Carvalho | |
| 82 | Watermark Singular-Values Encryption and Embedding in the Frequency Domain | 989 |
| | Chady El Moucary and Bachar El Hassan | |
| 83 | Business Intelligence Made Simple | 1001 |
| | Vasso Stylianou, Andreas Savva and Spyros Spyrou | |

| | | |
|-----------|---|-------------|
| 84 | A Process Model for Supporting the Management of Distance Learning Courses Through an Agile Approach | 1013 |
| | Amélia Acácia M. Batista, Zair Abdelouahab, Denivaldo Lopes and Pedro Santos Neto | |
| 85 | Numerical Modeling of Electromagnetic Induction Heating Process Using an Inductor with Constant Step Between Turns | 1027 |
| | Mihaela Novac, Ovidiu Novac, Mircea Gordan and Cornelia Gordan | |
| 86 | Satisficing-Based Approach to Resolve Feature Interactions in Control Systems. | 1039 |
| | Jan Corfixen Sørensen and Bo Nørregaard Jørgensen | |
| 87 | Properties Evaluation of an Approach Based on Probability-Possibility Transformation | 1053 |
| | M. Pota, M. Esposito and G. De Pietro | |
| 88 | Functional Verification of Class Invariants in CleanJava | 1067 |
| | Carmen Avila and Yoonsik Cheon | |
| 89 | Normalization Rules of the Object-Oriented Data Model | 1077 |
| | Vojtěch Merunka and Jakub Tůma | |
| 90 | Location Based Overlapping Mobility Aware Network Model | 1091 |
| | Abdul Razaque, Aziz Alotaibi and Khaled Elliethy | |
| 91 | Expert System for Evaluating Learning Management Systems Based on Traceability | 1103 |
| | E. Valdez-Silva, P. Y. Reyes, M. A. Alvarez, J. Rojas and V. Menendez-Dominguez | |
| 92 | How Variability Helps to Make Components More Flexible and Reusable | 1115 |
| | Yusuf Altunel and Abdül Halim Zaim | |
| 93 | Computing and Automation in the AEC Industry: Early Steps Towards a Mass Customized Architecture | 1129 |
| | Neander Silva, Diogo Santos and Ecilamar Lima | |

| | | |
|------------|---|-------------|
| 94 | Three Dimensional SPMD Matrix–Matrix Multiplication Algorithm and a Stacked Many-Core Processor Architecture | 1139 |
| | Ahmed S. Zekri | |
| 95 | Face: Fractal Analysis in Cell Engineering | 1151 |
| | K. P. Lam, D. J. Collins and J. B. Richardson | |
| 96 | A First Implementation of the Delay Based Routing Protocol. | 1165 |
| | Eric Gamess, Daniel Gámez and Paul Marrero | |
| 97 | Different Aspects of Data Stream Clustering | 1181 |
| | Madjid Khalilian, Norwati Mustapha, Md Nasir Sulaiman and Ali Mamat | |
| 98 | Teaching Computer Ethics Via Current News Articles. | 1193 |
| | Reva Freedman | |
| 99 | Designing and Integrating a New Model of Semi-Online Vehicle’s Fines Control System | 1205 |
| | Anas Al-okaily, Qassim Bani Hani, Laiali Almazaydeh, Omar Abuzaghle and Zenon Chaczko | |
| 100 | State Diagnosis of a Lignite Deposit by Monitoring its Surface Temperature with a Thermovision Camera | 1219 |
| | Alina Dinca | |
| 101 | A Framework Intelligent Mobile for Diagnosis Contact Lenses by Applying Case Based Reasoning | 1233 |
| | Eljilani Mohammed | |

Chapter 1

Change Rate Concepts and their Realization in the MM&S: A Computer Program for Modeling and Simulation of Dynamic Systems

Nguyen Van Sinh

Abstract The concept of “four element groups” means that all elements of a dynamic system can be divided into four groups: (1) constant elements, (2) state elements, (3) intermediate elements, (4) listed elements. This concept is realized in my MM&S-computer program with two facts. The first one is that four above mentioned element groups are correspondingly assigned four symbols: circle, square, rhombus, circle with three signs (these signs signal how we should handle this listed element at time point where its value has not been declared). These symbols are used to draw simulation scheme of interaction of the system elements. The concept of “change rate” means that every state element has a change rate as its attribute. Other elements of the system affect current state element by affecting its change rate. The current state element affects other elements of the system with its value. This concept is realized in my MM&S-computer program with the fact, that the links connect the state elements directly. In case if a state element has an incoming link, we understand that the change rate of this state element is being affected. Once change rate is an attribute of the state element, the value of the state element can be automatically used for the calculation of change rate. Realization of these above concepts make the simulation scheme of a dynamic system more clear and simple. This paper gives the reasoning for these concepts and also describes the model formats, model calculation in MM&S-computer program.

N. V. Sinh (✉)

Institute of Ecology and Biological Resources, 18 Hoang Quoc Viet,
Nghia Do, Cau Giay, Hanoi, Vietnam
e-mail: vansinh.nguyen@iebr.ac.vn

1.1 Introduction

Dynamic systems have been studied since long time ago [1–6]. The system elements have also been classified [1, 6, 7]. However the modeling and simulation software have been based on the black-box concept with inputs and outputs. An example is the widely used STELLA software of the ISEE SYSTEMS [8] (formerly High Performance Systems Corporation) [7]. The simulation diagram that is based on this black-box concept can not provide a good visualization of the system structure: even a real state element of the system must be presented here as a black-box with input and output; one can not recognize from the simulation diagram, whether an element is a constant or intermediate one, because they have the same symbol (a circle).

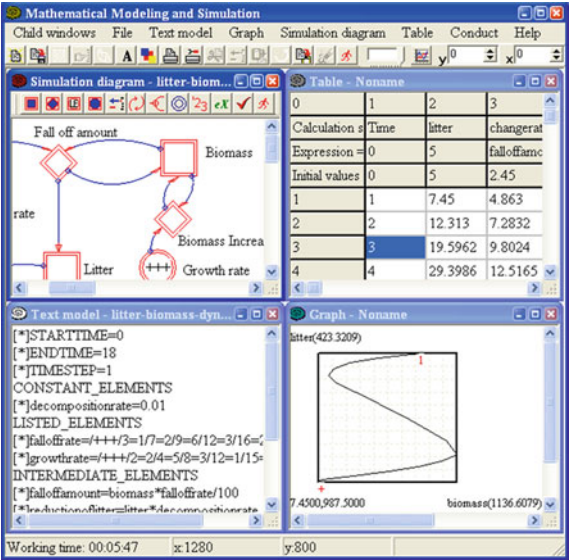
In a dynamic system we can find elements of different “mathematical nature”. Some elements do not change their value forever or at least during the time we observe the system. Some elements change their value but we can determine their value at any time by measuring, weighting, counting,..., though it is sometime very difficult. Some elements change their value and their value at a time can only be calculated from the value of the other elements. And finally, some elements change their value over time but their value are given at all or some time points of the time period when we observe the system (the values of the element are listed). Based on these facts a concept of “four element groups” can be formulated which means that all elements of a dynamic system can be divided into four groups:

1. Constant elements,
2. State elements,
3. Intermediate elements,
4. Listed elements.

Each state element changes its value over the time. The change rate of a state element can be negative or positive at a time. Not all the state elements have inflow and outflow. Instead, inflow and outflow are only typical for state elements of mechanical nature, the state elements of biological nature can grow. Based on these facts a concept of “change rate” can be formulated which means that every state element has a change rate as its attribute. Other elements of the system affect current state element by affecting its change rate and the current state element affects other elements of the system with its value.

These two concepts make the basis for the new design of the MM&S—a computer program for modeling and simulation of dynamic systems. The program is free available at the website of the Institute of Ecology and Biological Resources [9]. Further in this paper is the introduction to the new version of MM&S computer program as an explanation of how these above two concepts have been implemented.

Fig. 1.1 Interface of the MM&S



1.2 Materials and Methods

Delphi XE Professional Workstation ESD (item number: 2010111885211109) of the Embarcadero company [10] has been used to create MM&S computer program.

Four child windows have been designed for handling different tasks:

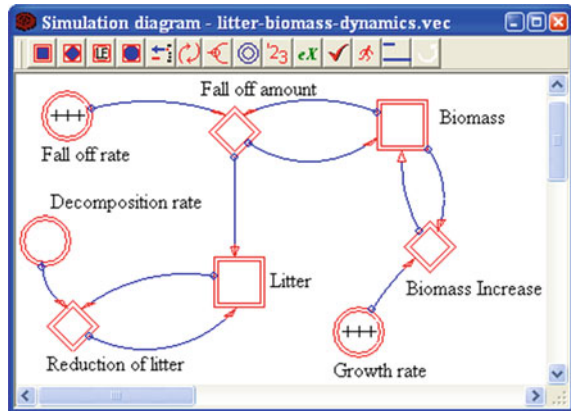
1. Child window with text editor for processing text model;
2. Child window with paint box for drawing and viewing simulation graphs;
3. Child window with paint box for processing simulation diagram and integrated model;
4. Child window with table grid for displaying results of simulation calculation.

All these child windows are managed by a multiple document interface—the main window (Fig. 1.1).

To visualize the system structure on the simulation diagram, four images have been used in MM&S to represent elements of the four above mentioned element groups: a square for state elements, a rhombus for intermediate elements, a circle for constant elements and a circle with three plus/minus signs inside for listed elements (Fig. 1.2). This is the implementation of the first concept (the “four element groups” concept) in simulation diagram. From the simulation diagram we can recognize the element group of a certain element through its symbol.

The links in the simulation diagram show the interactions between system elements. The state elements don’t have input and output, and incoming links of a state element specify the elements that affect its change rate. This means that the expression for calculating of the change rate of this state element has to include

Fig. 1.2 The simulation diagram child window



these affecting elements. The outgoing links of a state element specify the elements that are affected by this state element. This is the implementation of the second concept (the “change rate” concept) in simulation diagram.

1.3 User Interface of MM&S

1.3.1 Declared file formats of MM&S

MM&S declared following file formats:

1. The Text model format: the extension is ‘.ptm’; the symbol is:



2. The Simulation graph format: the extension is ‘.stm’; the symbol is:



3. The Simulation diagram format: the extension is ‘.vec’; the symbol is:



4. The Table format: the extension is ‘.tbl’, the symbol is



Once the MM&S program has been installed with using installation file, double clicking on a file name of one of this file types in Windows Explorer of Microsoft Corporation will cause MM&S starting with opening the file.

1.3.2 Buttons of MM&S and their functions

In MM&S we have two toolbars: one on the main window and the other on the simulation diagram child window. The functions of the most important buttons are as following:



Open button: display an open dialog box for choosing a file for its opening.



Save button: display an save dialog box for choosing a file for its opening.



Track bar: change the graph drawing speed (if graph child window is active).



Graph button: start the graph drawing procedure.



Up-down bar: change the size of the graph (if graph child window is active), or change the size of simulation scheme (if simulation scheme child window is active), or change the cell size of the table (if the table child window is active).



Symbol buttons: choose symbol to draw state element, intermediate element, listed element, constant element in the simulation scheme.



Link button: start drawing a link.



Delete button: start deleting an element in the simulation scheme.



Check button: start checking the integrated model of the current simulation scheme.



Run button: run a text model in a saved file (if clicking on this symbol on the toolbar of the main window) or run the integrated model of the simulation scheme (if clicking on this symbol on the toolbar of the currently active simulation scheme child window).



Export button: to export the integrated model of the simulation scheme to a file (*.ptm).



Switching button: to switch between displaying full element name and variable name.

1.3.3 Text Model Child Window

To create a new text model child window, from the main menu we choose ‘Text model/New window’. In this child window we can enter or open a model of text format and edit it. In a model each element is represented by a variable.

The format of a text model in MM&S (see Fig. 1.1: the left bottom child window) is as following:

- The first three lines are dedicated to declaring the time frame of the model: the time when the simulation starts, the time when the simulation ends, and the duration of a simulation step. Each line starts with brackets and one star inside, then the key words (STARTTIME, ENDTIME, and TIMESTEP), equal sign, and the time values.
- The next part of the model begins with a keyword ‘CONSTANT_ELEMENTS’. It signals that all constant elements of the model will be declared here. Every constant element is declared in one line: starting with brackets and one star inside, the variable name of the constant element, equal sign, and the value of the constant element at the end.
- The third part of the model begins with a keyword ‘LISTED_ELEMENTS’. All listed elements are declared in this part. Every listed element is declared in one line: starting with brackets and one star inside that are followed by the variable name of the listed element, equal sign, forward slash and three plus/minus signs, and the declaration of the values of the listed element. Each value declaration begins with one forward slash sign that is followed by the time value, equal sign, and then the value of the listed element. The three plus/minus signs signal the necessity of calculation of missing values based on the available ones for the time interval before the first available value (the first sign), for the time interval

between the first and the last available values (the second sign), and for the time interval after the last available value (the third sign). A plus sign means that the missing values should be calculated, the minus sign means that the value of zero should be assigned to the missing values in current interval.

- The fourth part of the model begins with a keyword ‘INTERMEDIATE_ELEMENTS’. All intermediate elements are declared in this part. Every intermediate element is declared in one line: starting with brackets and one star inside that are followed by the variable name of the intermediate element, equal sign, and the expression for calculating the intermediate variable at the end.
- The last part of the model begins with a keyword ‘STATE_ELEMENTS’. In this part every state element and its change rate are declared in two lines. The state element is declared in the first line: starting with brackets and one star inside, then the variable name of the state element, equal sign, and the initial value of the state element at the end. The second line is used for declaration of the change rate: starting with brackets and two stars inside, the name of the change rate (normally built by combining prefix ‘changerate_’ and the variable name of the state element), equal sign, and the expression for calculating the change rate at the end.

Following is the text model of Biomass-Litter dynamics:

```
[*]STARTTIME = 0
[*]ENDTIME = 18
[*]TIMESTEP = 1
CONSTANT_ELEMENTS
[*]decompositionrate = 0.01
LISTED_ELEMENTS
[*]falloffrate =/+++/3 = 1/7 = 2/9 = 6/12 = 3/16 = 2
[*]growthrate =/+++/2 = 2/4 = 5/8 = 3/12 = 1/15 = 4
INTERMEDIATE_ELEMENTS
[*]falloffamount = biomass*falloffrate/100
[*]reductionoflitter = litter*decompositionrate
[*]biomassincrease = biomass*growthrate/100
STATE_ELEMENTS
[*]litter = 5
[**]changerate_litter = falloffamount-reductionoflitter
[*]biomass = 1000
[**]changerate_biomass = biomassincrease-falloffamount
```

As we see, the format of text model has clearly shown the implementation of the “four element groups” and of the “change rate” concepts.

The text model should be saved to a file with extension ‘.ptm’. We can run the text model from a file by clicking on the run button of the main window toolbar or by choosing submenu item ‘Conduct/Run a model’.

Fig. 1.3 Dialog box for constant element in simulation diagram

Information on constant element

Element name:

Variable name:

Value: Order:

Elements to be affected:

1.3.4 Simulation Diagram Child Window

This child window has its own toolbar with buttons for drawing simulation diagram (see Fig. 1.2). To draw an element symbol we click on the button with that symbol and then click on the place in the child window where we want to put it. To draw a link between two elements we click the left mouse button on the affecting element and keep the left mouse button down while moving the mouse to the affected element, then release the left mouse button. The start of a link is marked with a blue circle and its end is marked with a red arrow. To move an element or an end of a link, click on it and hold the left mouse button down while moving the mouse. To delete an element symbol, we click on the delete button then click on element we want to delete. To delete a link, we click on delete button then then click on the begin or the end of the link.

To enter the model into the simulation diagram, we double click on the symbols of system elements: in the appeared dialog boxes we type in the element names, variable names, values and expressions for calculating variables and change rates (Figs. 1.3, 1.4, 1.5, 1.6).

The simulation diagram should be saved to a file with extension '.vec'. To check the completeness of the simulation scheme we click on the check button. Once the simulation diagram is complete, we can run the model from the simulation diagram by clicking run button on the toolbar of the simulation diagram child window, or we can export the model from the simulation diagram to a text model file by clicking export button.

Fig. 1.4 Dialog box for listed element in simulation diagram

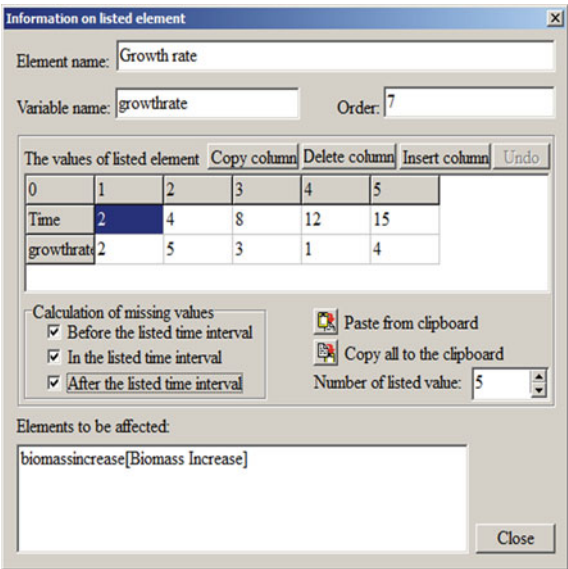
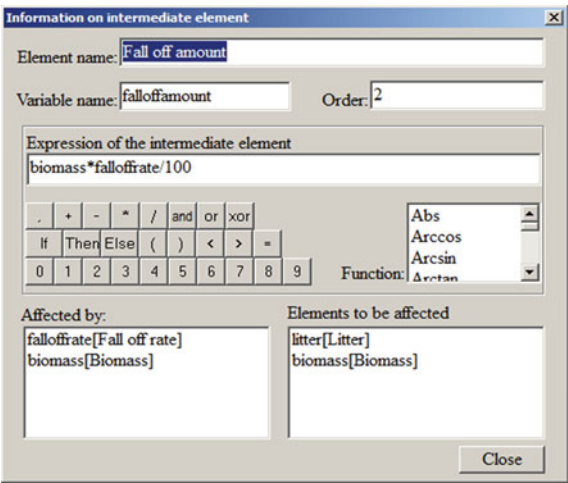


Fig. 1.5 Dialog box for intermediate element in simulation diagram



1.3.5 Simulation Graph Child Window

In MM&S we can simulate changes of the elements of a system by drawing time graph or phase graph (Figs. 1.7, 1.8). After simulation calculation we can draw graphs. To do this we can choose ‘Conduct/Draw a graph’ menu item or click on the graph button on the toolbar of the main window.

Several computer graphic techniques have been used in MM&S to enhance visualization: we can better trace the changes of system elements by changing the

Fig. 1.6 Dialog box for state element in simulation diagram

Fig. 1.7 Phase graph of biomass-litter dynamics

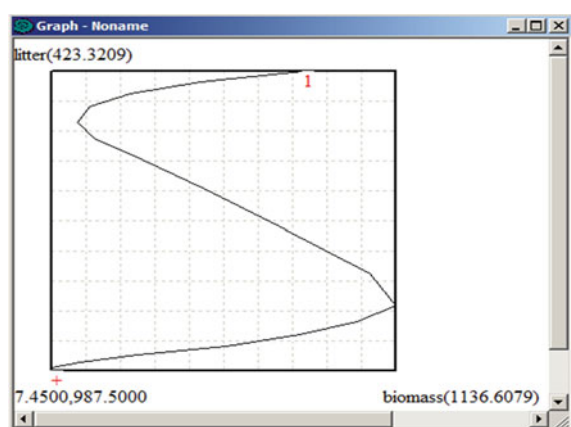


Fig. 1.8 Time graph of biomass and litter

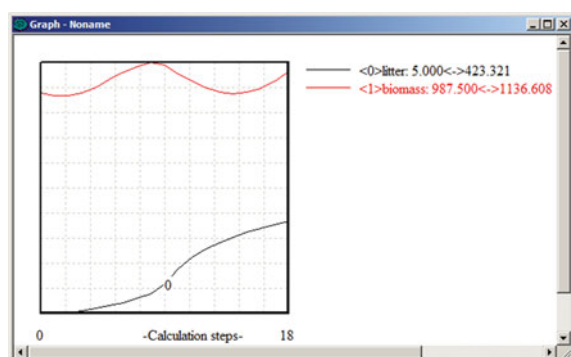
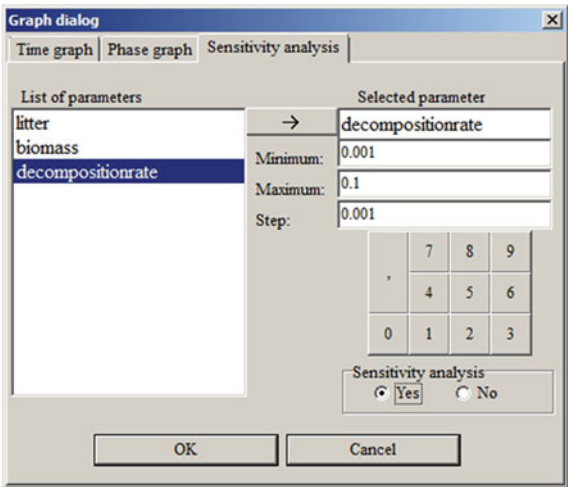


Fig. 1.9 Dialog box for drawing graph and sensitivity analysis



graph scale (with using up-down bar), by changing the graph drawing speed (with using track bar), or by pausing the graph drawing process (by clicking on the graph area while the graph is being drawn in slower mode).

By setting range for the initial value of a state element or for the value of a constant element (Fig. 1.9) before drawing graph, we can make analysis on sensitivity of other system elements to the changes of these elements. The software will conduct simulation calculation for all the changing range of the constant element or of the initial value of the state element that is used for sensitivity analysis and draw graph after each calculation. The effect of changing graphs shows us the sensitivity of the system elements to the changes of the element that has been chosen to make sensitivity analysis. While doing the sensitivity analysis we can also pause drawing graph or draw graph in a slower mode to have a closer look on the changes of the system elements.

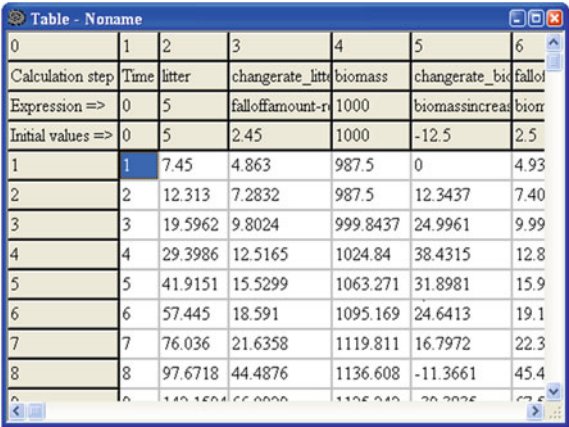
The graph can be saved in the picture format (*.bmp) or in the stream format (*.stm). When the graph has been saved in the stream format, we can open it in MM&S and draw in slower mode, pause drawing graph or change the size of the graph, without simulation calculation (standalone simulation graph).

1.3.6 Table Child Window

After conducting simulation calculations, MM&S automatically post all the results in a new table child window (Fig. 1.10). The table can be saved to a file with extension ‘*.tbl’.

In the first fixed row we can see the number of the column. The names of the variables are displayed in the second fixed row. The third fixed row is used for displaying the values of constant elements, or the initial values of state elements,

Fig. 1.10 The table child window



| 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|-------------------|------|----------|-------------------|----------|--------------------|--------|
| Calculation step | Time | litter | changerate_litter | biomass | changerate_biomass | fallor |
| Expression => | 0 | 5 | falloramount-r | 1000 | biomassincreas | bior |
| Initial values => | 0 | 5 | 2.45 | 1000 | -12.5 | 2.5 |
| 1 | 1 | 7.45 | 4.863 | 987.5 | 0 | 4.93 |
| 2 | 2 | 12.313 | 7.2832 | 987.5 | 12.3437 | 7.40 |
| 3 | 3 | 19.5962 | 9.8024 | 999.8437 | 24.9961 | 9.99 |
| 4 | 4 | 29.3986 | 12.5165 | 1024.84 | 38.4315 | 12.8 |
| 5 | 5 | 41.9151 | 15.5299 | 1063.271 | 31.8981 | 15.9 |
| 6 | 6 | 57.445 | 18.591 | 1095.169 | 24.6413 | 19.1 |
| 7 | 7 | 76.036 | 21.6358 | 1119.811 | 16.7972 | 22.3 |
| 8 | 8 | 97.6718 | 44.4876 | 1136.608 | -11.3661 | 45.4 |
| 9 | 9 | 119.1504 | 66.0000 | 1105.040 | 20.2026 | 67.5 |

or the expressions for calculating the values of the intermediate elements or of the change rates of state elements. The initial values of system elements are calculated if needed and displayed in the fourth fixed row.

The first column of the table contains the number of the calculation steps. The other columns contain the results of simulation calculation.

We can copy the table and past to other word processing software, or save the table in a text file by choosing menu ‘Table/Save in text file’, or save the table in a Excel file by choosing menu ‘Table/Save in excel file’.

1.4 Conclusions

By using four symbols to represent elements of the four element groups (a square for state elements, a rhombus for intermediate elements, a circle with three plus/minus signs inside for listed elements, and a circle for constant elements) MM&S enhances very much the visualization of the system structure. From the simulation diagram we can recognize the “mathematical nature” of each system element.

Change rate as an attribute of a state element in MM&S replaces the input and output of the softwares that have been designed based on the black box concept. In MM&S the links connect the state elements directly. In case if a state element has an incoming link, we understand that the change rate of this state element is being affected.

MM&S allows two formats of models: text model and model incorporated in the simulation diagram.

The graphic techniques (slower drawing of graph, pausing drawing of graph, pausing sensitivity analysis, or enlarging graph) allow user to better trace the change of system elements while drawing graphs or doing sensibility analysis.

Standalone simulation graph makes the results of the simulation more portable. The results can be saved in text or excel formats what makes it easy to report the modeling and simulation results.

Acknowledgments The Vietnam Academy of Science and Technology has supported this work.

References

1. Bossel H (1992) Modellbildung und simulation. Friedr. Vieweg & Sohn Verlagsgesellschaft mbH: Braunschweig/Wiesbaden, Deutschland
2. Bruenig EF, Bossel H, Elpel K-P, Grossmann W-D, Schneider TW, Wang Z-H, Yu Z-Y (1986) Ecologic-socioeconomic system analysis and simulation: a guide for application of system analysis to the conservation, utilization and development of tropical and subtropical land resources in china. Library of the Federal Research Centre for Forestry and Forest Products, Hamburg, Germany
3. Forrester JW (1989) The beginning of system dynamics. Banquet Talk at the international meeting of the system dynamics society, Stuttgart, 13 July 1989
4. Forrester JW (1994) System dynamics, systems thinking, and soft OR. Syst Dyn Rev 10(2):245–256
5. Sinh NV, Manh NH, Hung NM (2011) Modeling biomass-litter dynamics with the MM&S software. In: Proceedings of the 4th national conference ‘ecology and biological resources’, Hanoi, Vietnam, 21 Oct 2011, pp 1784–1791. (in Vietnamese, ISSN: 1859–4425)
6. Sinh NV (2006) An effort to enhance the computer simulation of dynamic systems: an example with mini-world model. In: Proceeding of the IUFRO international conference: ‘PATTERNS AND PROCESSES IN FOREST LANDSCAPES—consequences of human management’, Bari, 26–29 Sept 2006 (ISBN-10: 88-87553-11-4; ISBN-13: 978-88-87553-11-6)
7. Bossel H (2007) Systems and models: complexity, dynamics, evolution, sustainability. Books on Demand GmbH, Norderstedt
8. International Society for Ecological Economics (2004) Software reference guide: STELLA software technical documentation. ISEE Systems
9. <http://iebr.ac.vn/pages/1mms.asp>: Website for MM&S computer program. Access 16 Nov 2011
10. <http://www.embarcadero.com>: Website of the Embarcadero company. Access 12 Aug 2011

Chapter 2

A Software Architecture for Inventory Management System

Taner Arsan, Emrah Başkan, Emrah Ar and Zeki Bozkuş

Abstract Inventory Management is one of the basic problems in almost every company. Before computer age and integration, paper tables and paperwork solutions were being used as inventory management tools. These were very far from being a solution, took so much time, even needed employees just for this section of organization. There was no efficient solution available in the many companies during these days. Every process was based on paperwork, human fault rate was high, the process and the tracing the inventory losses were not possible, and there was no efficient logging systems. After the computer age, every process is started to be integrated into electronic environment. And now we have qualified technology to implement new solutions to these problems. Software based systems bring the advantages of having the most efficient control with less effort and employees. These developments provide new solutions for also inventory management systems in this context. In this paper, a new solution for Inventory Management System (IMS) is designed and implemented. Most importantly, this system is designed for Kadir Has University and used as Inventory Management System.

2.1 Introduction

Inventory Management is one of the basic problems for a company. It may cause a lot of paperwork, if there is no automated system available. Implementing such a system is possible but there are a lot of preliminary works such as determination of

T. Arsan (✉) · E. Başkan · E. Ar · Z. Bozkuş
Department of Computer Engineering, Kadir Has University, Cibali,
Istanbul, 34230 Turkey
e-mail: arsan@khas.edu.tr

the requirements, system structure decision—software requirements, barcode system selection and determination of the software tools.

2.1.1 Determination of the Requirements

Inventory Management System (IMS) is generally used by IT Office/Department or Accounting Office of a company or a university. Therefore, searching the basic needs for implementation is the first step of IMS design. Several meetings with IT Office and Accounting Office are arranged. Accounting Office needs detailed reporting tools, detailed categorization and declaration of specifications on each item, purchasing and billing info. The Information Technologies Office needs another module except the requirements of Accounting Office. The module is about the interior maintenance and exterior product service flow. For interior maintenance flow, there will be a section. This section will be available for all users. Basically, a maintenance request will be created by the users, and the IT Office will respond to these requests. Finally, it is necessary to consider the end user's needs that are also important part of the IMS software design. This led us to use barcode based system.

2.1.2 Software Requirements

After gathering all the requirements and information, we achieve next state which is quite important for the project. If the decision is not well organized, it may cause re-organization and programming the algorithms at the beginning.

At the beginning, first thing to decide is the general structure. We decide to build the system browser based. The system is going to be used by a large number of users, so it should have been designed to be reachable for every single user, instead of having a desktop application for every user. Also if we consider that Kadir Has University already has a similar system—Information Management System, it would be easy to integrate. By using the user database of the existing systems, the login parameters are ready to use for IMS. So, we would provide the portability. The next step is to decide the language to implement.

There are several parameters to decide in realization period. First is the compatibility, and then came the time which is related with ease of implementation. In the light of these redirections, we decided to use PHP as the language to code the system. It is compatible with the other software and hardware, and it is easy to integrate everything. PHP is an open source environment and there are a large range of resources to get help. Information Management System is already developed by using PHP. We also decide to use AJAX to obtain a faster system. In AJAX, only the needed data is being requested from the server, this method speeds up the loading time. While thinking how to code with PHP, we decide to use

Model View Controller based structure. This structure separates the visual design. We also use Code Igniter Framework for decreasing the amount of code lines. It gives ability to program faster with less effort. Code Igniter also support MVC structure. So we decide to PHP, MySQL and APACHE trio. We used MySQL as database. It is the best fit for PHP and it has the open source advantages. We can also show the comments of ORACLE as a reference to our right choice [1]. We used innodb engine, because it is transaction based, which is necessary for the project. To get the user data, we have to communicate with the database of University, which is programmed by ORACLE. We need to have the login information to have a fully integrated system, and also several details such as e-mail addresses and so on.

2.1.3 Barcode System

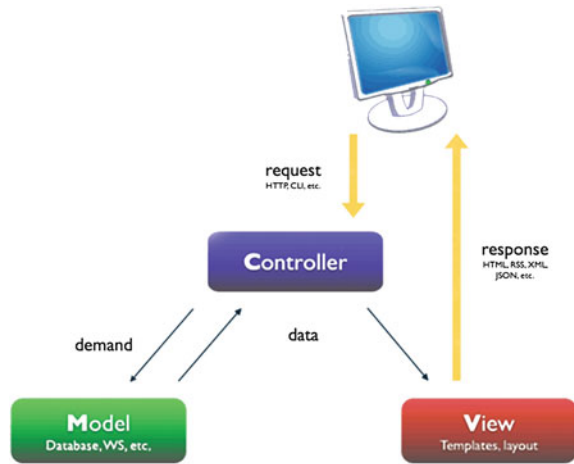
A barcode is an optical machine readable representation of data, which shows certain data on certain products. There are different types of barcodes, some of them have to be world wide unique and need license. We chose “code39” because first of all, code39 doesnot need license, have the possibility to print for unlimited products, supports full ASCII, number, letter and special character usage (capacity of 26 capital letters, 10 numbers, 7 special characters). And also it was the easiest to read and was the best fit for the barcode generators for PHP. As the barcode generator, we found one compatible with PHP. It is using PHP’s GD Graphics Library. We got the script and modified it to be able to print serial barcodes. It prints barcodes in increasing order (in multiple prints) on the screen, and sends to the printer. The script is distributed under GNU License. And the hardware part of the system is out of a label printer and a barcode scanner. Barcode scanner is a USB connected model. It scans the barcode, writes to the field where the cursor is, and presses enter. This function eases the usage. So the only extra hardware to be used by the staff will be a barcode scanner after the system is implemented to the enterprise.

2.1.4 Software Tools

2.1.4.1 PHP

As a derivation of Perl, PHP, is a server side, user interactive, programming language, works nearly in on all platforms. We can say that it is a general purpose scripting language. It can be embedded into html. It can use various databases such as MySQL, SQL, Oracle, MS SQL etc. Also contains many server interfaces. Open source is one of the best specifications of PHP. Among several frameworks, the most popular one is zen [2].

Fig. 2.1 Internal structure of MCV architecture



2.1.4.2 JavaScript

JavaScript is a scripting language used to enable programmatic access to computational objects within a host environment. Generally used to implement dynamic web sites. For more information <http://en.wikipedia.org/wiki/JavaScript>.

2.1.4.3 Asynchronous Javascript and XML

Asynchronous Javascript and XML (AJAX) is being used to build applications which are using JavaScript and XMLHttpRequest. It is mostly used in avoiding to load the whole page, just to load the needed part. By using XMLHttpRequest, it is possible to do more than one independent process.

2.1.4.4 Model–View–Controller

Model–View–Controller (MVC) is a software architecture which provides to implement the visual, data and processing code parts independent. For example the Model unit is a collection of classes in communication with the other parts. It is the process unit, processes the task ordered by the Control Unit. View unit is the place to deal with the presentation of the data to the end user. It can get the data from both Model and Controller unit. Also it can send interactive data to both units. Controller part is the main part of the structure. If we want to concretize the architecture, C is the brain, M is the nerves and V is the move. Internal structure of MCV Architecture is given in Fig. 2.1.

2.1.4.5 Code Igniter Framework

CodeIgniter is a framework with a small footprint, requires nearly zero configuration, does not require the use of command line and provides to use large scale of libraries. CodeIgniter also provided us object oriented programming [3].

2.1.4.6 Oracle

Oracle RDBMS is a relational database management system. It is one of the strongest software in this section and also one of the largest IT companies on the world.

2.1.4.7 MySQL

MySQL is the most popular open source database software. It is easy to use, fast and reliable. Also it is a good match with PHP [4].

2.1.4.8 phpMyAdmin

It is a software, coded with PHP. The main function of this software is to manage MySQL database through Internet. It can create databases, add/edit/delete tables, run SQL queries, manage user authorization and manage field keys are some of the features [5].

2.1.4.9 InnoDB

InnoDB is a standard database engine in all the packages distributed by MySQL. The main difference of this engine is it is compatible with ACID, and the important ones for our project, it is transaction based and supports foreign key [6].

2.1.4.10 Apache

Apache is a GNU licenced open source web server program. It can run on platforms such as Unix, Linux, Solaris, Mac OS X, Microsoft Windows. Even the usage of Apache decreases, still it is the most popular web server program on web [7].

2.2 Implementation

In this section, there is a closer look to implementation steps, work flow and software architecture. All the data and process flow will get clear. Even it does not seem so complex on the interface level, there are many process running behind to provide efficient function of the system [8].

2.2.1 Database Design

The most important component of the system is the database design. It can be easily realized, if we consider the amount and the scale of the inventories. To have a strong and flexible system, the design must be well studied.

After some studies including investigation, workflow scenarios, traces and design, we obtain the following tables.

1. Attributes
2. Categories
3. ci Sessions
4. Content
5. Debits
6. Inventory
7. Inventory Attribute Values
8. Jobs
9. Locations
10. Providers
11. Services
12. Service Records
13. Tech Users

We start with categories and locations tables. We needed the ability of listing unlimited category and location of items and setting unlimited hierarchy. So we started to search for sufficient algorithms. The research resulted with Nested Sets. Nested Sets is a model used to represent a set of data organized into a hierarchy is useful in a computer database management context. One way hierarchical data has been commonly represented within a database is with the Adjacency List Model, which takes a set of data nodes and recursively attributes parent nodes with their respective child nodes. The Nested Set Model takes the Adjacency List Model a step further by modeling subordination. This is done by assigning each node a beginning and ending node hierarchy number based on the total amount of data in the hierarchical tree.

Two important structures of the system is the database and the Graphic User Interface (GUI). In between, JavaScript and PHP provides the connection and interaction. GUI is located in the view section. With XHTML and CSS, the visual

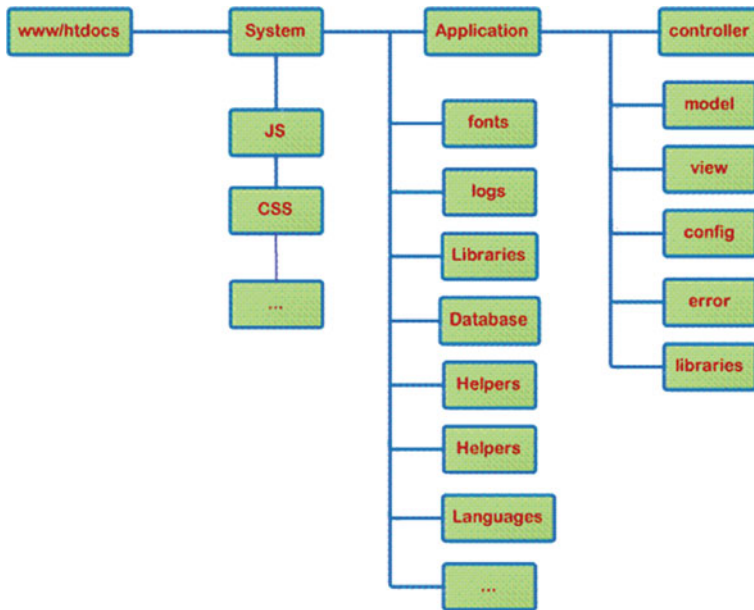


Fig. 2.2 Basic directory codeIgniter framework

part is generated. Javascript updates the HTML by the data sent by other layers and connects the GUI with working PHP structure. We used an Admin Panel called “CompleteLiquid admin control panel”. It was suitable for our MVC structure. With CSS and other related modifications, we could arrange it to our system. In the controller layer, dataflow decision codes take place which are Javascript and PHP. And in the Model section, PHP logic base code is placed.

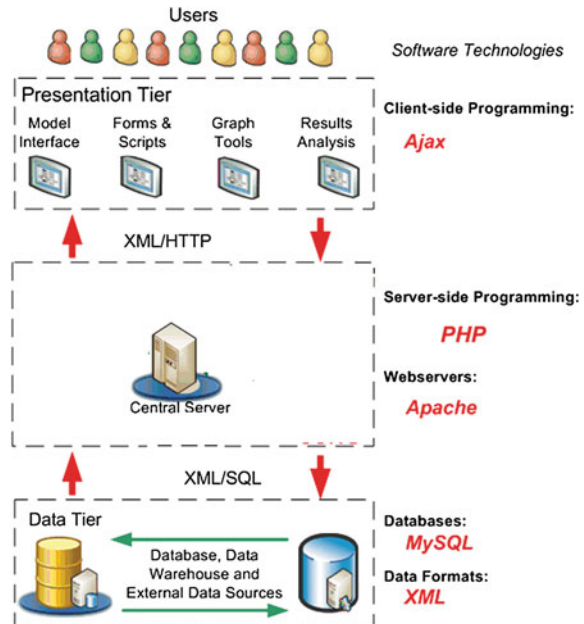
All codes are produced with CodeIgniter framework. The basic directory of the framework can be shown in Fig. 2.2, and a general look of IMS architecture is given in Fig. 2.3.

2.2.2 Modules

IMS is designed as several modules, separated by their specific roles and functions. In this part, the function of each module will be explained step by step. There are six modules in the system.

1. Definitions/Setup
2. Inventory Management
3. Service Management
4. Maintenance Management

Fig. 2.3 General IMS architecture



5. Debit Management
6. Deposit Management

2.2.2.1 Definitions/Setup

This module is the most important part of the system. Because if the definition of the enterprise is not implemented correctly to the system, it won't work efficiently. To have a successful, operational system, introducing the enterprise to the system is a must. And this is the first step of implementing the system to the enterprise.

In this module, there are four sections, which are mentioned below.

- User Setup
- Location Hierarchy
- Categorization
- Suppliers and Maintainers

(a) User Setup

In User Setup Menu, System Admin can create, edit or delete a user from the system. The table below is the user creation form. System admin can choose the role of the user from the picklist on the bottom.

(b) Location Hierarchy

The Location Hierarchy section is the place where you set the ownership hierarchy of the enterprise. System Admin can create one sub unit under another by choosing a unit.

(c) Categorization

In this section, a categorization of the inventories in the enterprise is being created. First step is the main categorization. The second step is related titles, and the last one is the specific keyword. User can add subcategories under each category after choosing it.

(d) Suppliers and Maintainers

In this section, System Admin can define inventory suppliers and the maintainers of each supplier. To define a maintainer, user can choose related supplier, and add maintainers under it.

2.2.2.2 Inventory Management

This Module is the center for all processes about incoming inventories. Module has two sections:

- New Registry
- Instant Registry

(a) New Registry

If the inventory is ready to record, this is the section to use. First step is to print a new barcode for the new inventory. The system pops out an alert. After this alert, there comes the next page, where the user has a button to print to confirm the barcode by scanning it into the field.

After confirming the barcode, the next page is an information form. There are Category, Supplier, barcode number (auto entered), serial number, price, receipt date and extra info fields to be filled. Then, the next step is the product specifications part. The first part has the title group of related inventory category. The user can switch between titles and choose related specifications from the second list. Then the chosen specifications are added to the last field as related keywords of the inventory. And after clicking the button, the entry is on the database, and in ready state for other actions on the system.

(b) Instant Registry

This section allows user to print several barcodes when the inventory is not ready to record. For example, it may not be possible to bring label printer everywhere. So the user can print as much as needed, the system creates empty fields in the database to be filled in inventory registration process. When these barcodes are scanned, register page comes to complete the register, and the normal registry flow continues.

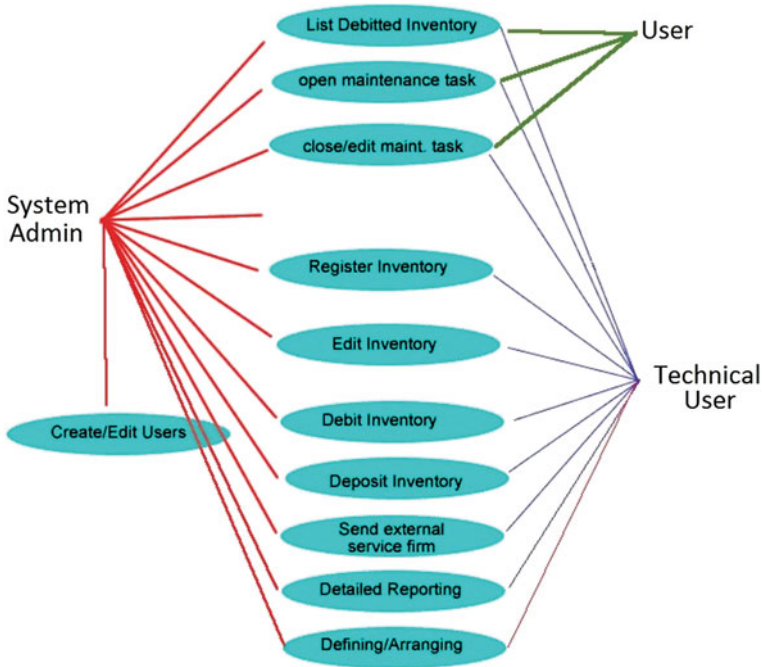


Fig. 2.4 Use case diagram of IMS

2.2.2.3 Service Management

Interior maintenance may not be sufficient, or may cause warranty problems, when inventories have technical problems. In this situation, they may need to be fixed by authorized maintainers. This module is for management of exterior maintenance.

There are three search fields in this section with Barcode No, Serial No parameters to reach the inventory or User Name parameter to reach the inventories of a specific user. After reaching the inventory, the window below comes to take action.

The user can view the service history of the inventory, and open a new service record to send the inventory to service. There is a pick list to select the service firm, and an info box to write a brief explanation. Then, another “close service record” link appears on the inventory information table.

When the inventory is returned from the service firm, user reaches the table below to close the record. User can write the information, process done to the text field and closes the service record, which will be seen in service history from that moment. Also user can use the “User Name” search box to reach a specific user, and list the inventories on users debit. Then, the same program flow can be used to take action.

2.2.2.4 Maintenance Management

This module is for interior inventory maintenance flow. There are two sections: first is to create new maintenance request for device users, the other is maintenance task listing for related worker.

The user will fill the required fields and then the request will be an active task for maintenance staff and will drop in their task table that they have on their interface if a staff member is assigned to the task. Otherwise, the task will drop into task pool. Then staff deals with the problem, and closes the task after writing info about actions taken. If the problem gets fixed before maintenance arrive, the user also can close the task.

2.2.2.5 Debit Management

This module sets the relationship between users and inventories. Basically, assign an inventory to a user, and performs other operations based on debiting.

2.2.2.6 Deposit Management

This module manages the barrowed items. Assign the inventories to desired person. There is also a chance to list deposited inventories in this module. It is a manner of tracking.

2.2.3 Users

Users have an authority hierarchy according to their section, and they are only allowed to use related modules except system admin. System admin has the authority to use every module in the system. In order to prevent improper use, the system logs every act, every edit. Although users can only use their modules, they can also use some other features of the system. There are three user types and use case diagram of the IMS is shown in Fig. 2.4:

1. System Admin
2. Technical Users, Accounting: Has authority on Inventory Management, Deposit Management, Debit Management, Service Management and also reporting features.
3. Basic Users: Basic users are only able to report their debit inventories and request service for them.

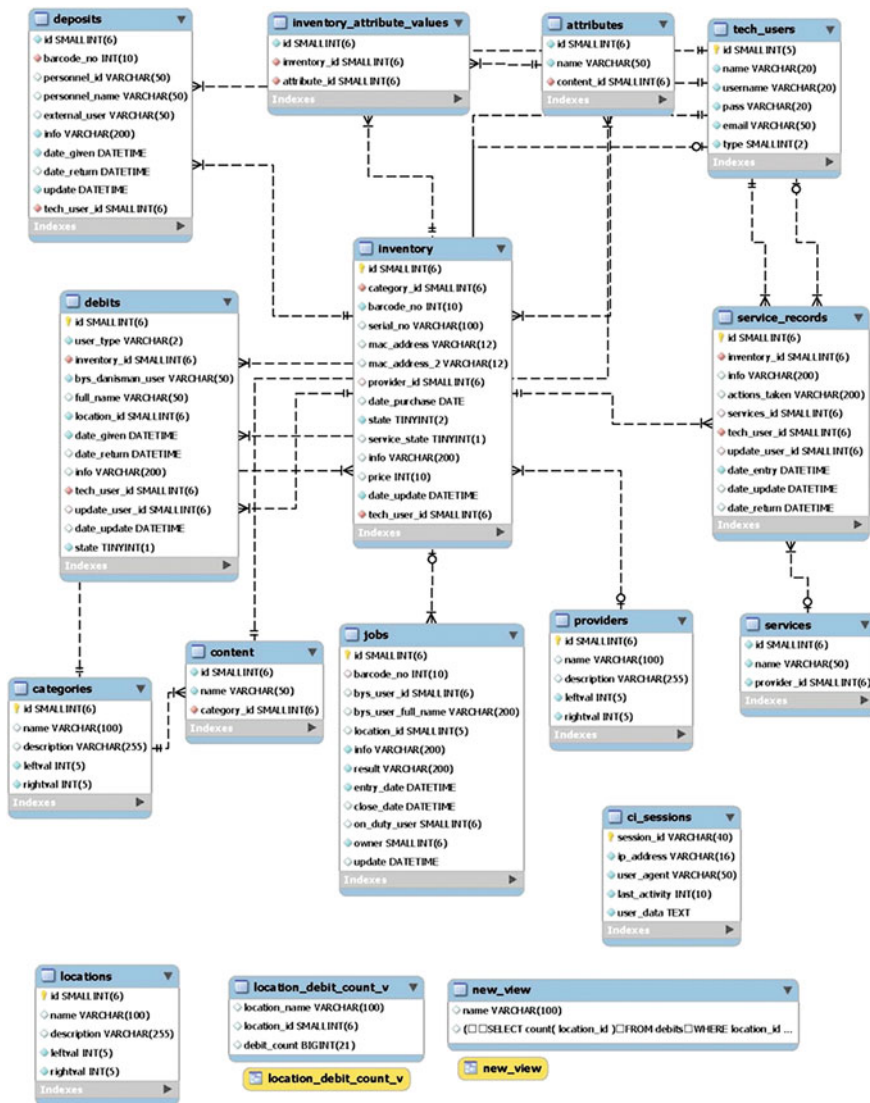


Fig. 2.5 Software architecture of—IMS

2.2.4 Login

The first step is “login” to the system. There are two login alternatives. First one is the system management, the other one is the end users. The login data is gathered from University database, so the users can login the system with their existing username and password.

Complete software architecture of IMS software is given in Fig. 2.5.

2.3 Conclusions

In the light of software engineering methods, we gather requirements, analyze and figure out the workflow, design methods and structures, construct scenarios, make tests, code the software, debug the faults and bugs, and finally we obtain a new IMS software. Our first aim is to develop the IMS software that is able to meet the requirements gathered. During the development process, there are many feedbacks, led us to re-design.

The software has the ability to track, to keep history, to give detailed reporting for each inventory. Also modules of the software manage the services and operations for inventories and users. Briefly, we could satisfy the user's requirements. It is also a scalable and a flexible solution. No matter how fast business's growing, proposed software can adapt to meet or exceed the requirements.

A new, functionally content rich software architecture model of an IMS is developed, presented, implemented and discussed. This level of implementation detail does not commonly appear in the literature. We believe this is a significant contribution.

References

1. Jason Gilmore W (2010) Beginning PHP and MySQL: from novice to professional, 4th edn. Apress, Berkely
2. Quigley E, Gargenta M (2006) PHP and MySQL by Example. Prentice Hall PTR, Upper Saddle River
3. Griffith A (2010) Codeigniter 1.7 Professional Development, Packt Publishing, Birmingham
4. DuBois P, Hinz S, Pedersen C (2005) MySQL 5.0 Certification Study Guide, MySQL Press
5. Zandstra M (2007) PHP Objects, Patterns, and Practice, 2nd edn. Apress, Berkely
6. Schwartz B, Zaitsev P, Tkachenko V, Zawodny J, Lentz A, Derek J. Balling (2008) High performance Mysql, 2nd Edn. O'Reilly, Sebastopol
7. Fultus Corporation (2010) Apache HTTP Server 2.2 Official Documentation—Volume II. Security and Server Programs. Fultus Corporation, Palo Alto
8. Chaffer J, Swedberg K (2007) JQuery reference guide: a comprehensive exploration of the popular javascript library. Packt Publishing, Birmingham

Chapter 3

Libraries Opt for More Online Sources

Zeenath Reza Khan and Sreejith Balasubramanian

Abstract A decade into the twenty-first century and the frenzy to stay in the forefront of discovering and adopting new technologies globally, in offices, malls and houses has no signs of declining. Education field is no different. With the rapid convergence of the information age and the boom in technology usage, information technology has taken a firm place in classrooms across borders. With this demand, the demand for academic text, publications, and other sources at students' finger tips is at a record high (American Library Association, 2011). As the demand for online sources increase, so do the sources themselves, or do they? This study looks closely at nearly 20 different tertiary education institutions and determines if, at all, the online sources have increased as perceived.

3.1 Introduction

Technology is a part and parcel of education in this century. E-learning, or part-there-of, is the new trend in and out of classrooms. Educators and academic institutions are trying more and more to introduce various technology-based components in their pedagogy in order to successfully reach students and help them learn. From using blended-learning tools such as Black Board, to using

Z. R. Khan (✉)

Faculty of Computer Science and Engineering, University of Wollongong,
Dubai, United Arab Emirates
e-mail: zeenathkhan@uowdubai.ac.ae

S. Balasubramanian

Faculty of Business and Management, University of Wollongong,
Dubai, United Arab Emirates
e-mail: sreejithsubramanian@uowdubai.ac.ae

Podcasts and so on, tertiary educators have tried to move away from traditional teaching techniques to automate their teaching tools, teaching environments and resources. So why not libraries?

Libraries have always played a central role in facilitating teaching and learning for both students and teachers. Books, periodicals, research papers, journal articles and catalogues have been the basis of research and education over centuries.

This study looks closely at nearly 20 academic institutions, their traditional versus online sources, and tries to establish a rate of increase. As this is part of a research grant study titled *Implications of increased online-sources and readily-available e-technology on students' attitudes towards e-cheating in the UAE* from 2008 to 2010. The result of this research is expected to lead authors to a subsequent study into the impact, if any, of 'increased online resources' to students' attitude towards e-cheating. However, that is beyond the parameters of this paper, and so this study will aim to establish whether there is a marked increase in online sources over the years among various academic institutions.

3.2 Libraries and Literacy

It is believed that libraries, in their most basic form of simple collections of knowledge that has been written down, has existed as far back as civilization itself [1]. Historical findings of written knowledge date back to over 5000 BC in the form of parchments, papyrus and such [1].

Because learning and literacy go hand-in-hand, it is believed that literacy necessitates the establishment of a library [2–5]. Some contenders initially believed libraries to play only a supporting role to education; however, a stronger position eventually surfaced that positioned libraries as learning centers, with education an essential part of their mission [6]. “The American Library Association’s official position on the role of libraries in the area of literacy encourages library involvement and places no limitation on how libraries should be involved in literacy education” [6]. Thereby, during the 1970s and 1980s, “many academic libraries in the United Kingdom, Canada, the United States, Germany, Scandinavia and Australia started ...programs of user education, bibliographic instruction, or reader education” [7] that have continued well into the twenty-first century.

3.2.1 Why Libraries?

In 2002, the Education Commissioner of Massachusetts, United States of America, David P. Driscoll noted that ‘a Simmons College study released in October 2000 found a strong link between school libraries and student achievement’ [8–16].

A recent study published by United Kingdom’s National Literacy trust suggests that students ‘who use their local public library are twice as likely to be above

average readers' compared to their counterparts [17]. Another study conducted by Haruki Nagata, Akira Toda and Paivi Kytomaki states that "using the library for its materials or research purposes has a direct connection to students' achievement of educational outcomes" [18]. Evidently, studies show statistically that "in some subjects, students who read more, measured in terms of borrowing books and accessing electronic resources, achieve better grades" [19]. Needless to say, libraries have been proven to play a central role in the academic lives of students.

3.2.2 Going Online

To fulfill reader demands and aid in literacy, libraries are adopting newer technologies, better modes of reaching their readers. Over the past decade that has meant getting their readings to become electronic. There is a strong trend of libraries 'embracing digital collections, though most libraries will continue to offer both print and digital collections for many years to come' [20]. This is further highlighted by Jagdish Arora in his study that "the libraries will not become digital libraries [completely], but will rather acquire access to ever growing digital collections on behalf of their users" [21]. Although for most libraries, electronic sources are a small percentage of the items made available to readers, they do represent the fastest-growing media today [22]. This is primarily because more and more people are becoming Internet users everyday across the globe. With powerful search engines such as Google, GoogleBooks and so on, readers want everything at their fingertips. Study by Carol Tanopir has found that about 75 % of most Internet users 'say they are library users and 60 % of library users are Internet users' [20]. "Coincided with availability of software, hardware and networking technology, the advent of world wide web, its ever increasing usage and highly evolved browsers have paved the way for creation of digital libraries" [21].

According to Tanopir,

"Libraries prefer digital collections for many reasons, including, but not limited to, the following: digital journals can be linked from and to indexing and abstracting databases; access can be from the user's home, office, or dormitory whether or not the physical library is open; the library can get usage statistics that are not available for print collections; and digital collections save space and are relatively easy to maintain." [20].

As mentioned by references [20] and [23], going online, and using digital sources rather than print also help reduce costs for libraries in terms of processing and space costs which is a great incentive for libraries.

3.3 Gap in the Study

During the course of this study, authors have found ample research that mention how important libraries are to learning and literacy; the kinds of policies that are in place to help libraries cater to enhancing the literacy levels of their communities;

of how and why libraries are going from traditional print sources to electronic media.

But very few, if any, studies have been found that statistically prove that there indeed is an increase in online sources for students in terms of library databases, library membership to other online sources and so on.

Particularly in the United Arab Emirates, a considerably new nation founded in late 1971, its boom in the technology and education sector has been taking place at a tremendous rate in the last decade or so. 'The UAE ranked the 1st of all the Arab states in the 2009–2010 Networked Readiness Index (NRI) study issued by the World Economic Forum, and 23rd among all 133 countries assessed' [25]. The World Summit on Information Society that was held in Geneva in [24] also reported the UAE as first among the Arab states in terms of the individual indicators analyzed such as 'Internet user rates' and 'International internet bandwidth capacity' [25]. In the last decade, over twenty new academic institutions opened their doors to students from across the globe. Yet, authors have found no research that has been reported that highlights the growing trends of libraries increasing their online sources in the country.

3.4 Looking Closely at Some Academic Libraries in the UAE

To gather substantial information that may help answer the research question, '*have online sources increased in libraries and at what rate over the last five years*', a primary research was conducted to identify the increased online resources in libraries of universities in UAE.

Over 25 academic institutions were approached, of whom, 3 rejected participation on the grounds that they were not allowed to take part in any outside research studies, and 2 rejected participation because they were fairly new and had not yet established a proper physical library, much less an online presence. The remaining 20 institutions—that represented both accredited and non accredited colleges, institutions, universities, and affiliated, non-affiliated and direct campuses—participated under the condition of strict confidentiality due to severe competition in the market.

A survey tool was developed that was completely anonymous in nature, and focused on each institution's collection of both traditional library resources such as text books, journals and catalogues, versus online resources such as online databases, e-readings material online resource materials for subjects (all either own collection or attained through membership), irrespective of syllabi taught or courses offered. The survey was designed to collect data for a period of five years. However, as the results were drawn in, authors had to minimize the time period to three years, as a large portion of the institutions had either not been around that long, or not had a substantial presence in the market long enough to show a significant difference in library collections.

Table 3.1 Summary of 20 academic institutions in the UAE recording their average collection of resources in the libraries over a period of three years from 2008 to 2010

| Traditional Library resources | Average collection (2008) | Δ | Average collection (2009) | Δ | Average collection (2010) | Total Percentage increase from 2008 to 2010 |
|-------------------------------|---------------------------|-------|---------------------------|------|---------------------------|---|
| Text books | 10009 | Δ | 10689 | Δ | 12008 | Δ |
| Δ | Δ | 680 | Δ | 1319 | Δ | 16.65 |
| Journals | 1412 | Δ | 1678 | Δ | 1999 | Δ |
| Δ | Δ | 266 | Δ | 321 | Δ | 29.36 |
| Catalogues | 250 | Δ | 892 | Δ | 1001 | Δ |
| Δ | Δ | 642 | Δ | 109 | Δ | 75.02 |
| Online Library resources | Δ | Δ | Δ | Δ | Δ | Δ |
| Online Databases | 4000 | Δ | 15000 | Δ | 24897 | Δ |
| Δ | Δ | 11000 | Δ | 9897 | Δ | 83.93 |
| E-reading materials | 900 | Δ | 6700 | Δ | 10000 | Δ |
| Δ | Δ | 5800 | Δ | 3300 | Δ | 91.00 |
| Online resources for subjects | 334 | Δ | 2700 | Δ | 6384 | Δ |
| Δ | Δ | 2366 | Δ | 3684 | Δ | 94.77 |

Authors approached libraries to fill in the online survey that was made available on an online-survey tool and a link presented along with a participation information sheet.

Before analyzing the results, the authors made assumptions based on the hypothesis that:

- *the increase in online databases will outweigh all other resources both traditional and digital*
- *rate of increase for traditional resources will be slower than digital*
- *there is a significant increase in the online resources in universities across UAE.*

3.5 Libraries and Online Resources in the UAE

The survey that was completed in five months produced the results illustrated in Table 3.1.

When comparing the results of the traditional versus the online sources, the graphs below illustrate that both types of resources have increased over the past three years. As Fig. 3.1 shows, most academic institutions upped their collections

Fig. 3.1 Graph showing the increase in collections of traditional resources

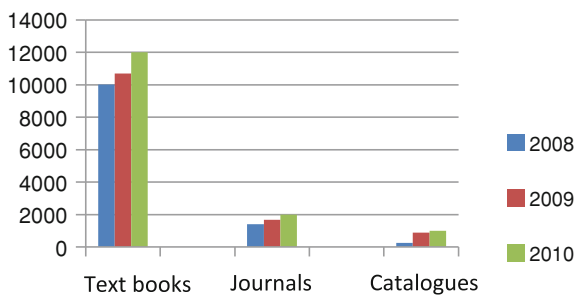
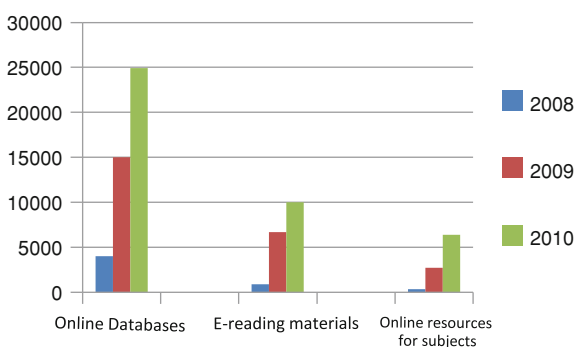


Fig. 3.2 Graph showing the increase in collections of digital resources



of text books by approximately 16 % in three years, whereas, journal collections in print were increased by 29 % and catalogues by 75 %. As Fig. 3.2 demonstrates, online databases that include direct collection of sources and access to databases through membership to journals, conference papers, catalogues and such increased by 83 % in three years, while e-reading materials' collection increased by 91 % and online resources for subjects increased by about 95 %. While the actual average amounts remain maximum for textbooks and online databases at about 12,000 books and 25,000 online databases respectively in 2010, the study shows a smaller percentage of increase in these two categories, although it was expected that these would be the two categories that would have the highest percentage of increase, particularly online databases.

The rate of increase in collection of digital resources when compared to traditional resources is high. Where the rate of increase for traditional resources range from 16 %—a maximum of 75 % increase, the online source collection increase ranges from 83 % to about 95 % increase over the three years studied.

This finding supports the assumption made that the rate of increase for traditional resources is slower than the rate of increase for digital resources. This ultimately proves the hypothesis that overall online sources have increased in majority of the academic institutions.

3.6 Conclusion

Libraries have always been pushed to the forefront of knowledge gathering, acquirement and to lend themselves to social literacy since the beginning of civilization. It is no wonder then that as the information age reaches a feverish high a decade into the twenty-first century, libraries are frantically trying to keep up with the various media preferred and used by readers.

The authors had set out to find some statistical evidence to answer the research question: *have online sources increased in libraries and at what rate?* It is believed that given the limited scope of this paper, the results do shed evidence to answer the question in the positive. Despite the limitation of permission, nature of confidentiality and sensitivity of the study due to a perceived existence of fierce competitive market, the data collected clearly shows an overwhelming increase in academic libraries' collection of digital or online collections when compared to the traditional print resources and at a rate much higher than the rate of increase of traditional resources.

3.7 Continual Study

As the original purpose of the study was to add to the research grant topic, the findings will now be used as a factor, 'increased online sources', to see if the increase in online sources in academic institutions have any impact on students' attitude towards e-cheating. "Both the increase in the amount of information available and the improvement in its accessibility have had a huge impact on academics' information behavior" [25]. This will further be proved or disproved by the authors through a follow up study.

References

1. Krasner-Khait B (2001) Survivor—The history of the library, History Magazine (Online) <http://www.history-magazine.com/libraries.html>
2. Davidson J (1988) Adolescent illiteracy: what libraries can do to solve the problem—a report on the research of the project on adolescent literacy. J Youth Serv Libr:1(2):215–218 (EJ 373 763)
3. Humes B, Cameron C (1990) Library programs. Library literacy programs: analysis of funded projects 1989. Washington, DC: office of educational research and improvement. ED number pending; also available from superintendent of documents, U.S. Government printing office, Washington, DC 20402
4. Mathews AJ, Chute A, Cameron CA (1986) Meeting the literacy challenge: a federal perspective. Libr Trends 35(2):219–241 (EJ 347 674)
5. Scamber L (2003) The role of libraries in literacy education. US federal government. ERIC Digest (Online) <http://www.libraryinstruction.com/literacy-education.html>
6. Quezada (1990) Shaping national library literacy policy: a report from the Alexandria forum. Wilson Libr Bull 65(3):22–24, 158 (EJ number pending)

7. Virkus S (2003) Information literacy in Europe: a literature review. *Inf Res* 8(4):2–3 (Online) <http://informationr.net/ir/8-4/paper159.html>
8. Mass.gov (2002) Education commissioner highlights importance of libraries. Massachusetts department of elementary and secondary education (Online) <http://www.doe.mass.edu/news/news.aspx?id=761>
9. Abell J (1999) The impact of the use of school libraries on student achievement. *027.8 School Library Bulletin* 5(1)
10. Bingham JE (1994) A comparative study of curriculum integrated school library media programs. Achievement outcomes of sixth-grade student research papers. Doctoral dissertation, Kansas State University
11. Brien DP (1995) The teaching and learning processes involved in primary school children's research projects. Doctoral dissertation, University of New South Wales
12. Callison HL (1979) The impact of the school library media specialist on curriculum design and implementation. Doctoral Dissertation; University of Southern Carolina
13. Lance KC (1994) 'The impact of school library media centers on academic achievement', *School Library Media Quarterly*, 22(3):167–172
14. Oberg D (1995) Principal support: what does it mean to teacher-librarians? In: *Proceedings in sustaining the vision, a selection of conference papers, 24th international association of school librarianship conference*, Worcester College of Higher Education, Worcester, pp 17–25, 17–21 July 1995
15. Sivanesarajah Y, McNicholas C, Todd R (1993) Making sense of Science: An information skills approach. *Science Education News* 42:25–27
16. Todd R (1995) Integrated information skills instruction: does it make a difference? *Sch Libr Media Q* 23(2):133–138
17. Scholastic (2011) Libraries play crucial role in supporting literacy. Scholastic Ltd. (Online) <http://education.scholastic.co.uk/content/15441>
18. Nagata H, Toda A, Kytomaki P (2011) Students' patterns of library use and their outcomes. Research center for knowledge communities. Japan. (Online) <http://www.kc.tsukuba.ac.jp/div-comm/pdf/report0704.pdf>
19. Goodall D, Pattern D (2011) Academic library non/low use and undergraduate student achievement: a preliminary report of research in progress. *Libr Manage* 32(3):159–170
20. Tanopir C (2003) Use and users of electronic library resources: an overview and analysis of recent research studies. Council on library and information resources. (Online) <http://www.clir.org/pubs/reports/pub120/pub120.pdf>
21. Arora J (2001) Building digital libraries—an overview. *DESIDOC Bull Inf Tech* 21(6):3–24
22. Glanton D (2011) HarperCollins puts new limits on library e-books. *Los Angeles Times*. (Online) <http://articles.latimes.com/2011/mar/07/business/la-fi-ebooks-20110307>
23. Montgomery CH, King DW (2002) Comparing library and user related costs of print and electronic journal collections: a first step towards a comprehensive analysis. *D-Lib Mag* 8(10) (Online) doi:10.1045/october2002-montgomery
24. WSIS (2010) About UAE ICT. World summit on the information society. (Online) <http://www.wsis.ae/About-UAE-ICT.php>
25. Olle C, Borrego A, Librarians' perceptions on the use of electronic resources at Catalan academic libraries: results of a focus group. University of Barcelona. (Online) http://diposit.ub.edu/dspace/bitstream/2445/11162/1/Article_FG_Final.pdf

Chapter 4

Emerging Threats, Risk and Attacks in Distributed Systems: Cloud Computing

Isabel Del C. Leguías Ayala, Manuel Vega
and Miguel Vargas-Lombardo

Abstract Nowadays Cloud Computing provides anew paradigm to organizations, offering advantages, not only for its speed but also for the opportunity of save costs when implementing new applications, by just paying for the resources you use. This article identifies the threats, risks and attacks, also identifies their causes, in addition, proposed solutions from the National Institute of Standards Organization and Technology (NIST) and Cloud Security Alliance (CSA) are also mentioned here.

4.1 Introduction

There has been a boom in organizations adopting Cloud Computing as a way to expand and replace their ICT infrastructure in recent years. Cloud Computing [1] consists of three main service models, such as, software, platform and infrastructure that provides services through the Internet, where the user only pays for the used resources, providing multiple advantages, such as: reliability, availability, cost savings, flexibility and portability. As any technology, the infrastructure, software and services offered by a cloud are exposed to risks, vulnerabilities and threats; its security is responsibility of both, the service provider and the customer.

I. D. C. Leguías Ayala (✉) · M. Vega · M. Vargas-Lombardo
Technological University of Panama, calle domingo diaz, penonomé, Panama, Panama
e-mail: isabel.leguias@utp.ac.pa

M. Vega
e-mail: manuel.vega1@utp.ac.pa

M. Vargas-Lombardo
e-mail: miguel.vargas@utp.ac.pa

4.2 Cloud Computing

For [2], Cloud Computing has been defined as a model that allows users to quickly access a set of shared computing resources (e.g., networks, servers, storage and services) as on-demand service through the Internet or other computer network, it requires a minimum administrative effort or services provider interaction. According to [2], it is a paradigm of distributed computing on a large scale, driven by economies of scale, to a set of virtualized and dynamically-scalable, computing resources, managed computing power, platforms and services that are offered on demand to customers through the Internet. Cloud Computing consists of the following characteristics [3–6]:

- On-demand self-service: Services can be requested directly by the customer or user by means of Internet, for example, the user pays only for the time of using the service.
- Broad network access: Services are located in the cloud and are accessible from any desktop or mobile equipment with Internet access. For example, Smartphone, laptops, personal computers or PDAs.
- Resources pooling: Services in the cloud can be used by multiples users under a model of multi-tenancy at different places.
- Rapid elasticity: The quantity and quality of the services offered by the cloud can increase or decrease rapidly depending on the user needs.
- Measured service: It checks and optimizes automatically the use of resources by using a measuring capability suitable for the type of service (for example, storage, processing, bandwidth and active user accounts). Thus allow to control, monitor and report on resources and help to maintain transparency of the used resources for both, the service provider and the user.

4.2.1 Cloud Service Models

The service providers generally focus on a cloud type according to the functionality of service that they offer: infrastructure, platform or software [7]. However, there are no restrictions to offer several types of services at the same time. The following are the service models of the Cloud Computing [4, 8, 9]:

- Software As Service (SaaS): Applications are hosted and offered online via Internet by means of web browsers to provide a traditional desktop. Customers have control only over the applications supplied the service provider, the later is responsible of manage and maintain the applications, data and the underlying infrastructure. For example, Google Docs, Sales force CRM, SAP Business by Design, Zero Nines y Lotus Live.
- Platform as a Service (PaaS): The platform is provided as a service, which generally offers a virtualization environment, allowing developers to write their applications according to the specifications of a particular platform without

having to worry about the hardware. Developers deploy and run applications in the platform without controlling the underlying infrastructure. Examples of this type of service are Forge.com, Google App Engine, and Windows Azure.

- **Infrastructure As Service (IaaS):** This service model provides customers with computing resources as a service. Clients can obtain resources such as, servers, applications and network computers that are administered on-demand by the service provider, over this resource, clients can deploy applications including the operating system. In this service model, security is managed by the client. For example, Amazon Web Services, VMware, OpenCloud, EMC2 and Cleversafe.org.

4.2.2 Classification of the Cloud

The NIST defines four deployment models for Cloud Computing that are the following [9]:

- **Public Cloud:** In this type deployment model, multiple clients can quickly access shared computing resources offered by single service provider; and only pay for the operating resources. Although it presents indisputable advantages there are latent security threats, such as regulatory compliance and quality of service (QoS). Some public cloud examples are, Amazon Cloud, Google Apps, and Windows Azure.
- **Private Cloud:** Computing resources are managed and controlled by a private company. Often the data center is deployed and managed by internal staff or the service provider. Its main advantage is that the security, regulatory compliance and quality of service is carried out by the cloud owner. An example of a private cloud is the one owned by eBay.
- **Community Cloud:** The infrastructure used in this deployment model is shared by several organizations and is supported by a specific community that share interests (for example mission, security requirements, policy or compliance considerations). Example Alexandros Marinos [10] Digital Ecosystem.
- **Hybrid Cloud:** It is a combination of clouds (public, private or community) which are separated as entities but at the same time are united by standardized or proprietary technologies that allow the portability of applications and data. Example Bursting Cloud.

4.2.3 Virtualization

Virtualization is a software based technology widely used in Cloud Computing, which employs hardware and/or software simulation in order to run multiple operating systems and applications on top of a shared hardware architecture [11, 12].

The environment produced by this simulation is known as virtual machine (VM). There are different forms of virtualization which are classified according to the computing architecture layer. As an example there is the application virtualization where the simulated environment allows an application to use a virtual implementation of the application programming interface (API) letting the application to be executed on different platforms without changes in the application itself.

Full virtualization is another form of virtualization, where operating systems run on top of virtual hardware without being aware of the virtualization environment, in separate virtual machines called guest operating systems. Guest operating systems located in a single host are administered by the hypervisor, also known as virtual machine monitor (VMM), used to control the flow of instructions between the guest operating system and the physical hardware. For example.

Paravirtualization [13] is a virtualization technique where guest operating systems are aware of being executed on a virtual environment. This leads to the modification of the operating system kernel which now requires to replace non-virtualizables instructions with “hypercalls” which communicate with the hypervisor in a direct way.

4.3 Cloud Computing Threats

According to the study “Top Threats to Cloud Computing v1.0” made in 2010 by CSA, an international organization that drives the use of best practices for Cloud Computing security, there are seven major security threats that affect Cloud Computing infrastructure. Below are the seven threats contained in this report [14, 15]:

- Abuse and Nefarious Use of Cloud Computing: This type of threats concern mostly IaaS and PaaS Cloud Computing service models when they have a registration process that allow to anonymously activate this service, for example a malicious user that has a valid credit card can access the service, which may result in propagation of spammers, malware and other illegal activity by cyber criminals who could use the cloud as a means of operations. Example, IaaS services used by Zeus Botnet, and other malicious programs for command and control functions.

To mitigate this threats CSA recommends the following: implementing a stricter registration process, close credit card fraud monitoring, same as traffic from possible illegal activities by users, check and confirm public blacklists to determine traffic from/to such IP ranges.

- Insecure Interfaces and APIs: Service providers usually offer a set of interfaces and APIs to their users that allows them to control and interact with their resources. This allows customers to handle and monitor Cloud services executed through APIs. Thus security will strongly depend on how it has been implemented on the API itself. Examples of this type of threat are: uncontrolled or

anonymous access, reuse of tokens, and authentication without encryption.

It is recommended to ask and analyze the security model of the cloud provider interface, also ensure that authentication and access controls are implemented using safe data encryption methods.

- **Malicious Insiders:** In every organization, malicious insiders should be considered one of the most important threats, because they can access data and applications in your organization. In cloud environments this is no different to organizations, because security incidents may also occur carried out by dissatisfied employees or accidents by mistake or ignorance too. In most cases, the service provider manages active accounts and the customer is in charge of requesting activation and deletion of user accounts, this may cause a security hole when the client reports late or does not report at all about changes in organization's staff to the service provider for the necessary updating in active accounts. Incidents produced by this type of threat impact negatively on the organization's image and the assets of the company. Service providers must instruct their customers on methods to control these internal threats.

As a mitigation measure, according to the CSA, legal terms of reliability and non-disclosure agreements should be specified in employment contracts. Problems that may exist in the notification processes should also be checked and identified.

- **Shared Technology Issues:** These threats affect IaaS models, due to the design of the physical components used to provide this model (CPU, GPU, storage, networks, etc.) that were not designed for shared application architecture. In some cases a virtualization hypervisor can access the physical resources of its host, causing security incidents.

To prevent such incidents, it is recommended to establish a defense in depth strategy concerning mainly, on computing resources, storage and network. Also good security practices should be used to properly manage resources, so no activity from a IaaS customer interfere with some other customer activity. For example, an exploit or a malware in a virtual machine, capable of access resources of the host platform, may access more than one customer infrastructure.

CSA recommends the use of best practices for installation and configuration of hardware and its virtualization. Close supervision of environments including frequent vulnerability scans and configuration audits in order to detect any unwanted changes or malicious activity. Adequate service level agreements with service providers to include patching of applications that require it and vulnerabilities fixes.

- **Data Loss:** There are many ways in which information can be compromised. For example, an unauthorized data deletion or change without a proper backup policy in place, causing a loss of data. In the cloud, the risk of compromise data increases, thanks to an increased number of interactions between them caused by the architecture itself. This can cause losses to the organization image, economic mishaps and legal issues due to information leakage, or violation of security and privacy regulations, etc. For example, misuse of encryption keys,

inadequate authentication and/or authorization and weak audit.

The suggested recommendations to mitigate these threats are: APIs to implement a robust access control, using encryption to protect data traffic. Analyze that data is protected during design time, as during runtime. Provide effective mechanisms for key generation, storage, and destruction of information. Define policies to establish procedures for the destruction of persistent media before throwing it out, as well as make the respective security data backups.

- **Account or Service Hijacking:** In the cloud, if attackers accomplish to get credentials of an user, can access activities and transactions, as well as manipulate data, send and return forged information or redirects customers to malicious sites.

It is recommended to enforce policies to prohibit credential sharing between users and services. Implement authentication techniques whenever possible. Analyze sessions looking for illegal activity.

- **Unknown Risk Profile:** One of the main advantages of cloud infrastructure is to minimize the amount of software and hardware that organizations have to buy and maintain, letting them to focus on the business. While this results in cost savings, this should not be a reason for implementing a weak security due to a poor knowledge of the infrastructure in use.

The technical information of the platform should be studied and taken into account to define the security strategies to implement in a cloud. For example, define with whom share the infrastructure, as well as information from unauthorized access attempts is very important when deciding on security strategies. Against this threat, CSA recommends to demand service providers to disclose infrastructure details and to disclose applicable logs and data when a security incident occurs. Also, demand alerts on security incidents including relevant details for customers.

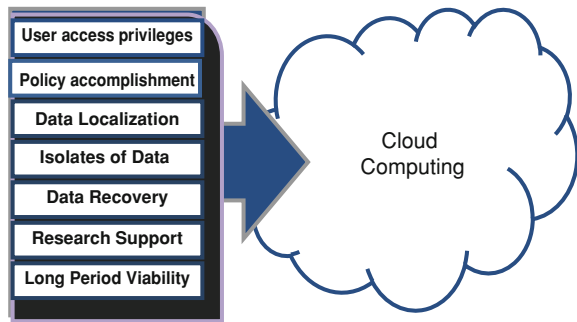
4.4 Cloud Risks

The large amount of data and the critical information stored in clouds are attractive targets for hackers. As a platform that offers different types of services, cloud characteristics are unique and require a risk assessment in some areas regarded as critical, which are: data integrity, recovery and privacy, evaluation of legal issues in regard to e-discovery, policies and standards enforcement, as well as audits [14, 16, 17]:

Garther depicts seven specific security concerns that customers should address with service providers before selecting any cloud provider (Fig. 1):

- **Privileged user access:** Management or processing of confidential information outside the organization brings a level of inherent risks, because these external services do not frequently carry out physical, logical, and personal controls, which require knowing who manages or has privileges on the data.

Fig. 4.1 Risk of cloud computing



- Policy compliance: Definitely, customers are responsible for the data security and integrity, even if it is stored outside the organization and managed by a service provider.
- As well as traditional organizations perform external audits to guarantee security, it is also necessary for cloud service providers to adopt these types of best practices.
- Data localization: Since Cloud Computing uses virtualization technologies on resources located locally and overseas. Clients should make sure with service providers who has jurisdiction and what regulations authority has over the data storage and processing, by making agreements for the data treatment, including contractual obligations to comply with privacy regulations of jurisdictions required by clients.
- Data isolation: In a Cloud data is generally placed in a shared environment alongside to data from other clients. Cloud providers should use effective encryption methods to guarantee data isolation between clients; they also should be liable for ensuring this isolation.
- Data recovery: Service providers should have security policies with data recovery methods in case of disasters. Cloud providers must have the ability to restore data completely in a maximum pre-established amount of time. Garther recommends replicating data across multiple infrastructures to avoid vulnerabilities in the event of a major failure.
- Investigative support: Malicious or illegal activities can be almost impossible to detect due to record activity from multiple clients in a single log or if it is deployed through a series of hosts and data centers. It is recommended that the service provider manage incident logs and data centrally, but appropriately segregated.
- Long-term viability: Ideally, adoption of cloud computing will gradually grow in the coming years, providing better availability and quality of service. However, due to unpredictable market changes a service provider is likely to be acquired by a bigger service provider. Customers need to make sure they will have the ability of.

4.4.1 Other Cloud Risks

- Failures in provider's security: In the field of cloud computing, service provider's security is very important, since they control hardware and hypervisors used to store data and run applications [14, 18].
- Attacks by other customers: In a Cloud Computing context, offered services are shared among customers. If client segregation breaks down, a client can access data from another or interfere in their applications.
- Availability and reliability issues: Data centers used in cloud computing are usually more reliable than enterprise data centers, but sometimes interruptions occur. Additionally, Internet is critical to cloud's reliability and availability.
- Perimeter security model broken: Many organizations use a perimeter security model with a robust security in the organization's network. This model has been weakening through years, especially by outsourcing and a highly mobile workforce. Security models for situations where critical applications and data are stored outside the security perimeter of the organization should be used.
- Integrating provider and customer security systems: Over decades, organizations have tried to develop a unified guide as well as other components for their security architecture, for example, automated provisioning, incident detection and response, etc. Service providers must integrate with these systems to avoid manual uncoordinated responses as in bad old days.

4.5 Attacks in Cloud Computing

Nowadays, many companies are migrating to Cloud Computing, this fact is perfectly known by cybercriminals, who have found a new target for their attacks in this increasing used technology.

4.5.1 Attack Vectors

Cloud infrastructure, based on virtual machines that share resources between multiple clients and guest virtual machines, gives rise to new threats. An important threat to cloud infrastructure is the possibility of a malicious code that can be executed from a virtual machine and affect other hypervisor or guest virtual machine [15].

Advanced administration features included in hypervisors, like the ability of live virtual machine migration, increase hypervisor's code length and complexity, thus their space of attack is also increased. Some of the attack vectors that these cybercriminals use are the following [19]:

- Denial of service attack (DoS): Several security professionals point out that cloud services are vulnerable to these attacks because these services are shared by many users letting this type of attack to impact on a large number of users and cause more damage. The multi-tenant infrastructure used in cloud computing has more specific threats associated to DoS attacks, like [20]: Shared resources consumption, where attacks prevent other users to make use of system resources as execution time, memory, storage and network interfaces, can result in a denial of service. Regarding virtual machine and hypervisor exploitation, where vulnerabilities in the virtualization environment or in the host OS can be exploited by an attacker to caused is ruption or instability to the system. For example, the devastating DoS attack suffered by Twitter in 2009.
- Side Channel Attacks: With this type of attack a malicious virtual machine is placed near a cloud server and then used to launch a side channel attack.
- Authentication attacks: Authentication is considered a sensitive point in both hosted and virtual services, because of its inherent weakness. There are different forms to authenticate users known as authentication factors. For example, users can be authenticated based on what the user knows, has or is. The main targets of cybercriminals are the mechanisms and methods used to ensure authentication processes.
- Man in the middle attacks: It is an attack type where the attacker is placed between two users (the sender and receiver) and intercept and modify messages without users to realize.

The same way in [20]:

- 1 Social networking attacks: The social networking sites have brought a higher risk of social engineering attacks. Cloud Computing is being targeted due to the high amount of information handled by its customers (organization and users). Attackers setup their identity to gain the trust and use online information to verify staff relationships and roles and prepare their attacks. An example of these attacks is the social engineering attack, which is performed against a user through the people he or she know and the social networks subscribed by the user.
- 2 Attack on mobile devices: The use of smart phone has been increasing in the latest years, same as the use of Cloud Computing, which is not only limited to laptops or desktop PCs, it also allows connectivity with mobile devices. Attacks, on mobile devices are also raising since they present similar vulnerabilities as traditional devices such as desktop PCs and laptops.
- 3 Service provider data control [21]: Treatment and handling of applications and data must consider their legal implications, which are complex and not well understood in most cases. Service level agreements must include measures to avoid possible lack of control and transparency when data is stored by third parties. Other concerns that also need to be considered are: due diligence, audit, economic data analysis and the availability effective cost.

- There are also other types of attacks which are [22]:
- 1 Wrapping Attack: This is a technique where a cyber criminal deceives parties in a communication during the translation of the SOAP message in the Transport Layer Service (TLS). The message body is duplicated and sent to the server as a legitimate user. The authentication of the message is then verified BY its signature value (which is also duplicated) and the message integrity is checked done. As a result, the cyber criminal is able to penetrate the cloud and may execute malicious code to break the normal operation of servers in the cloud.
 - 2 Malware-injection attack: In a malware injection attack, the attacker intends to inject malicious services or code so later these forged services appear listed as valid to users to run from the cloud. If the attacker succeeds, then the service will be subject to ev as dropping. This attack can be achieved through data modifications in an ingenious way to change functionality or causing dead locks, which make a legitimate user to wait for the end of a job that was not generated by the user. Here the attacker makes the first step by implementing a malicious service, so it runs in IaaS or SaaS cloud of servers. This type of attack is also known as a meta-data spoofing attack.
 - 3 Flooding attack: When an attacker has been authorized to make an application in the cloud, then he or she can easily create false data and setup these requests to the server. When processing such requests, a chain reaction is started compromising each cloud server one by one. The first compromised server checks the authenticity of the requested jobs, which consume CPU cycles and memory, resulting in services to stall and the off loading of services to another server, then the next server is compromised in the same way. The attacker is successful in flooding the cloud system when these requests propagate uncontrolled through the IaaS.
 - 4 Data Stealing: This attack is more traditional and common in respect to the violation of a user account. The user account and password have been stolen by any means. As a result, the sub sequent theft of confidential data or destruction of data can obstruct the storage integrity and security the cloud.
 - 5 Accountability Check: The payment method in cloud systems is based on resource consumption. When a client use cloud services, the usage duration, the amount of data transfer over the network and the CPU cycles per user are recorded and taken into account for billing purposes. Thus when an attacker has compromised the cloud with a forged service or malicious code, then the legitimate owner of the account is billed for resource consumption not originally intended by him or her. In many cases clients are not aware of the attack until they detect the real cause of the additional CPU usage, and after being billed for such usage.

4.6 Trust in Cloud Computing

When an organization decides to use Cloud Computing services it must trust control of security issues to a provider. The following four aspects of security must be shared with and in some cases controlled by service providers [3, 23]:

- **Internal access:** The data processed or stored outside the organization's security control boundary, raises organization's risk levels. Threats to information security are known by most organizations and the same also apply to services Cloud, meaning that internal threats go beyond those that can carry current employees, and should include external personnel such as contractors, organization affiliates or other parties who also have permission to access networks and operating systems within the company. Incidents may include different types of fraud or deception, as well as cause resource damage and theft of confidential information. The moving of information and applications done in the cloud by the service provider brings security risks to sensitive information managed by the provider and to its customers.
- **Compound Service:** Services in the Cloud can be made through various layers and other cloud services. For example, a SaaS provider can build its services using the structure of another IaaS or PaaS service provider. Service providers that use third parties or outsource some service should raise concerns, including control measures over third parties, defined responsibilities between parties, as well as a defined way to escalate problems that may occur. A customer to fully trust a service provider needs to know the agreements with third parties before reaching an agreement with a service provider; it is also required to maintain without the terms of these arrangements changes or to notify in advance of any anticipated changes. The responsibilities and performance guarantees can be a serious problem for composite services.
- **Visibility:** The transfer of services to the cloud passes control to the service provider to ensure the data and applications running on the systems of the organization. Administration, procedures as well as technical controls must be adjusted according to ones used in the organization's internal systems to avoid creating security holes. Service providers are free to give details of the security and privacy because that information can be used as a means of attack. There must also be agreements between the parties to ensure that policies and procedures are applied throughout the system lifecycle. Agreement should have a procedure for customers to gain visibility on provider's security controls and processes and its performance over time.
- **Risk management:** As mentioned earlier, cloud services are not in the direct control of the customer who really owns the information stored and/or processed by the cloud. This scenario brings complexity to proper risk management and assessment, which cannot be based on the total knowledge of the security controls employed by service providers and the measures of its effectiveness, thus level of trust should be based on other factors.

4.7 Case Study

The National Authority for Governmental Innovation (AIG) [24] is the entity responsible for the modernization of the Panamanian State. Among its objectives, there is the improvement of the entire system of Information and Communication

technology (ICT), in order to solve problems such as licensing, acquisition of new systems and applications, quality of service, as well as the reduction of operational costs. Among its projects of expansion, it has created a platform of computer services that will support the Panamanian governmental entities through the project called “NubeComputacional”, which infrastructure is already operational.

Since 2010 AIG initiated the migration of services and systems of Panamanian entities towards the cloud service. Some of these services are informatics portals, emails, public accountancy and other applications considered of critical importance. It also has contingency services to ensure the provision of services without interruption.

Currently there are State entities that use the governmental cloud. At the same time, these entities can acquire storage services and cloud data backup, allowing the reduction of operational costs and hardware maintenance.

One of the issues of special interest for the governmental entities using the cloud service, is the security of the sensitive information they handle, as well as the legal frame that regulates them, taking into account the threats from criminal attacks or political motivations [25]: there are articles that punish informatics crimes in the Panamanian Penal Code. Other inconveniences are the threats that are faced when using the cloud: The high consumption of resources by the users, and the great amount of requests to the service, make the service expand to the point of needing to outsource to external service providers that are outside the defined security policies and lastly the inappropriate definition of the contracts with the service providers, as well as the lack of follow up of these contracts.

Recently the creation of “CSIRT Panamá” [26] was approved, as entity in charge of providing answers to security incidents in such a way that the necessary actions are taken for the prevention, treatment, identification and resolution against security attacks. In this way, it will offer information to improve the security of the information and telecommunication systems of the governmental sector.

Taking into consideration the characteristics of this project, besides the laws mentioned previously, it's necessary to elaborate an strategic plan that considers the security and privacy of the information that it administers, before migrating from their traditional systems to the cloud. It must begin by migrating to the cloud the information that is not considered of sensible or critical importance, considering procedures that allow the incorporation of their operations in a gradual manner to the cloud, because of the complexity of the information that it handles problems could arise and the creation of new methods of evaluating and risks is imperative. This means that risks are to be determined and if it's necessary to redefine the requirements in case they don't adapt to their needs. The human factor and the regulatory frame should also be taken into consideration.

In the case of the computational cloud of the Panamanian State, we consider necessary that the following considerations are met [27]:

- The governmental cloud service should be structured in such a way as to prevent the shutdown of the entire service from an attack on a single part of the system, as well as various internet providers that guarantee the quality of communication.

- Define the levels of service necessary as a part of the requirements, whether it be as a means of evaluating parameters as availability, answer time, etc. These evaluations will allow measuring the performance of the services the cloud provides. In addition to identifying the controls and adequate them in such a way that the minimum performance level is met.
- Check the existent policies and procedure in terms of security and evaluate how these can help improve the different cloud service models.
- Have a Security Operations Center (SOC) that monitors the cloud where the alerts and other indicators are checked in case they are activated by incoming threats.
- Have a security staff that is prepared and capable of looking after the proper functionality and safety of the system.
- Make backup and data recovery plans in the event of a security incident.
- Perform backup and recovery tests periodically to ensure that segregation and controls are effective.
- In case that the Panamanian State decides to allow private companies to use to cloud, these companies have to be regulated by the law pertaining State contracting.
- Ensure that the main requirements of security, strength and legal framework are specified into the services requirement levels and written down at the moment the contract, with the provider of services of the governmental cloud, is defined.
- Guarantees that apply access controls of the personnel/provider of the cloud that allows a logical segregation of responsibilities.
- Perform training and education of security awareness for all the public employees at least once a year. A new personnel must be instructed in security awareness during their integration process and before they access the operating environment of the cloud.
- Identify and clearly define roles and user privileges.
- To have identity federations and Web Single Sign On.
- Implement strong access controls to the users in the cloud. Likewise to guarantee the authorized access to information management in the cloud.

4.8 Conclusion

Due to the rapid growth that Cloud Computing is taking is the important for organizations and customers to know its infrastructure and capability as well as its main inconveniences. One of the main disadvantages, is presented by its complex way to implement security procedures and methods. Data localization raises important legal considerations due to the involvement of different jurisdictions and regulations. Another aspect is enforcing compliance with security policies and rules, which require service providers and clients to establish clear policies and procedures of how to handle services and the protection provided to sensitive datastorage, as well as what the actions to take in the event of incident or attack that may affect cloud services will be.

The service provider must make use of strong access control mechanisms granting users with the minimum required privileges needed by them.

The Panamanian government has just started its migration to cloud services. To migrate to a governmental cloud is necessary to create a juridical or legal framework that should regulate thus the use and contracting of the cloud with the providers of services, as the security that both the infrastructure and the protection of the critical information that handles the government must have.

References

1. Dawoud W, Takouna I, Meinel C (2010) Infrastructure as a service security: challenges and solutions. Paper presented at the 7th international conference on informatics and systems (INFOS), pp 1–8
2. Zhao Y, Foster I, Raicu I, Lu S (2008) Cloud computing and grid computing 360-degree compared. In: Grid computing environments workshop (GCE'08), pp 1–10
3. W. J. (2011) Guidelines on security and privacy in public cloud computing. NIST J 1–60
4. Grance T, Mell P (2009) Definition of cloud computing. NIST J 1–7
5. Cloud Security Alliance (2009) Security guidance for critical areas of focus in cloud computing V2.1. J Ala Acad Sci 76
6. Khajeh-Hosseini A, Sriram I (2010) Research agenda in cloud technologies. Technical Report.
7. Schubert L, Jeffery K, Neidecker-Lutz B (2010) The future of cloud computing opportunities for European cloud computing beyond 2010. ACC 2011, Part IV, pp 1–71
8. Creese S, Goldsmith M, Auty M, Hopkin P (2010) Inadequacies of current risk controls for the cloud. In: 2nd IEEE international conference on cloud computing technology and science, pp 659–666
9. Chen Z, Yang J (2010) Cloud computing research and security issues. In: International Conference on Computational Intelligence and Software Engineering (CiSE), pp 1–3
10. Marinos A, Briscoe G (2009) Community cloud computing. First international conference on cloud computing. pp 1–12
11. Souppaya M, Scarfone K, Hoffman P (2010) Guide to security for full virtualization. Recommendation of the national institute of standards and technology (NIST), pp 1–35
12. Kong J (2010) A practical approach to improve the data privacy of virtual machines. 10th IEEE international conference on computer and information technology (CIT 2010). pp 936–941
13. Payal S, Amarnath J, Rajeev N, Ravi P (2010) Security in multi-tenancy cloud. In: International Conference on Computational Intelligence and Software Engineering (CiSE), pp 4–7
14. Liu DQ, Srinivasamurthy S (2010) Survey on cloud computing security. In: 2nd IEEE international conference on cloud computing technology and science. pp 8–12
15. Cloud Security Alliance (2010) Top threats to cloud computing V1.0. J Ala Acad Sci 14–16
16. INTECO (2011) Riesgos y amenazas en cloud computing. pp 1–32
17. Gartner (2008) Assessing the security risk of cloud computing.
18. Hanna S (2009) A security analysis of cloud computing. Cloud Comput J <http://cloudcomputing.sys-con.com/node/1203943>
19. Gregg M (2009) 10 security concerns for cloud computing. Global Knowledge. pp 1–7
20. Centre for the protection of national infrastructure (CPNI) (2010) Information security briefing cloud computing. <http://www.203.128.31.71/articles/9New7IBHp4z.pdf>
21. Amardeep S, Vem E (2011) Attacks and security in cloud computing. Int J Adv Eng Appl pp 300–302

22. Vrbsky SV, Zunnurhain K (2010) Security attacks and solutions in clouds. In: 2nd IEEE international conference on cloud computing technology and science. pp 1–4
23. Jansen WA (2011) Cloud hook: security and privacy issues in cloud computing. 44th Hawaii international conference on system sciences. pp 1–10
24. Amardeep Singh Er. VEM (2011) Attacks and security in cloud computing. *Int J Adv Eng Appl* pp 300–302
25. Nube computacional. Autoridad de Innovación Gubernamental de Panamá
26. Gaceta Oficial, República de Panamá, Ministerio de la Presidencia, Decreto Ejecutivo No. 709
27. ENISA (2011) Seguridad y resistencia en las nubes de la Administración Pública. pp 1–135

Chapter 5

Cognitive Antenna System for Sustainable Adaptive Radio Interfaces

Ligia Cremene and Nicolae Crişan

Abstract Communication systems are usually implemented on a heterogeneous infrastructure and must operate in environments with accelerated dynamics. Adaptation is thus a key feature of such a system. Long-term, sustainable, adaptive solutions did not receive much attention in the design phase of wireless communication systems. With the advent of LTE, which was designed as a highly flexible radio interface—created to evolve—there is room for disruptive solutions to be put in place. A new approach for the receiver is proposed, where the antenna takes an active role in characterising and eventually learning the operation environment. The proposed solution—a *Cognitive Antenna System (CAS)*, is based on two main mechanisms that we called *antenna vision (AV)* and *signal fishing (SF)*. In the core cognitive cycle ‘observe-decide-act’ we aim to improve the ‘observe’ part, which critically influences the whole decision process. The SF and AV mechanisms bring a set of advantages: higher received SNR, no additional noise, higher AoA estimation accuracy.

L. Cremene · N. Crişan (✉)

Department of Communications, Adaptive Systems Laboratory, Technical University of Cluj-Napoca, Cluj-Napoca, Romania
e-mail: Nicolae.Crisan@com.utcluj.ro

L. Cremene
Romanian Institute of Science and Technology,
Cluj-Napoca, Romania
e-mail: Ligia.Cremene@com.utcluj.ro

5.1 Introduction

One of the main challenges for wireless communication systems is to implement spectrum efficient radio interfaces, which can compensate for impairments of the radio channel and smartly surf a dynamic environment, while maintaining a low degree of complexity in signal processing.

The proposed new approaches and techniques should not try to improve the existing physical layer through incremental changes but should cover the design of comprehensive models and architectures that can cope with the challenges and opportunities of radio communication systems of future generation.

Our approach is that of a long-term, sustainable, dynamic adaptation for wireless receiver chains. The antenna is seen as a pivotal element in determining and assessing quality of a wireless communication link. Actually, the user-perceived network or equipment quality relies, ultimately, on antenna performance. Moreover, until recently, the antenna influence on channel measurements was considered a bias, whereas now it can be used to a benefit by integrating it into the channel analysis [1, 2]. The antenna should take an active role in characterizing and eventually learning the operation environment by first ensuring high accuracy and reliability in observing the environment.

The antenna, the actual RF delivery and reception enabler, has evolved at a much lower pace than baseband processing techniques. Smart antennas have been proved to significantly improve system performance in terms of capacity and reliability, for various communication systems.

Multiple independently tilting beams may now be supported for different information/signals on the same antenna array. This enables different operators or different access technologies to be combined onto a single antenna array, while maintaining some network design/planning independence.

On the other hand one can observe a delay in broad adoption of truly innovative techniques such as the adaptive antennas [3]. Even if effective solutions for adaptive antennas exist for more than a decade, they are not yet implemented on a large scale. Limited reconfigurable solutions were introduced: Cross-polarized antennas, multi-band antennas, and the adjustable tilt antennas (multi-tilt, VET (Variable Electric Tilt) and RET (Remote Electrical Tilt)).

Therefore a rethink in terms of adaptation and sustainable evolution of telecommunications systems and equipments seems necessary. This may involve a constant correlation between the mathematical idealization, the physical phenomenon, and the latest technology solutions.

In the mean time, the first release of LTE (Long-Term Evolution) as defined by the 3GPP, was issued [4]. LTE was designed as a highly flexible radio interface [5, 6] created to evolve. The first release of LTE provides a new flat radio-network architecture designed to simplify operation and to reduce cost. It brings together the most robust, reliable and capacity enhancing technologies like FDD, TDD, OFDM, MIMO, etc. The adaptation is to be achieved by numerous parameters that would ensure high flexibility.

Support for multi-antenna transmission was an integral part of LTE from the first release, and the channel quality measurements for link adaptation and scheduling are designed to cater for this [7]. The presence of at least downlink-receive diversity is assumed. More advanced multi-antenna schemes also are supported by LTE, including transmit diversity, spatial multiplexing (including single-user and multi-user multiple-input multiple-output (MIMO), with up to four antennas, and beamforming. In the uplink, both open-and closed-loop transmit-antenna selection are supported as optional features [7].

In this paper we present the conceptual model and preliminary simulations of a CAS. The CAS relies on the mechanisms of signal-fishing (SF)—proposed by the authors in [2], and antenna vision (AV)—recently developed. A CAS improves the reliability of the wireless link by performing radio scene analysis and responding to changes in the RF signal environment.

Cognitive systems may exhibit a reactive or proactive behavior based on environment features, external stimuli, user requirements, operational constraints and capabilities. The CAS exhibits a proactive behavior as the antenna array takes an active role in characterizing the radio scene.

Antenna vision is used as a tool for radio scene analysis and provides a signal-space representation. Such representations of the signal space at the receiver enable observation of significant behaviour features of the incoming signals. Antenna-based spatial characterization of the channel fading enables the detection of signal maxima in the antenna environment—SF.

The paper is structured as follows: Section 5.2 presents the CAS conceptual model. The two core mechanisms—SF and antenna vision—are described in Sect. 5.3. Section 5.4 presents and discusses the simulation results. The conclusions are presented in Sect. 5.5.

5.2 Cognitive Antenna System Model

The core functionality of a Cognitive Radio is based on the ‘observe-decide-act’ cycle [8]. A CAS also implements this cycle and is not completely autonomous, nor completely controlled by a cognitive Tx-Rx equipment. We see cognition implemented in a distributed manner across the receiver processing chain (cross-layered).

The model in Fig. 5.1 captures the main functionalities of a CAS and highlights the proposed mechanisms of SF and antenna vision. Cognition mechanisms like perception (sensing), learning and reasoning, knowledge and representation models, all feed the decision making process and finally the action.

A primary processing part, meant for short-term adaptation, is located at the antenna array and controller level. The antenna array performs both the ‘observe’ and ‘act’ parts based on its sensing and actuating capabilities. The antenna is actually a self-structuring array with sensorial and actuatorial memories. The antenna vision mechanism is based at this level.

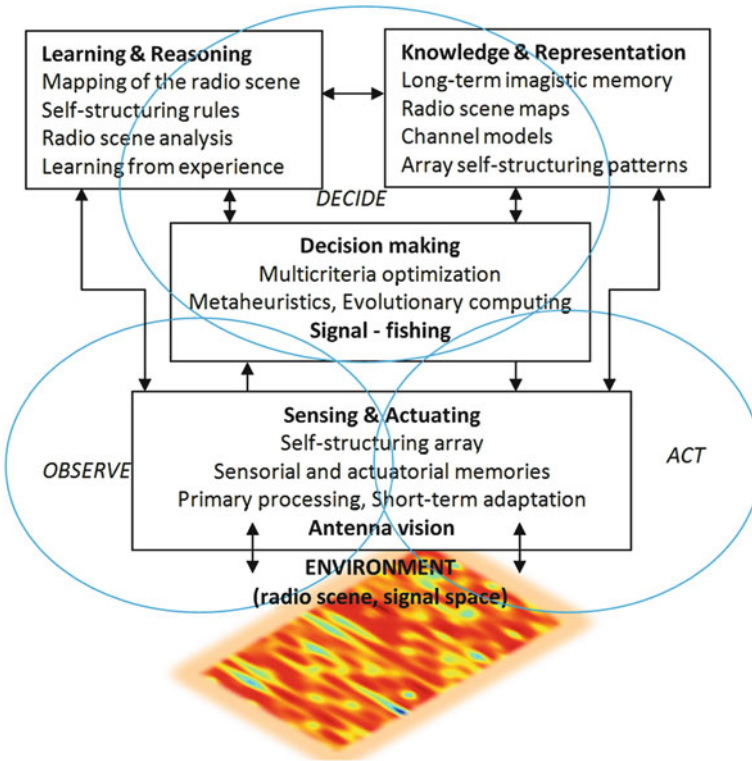


Fig. 5.1 Cognitive antenna conceptual model

Antenna vision is used as a tool for radio scene analysis and provides a representation of the signal space. Such representations of the signal space at the receiver enable observation of significant behaviour features of the incoming signals.

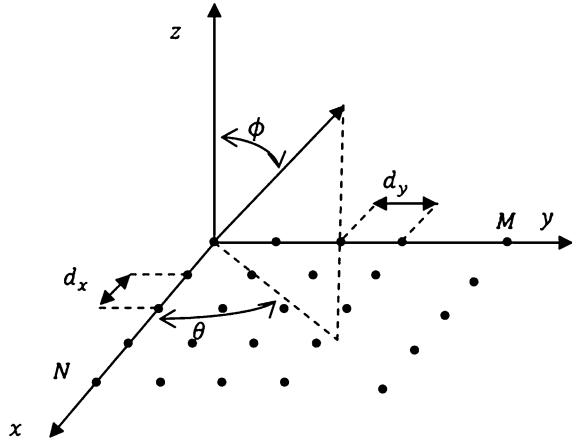
The learning and reasoning part is intimately linked to the knowledge and representation part. Mapping of the radio scene and learning self-structuring routines need to access and update knowledge and representation models such as long-term imagistic memory, radio scene maps, channel models, array self-structuring patterns.

The design of the decision making part needs special attention in the respect of maintaining a low complexity, especially that there are lots of computational intelligence tools available: metaheuristics, evolutionary computing multicriteria optimization. The signal-fishing mechanism is mainly based at this level.

SF addresses fading mitigation in adaptive multiple antenna receivers. It exploits the spatial component information of the channel in order to detect and exploit the channel signal maxima. This spatial method is meant to maximize the SNR by spatial decorrelation of the incoming fronts.

The next section describes the SF and AV mechanisms.

Fig. 5.2 The antenna is described as an $M \times N$ rectangular array with variable element spacing



5.3 The CAS Signal-fishing and Antenna Vision Mechanisms

SF is a new method of detecting and exploiting the signal maxima of the channel thus improving the received SNR.

The concept of antenna-based SF based on the associated concept of short-range, short-term fading characterization was proposed by the authors in [2]. The latter turns the selective channel problem into a flat fading problem, thus enabling the receiver signal processing part to perform better. By estimating or measuring the angle of arrival, phase shift, and amplitude of the incoming signal a short-term, short-range fading characterization is obtained. It is the antenna array that has the main role here. SF is an antenna-based spatial fading characterization method enabling the detection of signal maxima in the antenna environment.

An implementation of the proposed signal-fishing concept [9] using a genetic algorithm optimizer for antenna element positioning shows that the maximum signal levels of the channel can be detected and used to increase the mean received signal power. The channel fading effects are mitigated starting at the antenna level, an approach that does not introduce additional noise like current baseband processing techniques.

The antenna is described as an $M \times N$ rectangular array with element spacing $d_x = d_y = \lambda/2$ or $\lambda/4$) (Fig. 5.2).

Only two elements in the array are active at a given time. In order to search for the best positions of the active elements an evolutionary optimizer is used. This spatial method is meant to maximize the SNR by spatial decorrelation of the incoming fronts.

Activating the appropriate two array-elements results in the maximization of received signals R_{x1} and R_{x2} :

$$R_{x1}(a, b) = \sum_{i=1}^{n_p} \rho(\theta_i, 90^\circ) f_i \exp[j((a-1)(\beta d_x \cos \theta_i)) + j(b-1)(\beta d_y \sin \theta_i)]$$

$$R_{x2}(c, d) = \sum_{i=1}^{n_p} \rho(\theta_i, 90^\circ) f_i \exp[j((c-1)(\beta d_x \cos \theta_i)) + j(d-1)(\beta d_y \sin \theta_i)]$$

where

ρ —the antenna element factor

$f_i = |f_i| \exp(j\varphi_i)$ is the complex value associated to wavefront i .

a, b, c , and d are the coordinates of the activated receive elements R_{x1} and R_{x2}

β is the phase constant

θ_i is the angle of arrival (AoA) of the i th front.

The SF mechanism needs to find the maximum of the mean received signal power

$$P = |h_{11}|^2 + |h_{22}|^2$$

where h_{11} and h_{22} are the diagonal channel matrix coefficients.

An array controller translates the coordinates a, b, c, d into a binary signal suited for on/off element switching. The Packet Error Rate (PER) value is used as a trigger for the evolutionary search of the array element coordinates a, b, c and d . The angles of arrival are estimated based on the available R_{xx} matrix, using state-of-the-art AoA estimation methods (e.g. ESPRIT, MUSIC) [10, 11, 12].

Maximum signal levels of the channel can be detected and used during operation, to increase the mean received signal power. No additional noise is introduced.

Conducting radio scene analysis includes analyzing the signal space in terms of space, time, frequency, code and location. Antenna vision is used as a tool for radio scene analysis and provides a signal space representation. Such representations of the signal space around the antenna enable observation of significant behaviour features of the incoming signals. The proposal of a visual method for radio scene characterization (antenna vision) opens the way to using new computational intelligence tools in array signal processing.

5.4 Simulation Results

The results represent a sub-set of more extensive simulations and illustrate the improvement in (i) AoA estimation accuracy, (ii) generation/representation of the signal space, and (iii) signal maxima detection and exploitation.

Figure 5.3 captures the estimated pseudospectrum using the standard MUSIC algorithm on a uniform linear array (ULA). The angles of arrival are: AoA = [−40 −36 −24 24 36 40].

Fig. 5.3 Estimated pseudospectrum using the standard MUSIC algorithm on a uniform linear array (ULA)

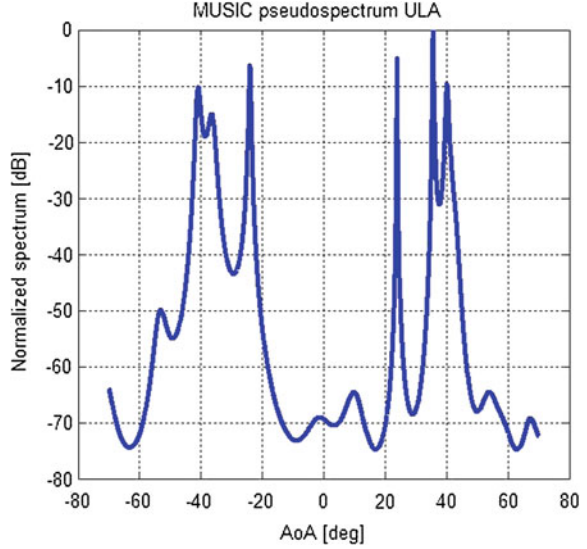


Fig. 5.4 Estimated pseudospectrum using the MUSIC algorithm together with the signal-fishing mechanism

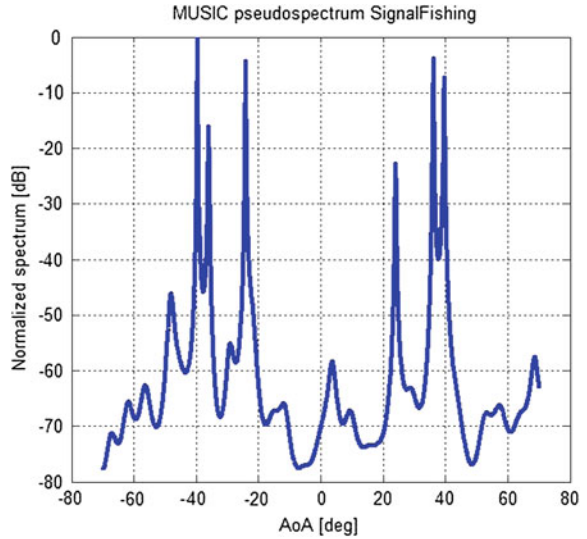


Figure 5.4 capture the estimated pseudospectrum using the MUSIC algorithm together with the signal-fishing mechanism. A significant improvement in accuracy can be noticed by comparing the two pseudospectra (Figs. 5.3, 5.4).

Figure 5.5 illustrates the array signal-space generated by the antenna vision mechanism. It can be noticed that certain array elements are better positioned in order to detect and use the signal maxima.

Figure 5.6 illustrates the signal maxima detection based on an evolutionary search of the radio scene. The solution pair is marked by the red dots (global

Fig. 5.5 The signal space view provided by the antenna-vision mechanism

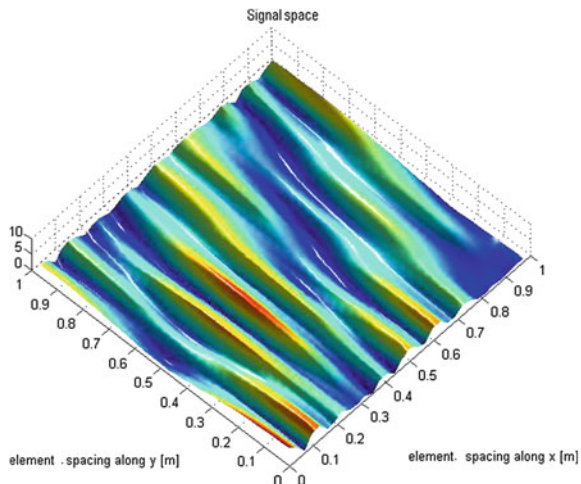
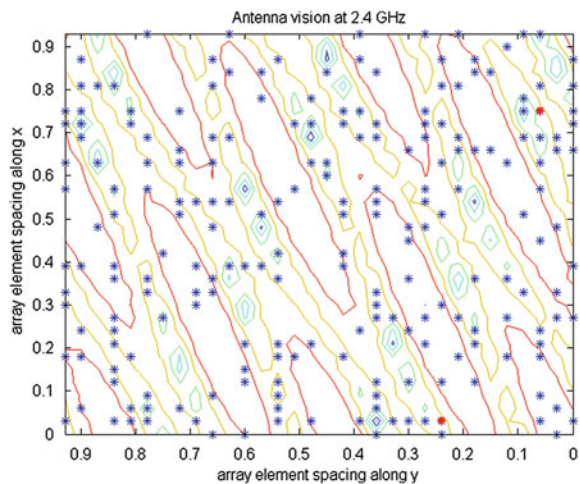


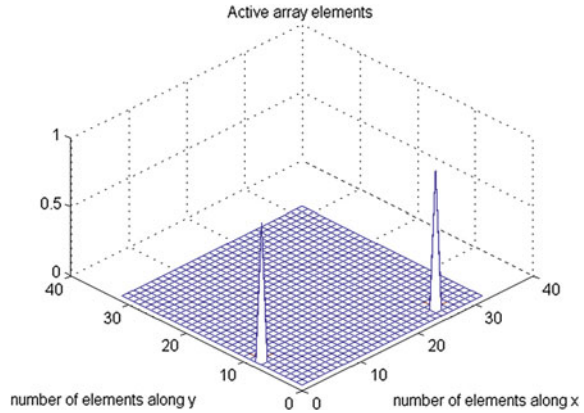
Fig. 5.6 Signal maxima detection based on the signal-fishing and antenna-vision mechanisms



maxima) and the blue stars indicate local maxima. The solution indicates the best coordinates (a, b) for receive element R_{x1} and (c, d) for R_{x2} . The corresponding activated array elements are shown in Fig. 5.7.

The gain of the signal-fishing mechanism compared to that of a fixed array (32×32) lies between 7 and 25 dB depending on the angles of arrival and on the search method.

Fig. 5.7 Array element activation for the signal-fishing mechanism



5.5 Conclusions

The premises for sustainable adaptive wireless receivers are set based on: (i) a systemic approach of the wireless receiver chain and (ii) the central role of the antenna in characterizing the transmission environment (radio scene). Maximum signal levels of the channel can be detected and used during operation by exploring the signal space provided through antenna vision mechanisms. Channel fading effects are mitigated starting at the antenna level—an approach that does not introduce additional noise like current baseband processing techniques.

The SF and antenna vision mechanisms bring a set of advantages: higher received SNR, no additional noise, higher AoA estimation accuracy. Thus the ‘observe’ part of the ‘observe-decide-act’ cycle is improved.

Future challenges in developing the CAS are: to develop the learning part of our antenna-based approach, extend the adaptive platform by correlating adaptive mechanisms from various levels of the processing chain, use other computational intelligence tools for radio scene characterization and learning.

Acknowledgments This paper was supported by CNCSIS-UEFISCDI, Romania, PD, project number 637/2010.

This work was also supported by the project “Develop and support multidisciplinary post-doctoral programs in primordial technical areas of national strategy of the research—development—innovation” 4D-POSTDOC, contract nr. POSDRU/89/1.5/S/52603, project co-funded from European Social Fund through Sectorial Operational Program Human Resources 2007-2013.

References

1. Mucchi L, Claudia Staderini J, Kyosti YP (2007) Modified spatial channel model for MIMO wireless systems. *EURASIP J Wirel Comm Netw* 2007:1–7
2. Crisan N, Cremene LC (2009) Antenna-based signal fishing, the fifth international conference on wireless and mobile communications—ICWMC’09, pp152–156, IEEE Computer Society Press, Cannes

3. Wang F, Ghosh A, Sankaran C, Fleming PJ, Hsieh F, Benes SJ (2008) Mobile WiMAX systems: performance and evolution. *IEEE Commun Mag* 46(10):41–49
4. 3GPP TS36.300 Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN), Overall description, (Stage 2 Release 8)
5. Dahlman E et al (2008) 3G evolution: HSPA and LTE for mobile broadband, 2nd edn., Academic Press, Salt Lake City
6. Stefania Sesia I, Baker TM (eds) (2009) LTE—the UMTS long term evolution. Wiley, New York, Ch
7. Parkvall S, Astely D (2009) The evolution of LTE towards IMT-advanced. *J Commun* 4(3):146–154
8. Doyle LE (2009) Essentials of cognitive radio. Cambridge University Press, New York
9. Crişan N, Cremene LC, Cremene M (2011) Software components for signal fishing based on GA element position optimizer. *Int J Comput Commun Control* 6(1):63–71, CCC Publications, Oradea
10. Tuncer E, Friedlander B (2009) Classical and modern direction of arrival estimation, Elsevier, Academic Press, Amsterdam
11. Rübsamen M, Gershman AB (2008) Root-MUSIC based direction-of-arrival estimation methods for arbitrary non-uniform arrays. In: *Proceedings ICASSP*, pp 2317–2320
12. Rübsamen M, Gershman AB (2009) Direction-of-arrival estimation for non-uniform sensor arrays: from manifold separation to Fourier domain MUSIC methods. *IEEE Trans Signal Process* 57(2):588–599

Chapter 6

Introducing the Concept of Information Pixels and the Storing Information Pixels Addresses Method as an Efficient Model for Document Storage

Mohammad A. ALGhalayini

Abstract Today, many institutions and organizations are facing serious problem due to the tremendously increasing size of documents, and this problem is further triggering the storage and retrieval problems due to the continuously growing space and efficiency requirements. This problem is becoming more complex with time and the increase in the size and number of documents in an organization. Therefore, there is a growing demand to address this problem. This demand and challenge can be met by developing a technique to enable specialized document imaging people to use when there is a need for storing documents images. Various techniques were developed and reported in the literature by different investigators. These techniques attempt to solve this problem to some extent but, most of the existing techniques still face the efficiency problem, in the case, the number and size of documents increase rapidly. The efficiency is further affected in the existing techniques when documents in a system are reorganized and then stored again. To handle these problems, we need special and efficient storage techniques for this type of information storage (IS) systems [1–7]. In this paper, we present an efficient storage technique for electronic documents. The proposed technique uses the Information Pixels concept to make the technique more efficient for certain image formats. In addition, we shall see how Storing Information Pixels Addresses (SIPA) method is an efficient method for document storage and as a result makes the document image storage relatively efficient for most image formats [8–12].

M. A. ALGhalayini (✉)

Developmental Programs Supervisor Assistant Vice Rectorate for Graduate Studies and Scientific Research, King Saud University, Riyadh, Saudi Arabia
e-mail: malghalayini@gmail.com

6.1 Introduction

In previous research paper,¹ we have analyzed and concluded that pdf image format is one of the best types to store scanned documents. Our next step was to divide the KSU² document into segments and store only the segment that contains the necessary information, namely, the body of the document. This excludes the header and footers, date and logo. This has proved to be efficient with respect to storage and saves at least 25 % storage space.³ In this research paper, we move a step further. We analyze the body segment of the document [13–15].

This is done by considering segment (5) of the KSU documents. In this segment, it is true with almost all documents that the shades are in black and white and there is no need for images to be stored neither as color images nor with higher memory usage, as is explained in previous papers.

The challenge is about how we can further reduce the storage size of the document. We should come up with a way that scrutinizes the document to a greater extent. This could be done by evaluating whether or not it is possible to make use of the fact that the whole document can be considered as an image, and since the information is in black, then the question should examine if we can store only the part of the segment that is black.

6.2 Analyzing the KSU Document Full Body Segment in Finer Detail

We segmented the KSU Document sheet into several segments depending on vital data which we actually need to store. These segments can be classified based on its content. Figure 6.1 below shows the segmentation pattern.

It is the body of the document that contains the necessary information. As explained previously, it is necessary to analyze this part in a way by which we can incidentally reduce the size of the document by considering the segment as an image.

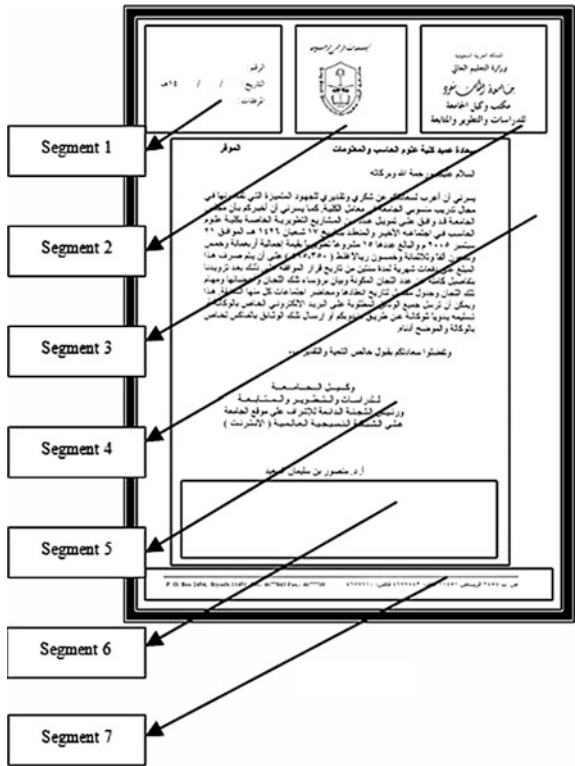
An image can be considered to consist of various shades. If the images are black and white, then the image can be considered as a gray scale image that

¹ Mohammad A. ALGhalayini and Abad Shah, “*Introducing The (POSSDI) Process The Process of Optimizing the Selection of The Scanned Document Images*”; CISSE 2006, The Second International Joint **Conferences** on Computer, Information, and Systems Sciences, and Engineering (CISSE 06).

² More about King Saud University available at: <http://www.ksu.edu.sa>.

³ Mohammad A. ALGhalayini and ELQasem ALNemah, “*An Efficient Storage and Retrieval Technique for Documents Using Symantec Document Segmentation (SDS) Approach*”; CISSE 2007, The Third International Joint **Conferences** on Computer, Information, and Systems Sciences, and Engineering (CISSE 07).

Fig. 6.1 The segmentation pattern



consists of (256) shades. When we further try to simplify this concept, we understand that the documents have written information and therefore can just be considered as black and white. The images are now considered at the pixel level. The advantages of considering these images at the pixel level are:

- The pixel is the building block of an image and therefore every image can be observed in minute detail based on the pixel values.
- It is the pixel values that the computer uses to store images. The number of bits used to represent pixel values can be manipulated and hence the size of the image.
- The spatial dimensions of a pixel can be predefined. The smaller the pixel, the better the resolution and the larger the pixel, the lesser the resolution. In addition, smaller pixels in an image with specific dimensions mean more number of pixels and therefore more storage space. Larger pixels in an image mean lesser number of pixels and less storage space; therefore, storage space and image quality is a tradeoff.

One of the best application softwares that can be used to check and simulate our proposed concept on images is MATLAB. We are going to use MATLAB to analyze the outcome of our logic on these images.

6.3 Introducing Storing Information Pixels Addresses Method

6.3.1 Concept of Pixel Based Segmentation

A typical KSU sheet is shown in Fig. 6.1. We make note of a few facts: The information written in the KSU sheet is in black. It does not contain any images. We observe that when the images are stored in any format, there is lots of space used up by the white portion of the sheet, which, apparently, stores no information and may also lead to an increase in memory usage.

This portion is not of any importance to us; therefore, we need a way to only store the information, that is, nothing but the information pixels which in turn is the black portion in the image. We shall call each black pixel an **information pixel**.

The concept of Storing Information Pixels Addresses (SIPA) will help save memory space. This can be done by taking into account only the pixels that are black in color which are the information pixels. It is important to consider the image as an ensemble of black and white shades of 256 types. This is obtained from the fact that every pixel is represented by 8 unsigned bits in MATLAB. We have to observe the image on a grayscale.

Another point to be made note of is that MATLAB does not interpret images based on colors. It does so based on the intensity of pixels. If the intensity of the pixel is zero, then it is black. If the intensity of a pixel is 255 which is the maximum value of an 8-bit pixel, it is white. We use this concept for identifying colors in MATLAB using pixel intensities.

The MATLAB program below is for a jpeg format image. This program is used to store the addresses of information pixels.

```
% Provides information of the image
Imfinfo ('C:\Users\ALGhalayini \Desktop\sample.jpg');
% Start counter to keep a tab of the time
Tic;
% Read the image in MATLAB
RGBImage=imread ('C:\Users\ALGhalayini \Desktop\sample.jpg','jpg')
% Convert it to gray format
GRAYImage=rgb2gray (RGBImage);
% Display the image
Figure (6.1);
Imshow (GRAYImage);
% Find the dimensions of the image
[m n]=size (GRAYImage);
total_pixels=m * n ;
% Store address of information pixels in Array1 and Array2
K=1;
```

```

for I=1 : m
for j=1 : n
if GRAYimage (I, j) < 128
Array1 (I, k)=I; Array2 (I, k)=j;
K=k + 1;
end
end
end
%Form N X 2 array that stores the address of information pixels
Array=[Array1' Array2'];
% End of time counter
Toc;

```

First, we execute the *imfinfo* command and make sure that the image format we are using is compatible with MATLAB. The image is then read into MATLAB using the *imread* command.

This command reads the image as a RGB image. The RGB color model is *additive* in nature. That is, when R, G and B light beams are added together, their light spectra add to make the final color's spectrum. Zero intensity for each component gives the darkest color (no light-considered *black*), and full intensity of each gives a white; the *quality* of this white depends on the nature of the primary light sources. When the intensities for all the components are the same, the result is a shade of gray, darker or lighter depending on the intensity. When the intensities are different, the result is a colorized. Since our image is black and white, we can use an image model that uses equal intensities of the 3 primary colors to produce shades between black and white. This is called the grayscale format.

6.3.2 Thresholding and Image Enhancement

Thresholding is the simplest form of image segmentation. It is a form of scaling the intensities in the image. It can be done on grayscale images.

During the thresholding process, individual pixels in an image are marked as "object" pixels if their value is greater than some threshold value (assuming an object to be brighter than the background) and as "background" pixels otherwise. Typically, an object pixel is given a value of "1" while a background pixel is given a value of "0." Finally, a binary image is created by coloring each pixel white or black, depending on a pixel's label. Since the grayscale image has values from 0 to 255, we can consider the threshold value to be 127, halfway between the highest and lowest intensities. In this scale, pixel values above 127 are considered white and below 127 are considered black. This process of thresholding also enhances the image because it removes unwanted gray shades in the image formed during scanning.

Table 6.1 Images with specific formats using MATLAB

| No. | Image type | Type meaning |
|-----|------------|----------------------------------|
| 1 | BMP | Windows bitmap |
| 2 | CUR | Cursor file |
| 3 | GIF | Graphics interchange format |
| 4 | HDF4 | Hierarchical data format |
| 5 | ICO | Icon file |
| 6 | JPEG | Joint photographic experts group |
| 7 | PBM | Portable bitmap |
| 8 | PCX | Windows paintbrush |
| 9 | PGM | Portable graymap |
| 10 | PNG | Portable network graphics |
| 11 | PPM | Portable pixmap |
| 12 | RAS | Sun raster |
| 13 | TIFF | Tagged image file format |
| 14 | XWD | X Window dump |

Let m be the number of pixels row wise and n is the number of pixels column wise. MATLAB has the ability to handle images with specific formats. These formats are presented in Table 6.1.

The execution time is recorded in MATLAB for each format. The code for JPEG format is displayed above. Observe that there should be only small differences in the code. It is in the part where the image is read. MATLAB must be given information about the type of format of the image it is supposed to read. This can be accomplished by replacing the filename in the *imread* command.

6.4 Applying Storing Information Pixels Addresses Method on the Whole KSU A4 Document Images

6.4.1 Program Algorithm

The program algorithm shown below is for the Whole KSU A4 document sheet.

1. clear all; close all; clc;
 - In this line, clear all: clears the memory and hence does not store values of variables used in any of the previous commands.
 - Close all: closes the execution of the previous commands.
 - Clc: Clears the screen.
2. The image is then converted from RGB format to Gray scale format.
3. The number of rows is taken to be m and the number of columns is taken to be n .
4. The total number of pixels is $m*n$.
5. for $I = 1 : m$


```

for j=1 : n
if GRAYimage (I, j) < 128
Array1 (I, k)=i; Array2 (I, k)=j;
K=k + 1;
end
end
end

```

The for-loops help check row-wise, and then column-wise, the value of the pixel. If the value of the pixel is less than 128, it is an information pixel (is black) and its address is stored in an array. Both the row and column addresses are stored. The arrays, Array1 and Array2 are of type double.

6. Array1 and Array2 are together combined to form Array that is a 2-dimensional array. The first column of the array represents the row address and the second column of the array represents the column address corresponding to each information pixel.
7. Tic and Toc are commands that are used to record the time taken for the whole execution.

6.4.2 Execution Results and Analysis

Table 6.2 represents the result of the analysis applied on the Whole KSU A4 page. It also shows the computed result using MATLAB. It is executed for four image format types:

- Jpeg
- Tif
- Bmp
- Gif

The result for these four formats is tabulated. The first column in Table 6.2 indicates the image formats on which the MATLAB program have been executed. The second column indicates the pixels per inch (resolution). For every image type, we execute the code for four different pixel densities 75, 100, 150, and 200. The pixel densities vary from less resolution with merely intelligible information to more resolution with better clarity. Third column shows the size of the original image in kilobytes corresponding to the image format in column 1 and pixel density in column 2. The fourth column shows the time in seconds taken to complete the execution. This process includes reading the image, locating and storing the address of information pixels. The fifth column indicates the total number of information pixels. The sixth column indicates the array dimension. The seventh column indicates the size of the resulting array. As indicated in the sixth

Table 6.2 The computed result using MATLAB

| Image format | dpi | Original image size(in KB) | Execution time (in s) | Number of information pixels | Array dimensions | Array size (in KB) | Compression ratio | Average compression ratio |
|--------------|-----|----------------------------|-----------------------|------------------------------|------------------|--------------------|-------------------|---------------------------|
| Jpeg | 75 | 80.00 | 2.43 | 544,617 | 3,728 | X 2 | 0.687 | 0.73 |
| | 100 | 112.00 | 2.17 | 981,618 | 7,719 | X 2 | 0.929 | |
| | 150 | 240.00 | 10.50 | 2,208,015 | 21,849 | X 2 | 0.703 | |
| | 200 | 384.00 | 31.22 | 3,926,472 | 41,548 | X 2 | 0.592 | |
| Bmp | 75 | 1,600.00 | 1.09 | 544,617 | 3,570 | X 2 | 28.687 | 22.53 |
| | 100 | 2,880.00 | 2.38 | 981,618 | 7,551 | X 2 | 24.412 | |
| | 150 | 6,480.00 | 11.23 | 2,208,015 | 21,633 | X 2 | 19.171 | |
| | 200 | 11,504.00 | 32.35 | 3,926,472 | 41,214 | X 2 | 17.864 | |
| Gif | 75 | 144.00 | 1.89 | 544,617 | 5,928 | X 2 | 1.555 | 1.23 |
| | 100 | 240.00 | 4.80 | 981,618 | 11,873 | X 2 | 1.294 | |
| | 150 | 496.00 | 19.60 | 2,208,015 | 30,092 | X 2 | 1.055 | |
| | 200 | 864.00 | 79.85 | 3,926,472 | 55,066 | X 2 | 1.004 | |
| Tif | 75 | 528.00 | 1.00 | 544,617 | 3,570 | X 2 | 9.467 | 6.14 |
| | 100 | 704.00 | 2.13 | 981,618 | 7,551 | X 2 | 5.967 | |
| | 150 | 1,616.00 | 10.31 | 2,208,015 | 21,633 | X 2 | 4.781 | |
| | 200 | 2,800.00 | 25.61 | 3,926,472 | 41,214 | X 2 | 4.348 | |

column, the array has two dimensions, one that represents the row address and the other that represents the column address.

Example: If an array has the element [1790, 34], it means that the 1790th row and 34th column is an information pixel. Since the number of pixels either row wise or column wise for even the least pixel density is greater than 255, the pixel addresses have to be represented with numbers greater than 255 at times and is therefore not of type integer whose range is from 0 to 255 but is of type double which indicates double precision and can go up to 65536. The size of the array is calculated by using the formula:

Size = $\varepsilon * 2 * 8 / (1024)$ Formula (6-1)

ε indicates the array dimension. Since each pixel is associated with a row and column pixel address, each of which are of type double, there are 2 integers of type double associated with every address. Every pixel address has a size of $2 * 8$ bytes because each number in the array is of type double which is 8 bytes. It has to be multiplied by the length of the array to get the whole array size. This is then divided by 1024 to get the size in KB. The eighth column gives the ratio of original image size to array size. If the value in this column is greater than 1, it means that the size of the array is smaller than the size of the image and using the concept of information pixels turns out to be successful. If the value is less than 1, it means that the size of the array is greater than the size of the image and the method fails in compressing the image further. The last column indicates the average percentage of compression for a specific format that averages over all pixel densities. The ratio of array to image compression is given by:

Ratio = Array Size/Image Size * 100

A few important observations that can be made from the table are:

- As the pixel density increases, the storage size of the image increases as a result of more pixels in the image.
- As the number of pixels and storage size increases, the execution time increases because the computational complexity increases.⁴
- The number of elements in the array used to store addresses of information pixels increases with increase in number of pixels or increase in dpi. This is because, as the dpi increases, pixels become smaller and represents smaller portions of the image.

For the JPEG image format, we observe that the applying SIPA method does not turn out to be successful. The average compression is 0.73 which means that

⁴ Even though time of execution time is shown in the table, we do not focus much on it since time may be affected by other factors like processor speed, RAM size, network reliability, Graphics memory size, etc.

the size of the array is 1/0.73 times the size of the image, that is, the size of the array is larger than the size of the image. Therefore, using SIPA method for JPEG image formats is not a good idea. Also, we can see that as the dpi increases, the compression ratio gets smaller. For 75 dpi, the array is 1/0.687 times the original image while for 200 dpi, the array is 1/0.592 times the original image. This shows that the inefficiency of this technique increases with increase in pixel density.

For the BMP image format, we observe that applying SIPA method of storing only information pixels is highly efficient. The average compression is 22.53 which means that the size of the array is just about 1/22.53 times the size of the original image, that is, the size of the array is very small compared to the size of the image. Therefore, using this scheme for BMP image formats is a good idea. Also, we can see that as the dpi increases, the compression ratio decreases. For 75 dpi, the compression ratio is 28.687 while for 200 dpi, the compression ratio is 17.864. This says that the efficiency of this technique decreases with increase in pixel density.

For the GIF image format, we observe that applying SIPA method does turns out to be good. The average compression ratio is 1.23 which means that the size of the array is about 1/1.23 times the size of the original image, that is, the array size is a little smaller than the image size. Therefore, using this scheme for GIF image formats is acceptable. Also, we can see that as the dpi increases, the compression ratio decreases. For 75 dpi, the compression ratio is 1.555 while for 200 dpi, the array to image size ratio is 1.004. This says that the efficiency of this technique decreases with increase in pixel density.

For the TIF image formats, we observe that applying SIPA method is very successful. The average compression is 6.14 which means that the size of the array is about 1/6.14 the size of the original image, that is, the size of the array is smaller than the size of the image. Therefore, using this scheme for TIF is a good idea. Also, we can see that as the dpi increases, the compression ratio decreases. For 75 dpi, the compression ratio is 9.467 while for 200 dpi, the compression ratio is 4.348. This says that the efficiency of this technique decreases with increase in pixel density.

From Table 6.2, we observe that irrespective of the image format, the array size for a specific dpi is almost the same.

This can be explained rationally as follows. When we work on images pixel wise in MATLAB, the location of information pixels in different image formats does not change. Therefore, the number of information pixels remains almost the same taking into consideration the fact that for a specific dpi, the pixel dimension is the same. The proof to this statement can be observed from the table. For example, for 75 dpi, the array dimension is 3728 for the JPEG, 3570 for both the BMP and the TIF, and it is 5928 in the GIF.

Table 6.3 Computed size of the array for all image formats for specific dpi

| dpi | Median array dimension | Array size (in KB) |
|-----|------------------------|--------------------|
| 75 | 3,649 | 57.01 |
| 100 | 7,635 | 119.29 |
| 150 | 21,741 | 339.70 |
| 200 | 41,381 | 646.57 |

6.5 Generalizing the Concept of Storing Information Pixels Addresses over all Image Formats

The disparity of the number of information pixels in GIF image can be attributed to the fact that GIF images are structured in a different way when compared to the other three image formats.

The images in other image formats have an R, G and B scale whereas GIF has only one dimension. We can now consider the median value as the number of information pixels in any image format in general. We are taking the median and not the average because we can avoid the disparity in values caused by the GIF image. The median can be calculated as follows.

Consider the values in ascending order:
3,728, 3,570, 3,570, 5,928

The middle value is the median. If the number of elements during the calculation of the mean is even, then we consider the average of the two middle values. Here the mean is therefore 3649. Similarly, we can calculate the mean for 100, 150 and 200 dpi.

We now plot a table to compute the size of the array for all image formats for specific dpi (Table 6.3).

We can now use this generalization of array sizes to all image formats. We now plot a table for the rest of the considered image formats (Table 6.4).

For the EXE image format, we observe that applying SIPA method turns out to be successful for dpi 75, 100, 150, and 200. The average compression ratio is 4.22, that is, size of the array is about (1/4.22) times the size of the original image, that is, the size of the array is smaller than the size of the image. Therefore, using this scheme for EXE images is a good idea. Also, we can see that as the dpi increases, the compression ratio decreases. For 75 dpi, the array to image size ratio is 8.7 while for 200 dpi, the array to image size ratio is 1.46. This says that the technique is acceptable when used for 200 dpi. As discussed in (Chap. 2), 150 dpi has an acceptable image quality and therefore this method can be applied on EXE image formats with 150 dpi for which it is successful.

For the FPX image format, we observe that applying SIPA method is efficient for all dpi except the 200. The average compression is about 1.66 which means that the size of the array is 1/1.66 times the size of the original image, that is, the array size is smaller than the image size. Therefore, using this scheme for FPX

Table 6.4 The rest of the considered image formats

| Image format | dpi | Image size (in KB) | Array size (in KB) | Compression ratio | Average compression ratio |
|--------------|-----|--------------------|--------------------|-------------------|---------------------------|
| .exe | 75 | 496 | 57.01 | 8.70 | 4.22 |
| | 100 | 544 | 119.2 | 4.56 | |
| | 150 | 736 | 339.7 | 2.17 | |
| | 200 | 944 | 646.5 | 1.46 | |
| .fpx | 75 | 160 | 57.01 | 2.81 | 1.66 |
| | 100 | 208 | 119.2 | 1.74 | |
| | 150 | 384 | 339.7 | 1.13 | |
| | 200 | 608 | 646.5 | 0.94 | |
| .htm | 75 | 192 | 57.01 | 3.37 | 2.08 |
| | 100 | 256 | 119.2 | 2.15 | |
| | 150 | 528 | 339.7 | 1.55 | |
| | 200 | 800 | 646.5 | 1.24 | |
| .max | 75 | 144 | 57.01 | 2.53 | 1.49 |
| | 100 | 192 | 119.2 | 1.61 | |
| | 150 | 336 | 339.7 | 0.99 | |
| | 200 | 544 | 646.5 | 0.84 | |
| .pdf | 75 | 96 | 57.01 | 1.68 | 1.11 |
| | 100 | 144 | 119.2 | 1.21 | |
| | 150 | 288 | 339.7 | 0.85 | |
| | 200 | 464 | 646.5 | 0.72 | |
| .png | 75 | 512 | 57.01 | 8.98 | 5.73 |
| | 100 | 640 | 119.2 | 5.37 | |
| | 150 | 152 | 339.7 | 4.47 | |
| | 200 | 264 | 646.5 | 4.08 | |
| tif class f | 75 | 32 | 57.01 | 0.56 | 0.29 |
| | 100 | 32 | 119.2 | 0.27 | |
| | 150 | 64 | 339.7 | 0.19 | |
| | 200 | 96 | 646.5 | 0.15 | |
| tif group 4 | 75 | 32 | 57.01 | 0.56 | 0.29 |
| | 100 | 32 | 119.2 | 0.27 | |
| | 150 | 64 | 339.7 | 0.19 | |
| | 200 | 96 | 646.5 | 0.15 | |
| tif lzw | 75 | 528 | 57.01 | 9.26 | 6.06 |
| | 100 | 704 | 119.29 | 5.90 | |
| | 150 | 161 | 339.70 | 4.76 | |
| | 200 | 280 | 646.57 | 4.33 | |
| tif uncomp | 75 | 160 | 57.01 | 28.06 | 22.28 |
| | 100 | 288 | 119.2 | 24.14 | |
| | 150 | 648 | 339.7 | 19.08 | |
| | 200 | 115 | 646.5 | 17.84 | |
| dcx | 75 | 131 | 57.0 | 23.01 | 14.44 |
| | 100 | 150 | 119.2 | 12.61 | |
| | 150 | 387 | 339.7 | 11.40 | |
| | 200 | 694 | 646.5 | 10.74 | |

(continued)

Table 6.4 (continued)

| Image format | dpi | Image size (in KB) | Array size (in KB) | Compression ratio | Average compression ratio |
|--------------|-----|--------------------|--------------------|-------------------|---------------------------|
| pcx | 75 | 13 | 57.0 | 23.01 | 14.44 |
| | 100 | 15 | 119.2 | 12.61 | |
| | 150 | 38 | 339.7 | 11.40 | |
| | 200 | 69 | 646.5 | 10.74 | |

images is acceptable. Also, we can see that as the dpi increases, the compression ratio gets smaller. For 75 dpi, the array to image size ratio is 1.74 while for 200 dpi, the array to image size ratio is 0.94.

For the HTM image format, we observe that applying SIPA method turns out to be successful for all dpi. The average compression is 2.08 which means that the size of the array is about 1/2.08 times the size of the original image, that is, the size of the array is about half of the size of the image. Therefore, using this scheme for HTM images is a good idea. Also, we can see that as the dpi increases, the compression ratio reduces. For 75 dpi, the array to image size ratio is 3.37 while for 200 dpi, the array to image size ratio is 1.24. This says that the technique is efficient when used for 75, 100, 150, and 200 dpi.

For the MAX image format, we observe that applying SIPA method is efficient only for 75 and 100 dpi. The average compression is about 1.49 which means that the size of the array is 1/1.49 times the size of the original image, that is, the array size is smaller than the image size; however, using this scheme for MAX images is not a good idea since images with resolution 75 and 100 dpi are unacceptable due to their quality and it is only for those dpi that this method is successful. Also, we can see that as the dpi increases, the compression ratio gets smaller. For 75 dpi, the array to image size ratio is 2.53 while for 200 dpi, the array to image size ratio is 0.84.

Although we previously stated that the PDF image format is preferred due to its relatively small size and better image quality, we observe that applying SIPA method is not very efficient but it is just acceptable. The average compression is 1.11 which means that the size of the array is about 1/1.11 times the size of the original image, that is, the size of the array is little smaller than the size of the image. Also, we can see that as the dpi increases, the compression ratio reduces. For 75 dpi, the array to image size ratio is 1.68 while for 200 dpi, the array to image size ratio is 0.72.

For the PNG image format, we observe that applying SIPA method is efficient. The average compression is 5.73 which means that the size of the array is about 1/5.73 times the size of the original image, that is, the size of the array is smaller than the size of the image. Also, we can see that as the dpi increases, the compression ratio reduces. For 75 dpi, the array to image size ratio is 8.98 while for 200 dpi, the array to image size ratio is 4.08.

For the TIF CLASS F and the TIF GROUP 4 image formats, we observe that the SIPA method is not efficient. The average compression is 0.29 which means that the size of the array is about $1/0.29$ times the size of the original image, that is, the size of the array is larger than the size of the image. Also, we can see that as the dpi increases, the compression ratio reduces. For 75 dpi, the array to image size ratio is 0.56 while for 200 dpi, the array to image size ratio is 0.15.

For the TIF LZW image format, we observe that applying SIPA method is efficient. The average compression is 6.06 which means that the size of the array is about $1/6.06$ times the size of the original image, that is, the size of the array is larger than the size of the image. Also, we can see that as the dpi increases, the compression ratio reduces. For 75 dpi, the array to image size ratio is 9.26 while for 200 dpi, the array to image size ratio is 4.33.

For the TIF UNCOMPRESSED image format, we observe that applying SIPA method is efficient. The average compression is 22.28 which means that the size of the array is about $1/22.28$ times the size of the original image, that is, the size of the array is larger than the size of the image. Also, we can see that as the dpi increases, the compression ratio reduces. For 75 dpi, the array to image size ratio is 28.06 while for 200 dpi, the array to image size ratio is 17.84.

For the DCX, and the PCX image formats, we observe that the SIPA method is efficient. The average compression is 14.44 which means that the size of the array is about $1/14.44$ times the size of the original image, that is, the size of the array is larger than the size of the image. Also, we can see that as the dpi increases, the compression ratio reduces. For 75 dpi, the array to image size ratio is 23.01 while for 200 dpi, the array to image size ratio is 10.74.

TIFF MULTI-PAGE CLASS F and TIFF MULTI-PAGE CLASS 4 have the same compression ratio and properties as TIFF CLASS F and TIFF CLASS 4 respectively. TIFF MULTI-PAGE LZW and TIFF MULTI-PAGE have the same compression ratio and properties as TIFF LZW. TIFF MULTI-PAGE UNCOMPRESSED has the same compression ratio and properties as TIFF UNCOMPRESSED.

6.6 Applying the Storing Information Pixels Addresses Method on the Full Body Segment

The only change in this program when compared to the previous program is the consideration of addresses for the information pixels only from the full body segment. In this paper, the concept of segmentation helps reduce memory occupied by information pixels because only segment (5) is considered and the information pixels in the other segments are ignored. This way, the total number of information pixels in the whole image reduces and hence the memory required for its storage. For this to be executed, we need to define the pixel boundaries for segment (5). This can be explained as follows: Length of the A4 sheet: 29.7 cm. Width of the A4 sheet is 21 cm.

6.7 Defining Pixel Boundaries for the Full Body Segment

The information pixels in segment (5) start 7.7 cm from the top of the sheet. Let 'm' represent the address of row pixels. Since the resolution of the image is 150 dpi and 1 cm = 0.39370078 inches, the start address of m is:

$$7.7 * 150 * 0.39370078 = 454.$$

The information pixels of segment (5) end 5 cm above the bottom of the sheet, which is 24.7 cm from the top of the sheet. Since the resolution of the image is 150 dpi and 1 cm = 0.39370078 inches, the end address of m is:

$$24.7 * 150 * 0.39370078 = 1459.$$

Let n represent the address of the column pixels. The information pixels in segment (5) start 2 cm to the right of the left end of the A4 sheet. Since the resolution of the image is 150 dpi and 1 cm = 0.39370078 inches, the start address of n is:

$$2 * 150 * 0.39370078 = 118.$$

The information pixels in segment (5) end 2 cm to the left of the right end of the A4 sheet, which is 19 cm to the right of the left end of the sheet. Since the resolution of the image is 150 dpi and 1 cm = 0.39370078, the end address of n is

$$19 * 150 * 0.39370078 = 1122.$$

Table 6.5 shows the starting and ending row and column pixels for different dpi as computed above.

6.7.1 Program Algorithm

A typical program for a JPEG image with 150 dpi that works only on the body of the KSU sheet is shown below.

```
clear all; close all; clc;
tic;
RGBImage=imread('C:\Users\ALGhalayini\Desktop\algha\A4_size.jpg\full page
256 colors 8bit 150.jpg','jpg');
GRAYImage=rgb2gray (RGBImage);
Figure (1);
Imshow (GRAYImage (454:1459,118:1122));
M=length (GRAYImage);
N=length (GRAYImage (1, :));
total_pixels=m * n;
k=1;
for I=454 : 1459
for j=118 : 1122
```

Table 6.5 The starting and ending row and column pixels for different dpi

| dpi | Starting row | Ending row | Starting column | Ending column |
|-----|--------------|------------|-----------------|---------------|
| 75 | 227 | 729 | 59 | 561 |
| 100 | 303 | 972 | 79 | 748 |
| 150 | 454 | 1459 | 118 | 1122 |
| 200 | 606 | 1944 | 157 | 1496 |

```

if GRAYimage (I, j) < 128
Array1(I, k)=I; Array2 (I,k)=j;
K=k + 1;
end
end
end
Array=[Array1' Array2'];
toc

```

We observe that the program is different from that of the JPEG 150 dpi A4 image size program in the 'for' loop. The snippet is shown below:

```

for I=454 : 1459
for j=118 : 1122
if GRAYimage (I, j) < 128
Array1 (I, k)=I; Array2 (I, k)=j;
K=k + 1;
end
end
end

```

The concept of selecting the starting and ending row and column which is nothing but the range of values of i and j is explained in Table 6.5. We now record the observations from these executions and display them in a table as below.

6.7.2 Execution Results and Analysis

Table 6.6 represents the result of the analysis applied on the Full Body Segment. It also shows that the technique of storing the information pixels addresses (SIPA) is successful for the bmp, the gif, and the tif image formats. We observe that Table 6.6 is very similar to Table 6.2. The only difference is that we consider the whole A4 KSU scanned document page and search for information pixels in Table 6.2 while we consider only the Full Body Segment of the A4 KSU scanned document page and search for information pixels in Table 6.6. Column 3 in Table 6.6 is the image size. Column 7 in Table 6.6 indicates the size of the array formed by storing information pixels contained in the body of the KSU document page while column 7 in Table 6.2 indicates the size of the array formed by storing information pixels contained in the whole KSU document page. By intuition, we

Table 6.6 The result of the analysis applied on the Full Body Segment

| Image format | dpi | Image size (in KB) | Execution time (in s) | Number of information pixels | Array dimensions | Array size (in KB) | Compression ratio | Average compression ratio |
|--------------|-----|--------------------|-----------------------|------------------------------|------------------|--------------------|-------------------|---------------------------|
| jpeg | 75 | 60 | 0.86 | 252,004 | 3,453.5 | X 2 | 53.96 | 1.11 |
| | 100 | 84 | 1.96 | 447,561 | 7,154.5 | X 2 | 111.79 | 0.75 |
| | 150 | 180 | 8.22 | 1,009,020 | 19,407.5 | X 2 | 303.24 | 0.59 |
| | 200 | 288 | 17.16 | 1,791,582 | 35,726 | X 2 | 558.22 | 0.52 |
| bmp | 75 | 992 | 0.98 | 252,004 | 3,313.5 | X 2 | 51.77 | 15.11 |
| | 100 | 1,744 | 1.94 | 447,561 | 7,022 | X 2 | 109.72 | 15.90 |
| | 150 | 3,904 | 8.22 | 1,009,020 | 19,304.5 | X 2 | 301.63 | 12.94 |
| | 200 | 6,944 | 16.91 | 1,791,582 | 35,696.5 | X 2 | 557.76 | 12.45 |
| gif | 75 | 96 | 1.05 | 252,004 | 3,928.5 | X 2 | 61.38 | 1.56 |
| | 100 | 144 | 2.37 | 447,561 | 7,987.5 | X 2 | 124.80 | 1.15 |
| | 150 | 304 | 9.19 | 1,009,020 | 20,360 | X 2 | 318.13 | 0.96 |
| | 200 | 592 | 17.10 | 1,791,582 | 36,695.5 | X 2 | 573.37 | 1.03 |
| tif | 75 | 304 | 0.84 | 252,004 | 3,313.5 | X 2 | 51.77 | 5.87 |
| | 100 | 496 | 1.87 | 447,561 | 7,022 | X 2 | 109.72 | 4.52 |
| | 150 | 992 | 8.18 | 1,009,020 | 19,304.5 | X 2 | 301.63 | 3.29 |
| | 200 | 2,912 | 17.24 | 1,791,582 | 35,696.5 | X 2 | 557.76 | 5.22 |

Table 6.7 Calculation of the median array dimension and the array size

| dpi | Median array dimension | Array size (in KB) |
|-----|------------------------|--------------------|
| 75 | 1,691.75 | 26.43 |
| 100 | 3,544.13 | 55.38 |
| 150 | 9,678.00 | 151.22 |
| 200 | 17,855.63 | 278.99 |

know that the size of the array used for storing information pixels addresses of the whole KSU document page should be larger than the size of the array used for storing information pixels of the Body Segment of KSU document page. This intuition is verified when we see that values in column 7 in Table 6.6 are lesser than corresponding values in Table 6.2.

**6.8 Generalization the Concept of Storing Information
Pixels Addresses over all Image Formats**

We now use **Formula (1)** to calculate the median array dimension and the array size and generalize it for all image formats as we did previously for the A4 size image formats (Table 6.7).

We now use the generalized array sizes for different dpi for all image types as explained previously for the A4 size image formats.

We recall that for all image formats and all dpi, whenever the compression ratio is less than 1, It implies that the array sizes formed by pixel based segmentation are larger than the image itself and therefore applying the SIPA technique is unsuccessful.

Table 6.8 shows that applying SIPA technique is successful for all considered image formats except the TIF Class F and TIF Class 4. It also shows the array to image average percentage for various image formats when scanned over the Full Body Segment only. TIFF MULTI-PAGE CLASS F and TIFF MULTI-PAGE CLASS 4 has the properties of TIFF CLASS F or TIFF CLASS 4.

TIFF MULTI-PAGE LZW and TIFF MULTI-PAGE has properties of TIFF LZW. TIFF MULTI-PAGE UNCOMPRESSED has properties of TIFF UNCOMPRESSED.

- From Tables 6.2 and 6.6 shown above, we note few important observations:
- As seen from the table, the time elapsed for all the cases for execution of the A4 image size is more than the time elapsed for execution of full body segments.
 - It is also observed that as the dpi increases, the time elapsed during the execution of the Full Body Segment is comparatively lesser than the time elapsed during the execution of A4 size image of the same dpi. This can be explained from the fact that the MATLAB code needs to run over lesser number of pixels during the execution of the Full Body Segment than the A4 size image.

Table 6.8 The applying SIPA technique

| Image format | dpi | Image size | Array size | Compression ratio | Average comp.ratio |
|-----------------------------|-----|------------|------------|-------------------|--------------------|
| .exe | 75 | 46 | 26.43 | 17.55 | 8.53 |
| | 10 | 51 | 55.38 | 9.25 | |
| | 15 | 64 | 151.2 | 4.23 | |
| | 20 | 86 | 278.9 | 3.10 | |
| .fpx | 75 | 12 | 26.43 | 4.54 | 2.72 |
| | 10 | 15 | 55.38 | 2.82 | |
| | 15 | 28 | 151.2 | 1.90 | |
| | 20 | 45 | 278.9 | 1.63 | |
| .htm | 75 | 16 | 26.43 | 6.05 | 3.73 |
| | 10 | 20 | 55.38 | 3.76 | |
| | 15 | 40 | 151.2 | 2.65 | |
| | 20 | 68 | 278.9 | 2.47 | |
| .max | 75 | 11 | 26.43 | 4.24 | 2.58 |
| | 10 | 16 | 55.38 | 2.89 | |
| | 15 | 24 | 151.2 | 1.59 | |
| | 20 | 44 | 278.9 | 1.61 | |
| .pdf | 75 | 72 | 26.43 | 2.72 | 1.84 |
| | 100 | 108 | 55.38 | 1.95 | |
| | 150 | 216 | 151.22 | 1.43 | |
| | 200 | 348 | 278.99 | 1.25 | |
| .png | 75 | 336 | 26.43 | 12.71 | 9.67 |
| | 100 | 512 | 55.38 | 9.25 | |
| | 150 | 976 | 151.22 | 6.45 | |
| | 200 | 286 | 278.99 | 10.27 | |
| tif class f | 75 | 16 | 26.43 | 0.61 | 0.46 |
| | 100 | 32 | 55.38 | 0.58 | |
| | 150 | 48 | 151.22 | 0.32 | |
| | 200 | 96 | 278.99 | 0.34 | |
| tif group 4 | 75 | 16 | 26.43 | 0.61 | 0.48 |
| | 100 | 32 | 55.38 | 0.58 | |
| | 150 | 48 | 151.22 | 0.32 | |
| | 200 | 112 | 278.99 | 0.40 | |
| tiflzw | 75 | 304 | 26.43 | 11.50 | 9.36 |
| | 100 | 496 | 55.38 | 8.96 | |
| | 150 | 992 | 151.22 | 6.56 | |
| | 200 | 291 | 278.99 | 10.44 | |
| tif _{uncompressed} | 75 | 992 | 26.43 | 37.53 | 29.97 |
| | 100 | 174 | 55.38 | 31.49 | |
| | 150 | 392 | 151.22 | 25.92 | |
| | 200 | 696 | 278.99 | 24.95 | |
| dcx | 75 | 752 | 26.43 | 28.45 | 23.76 |
| | 100 | 116 | 55.38 | 21.09 | |
| | 150 | 230 | 151.22 | 15.24 | |
| | 200 | 844 | 278.99 | 30.28 | |
| pcx | 75 | 752 | 26.43 | 28.45 | 23.76 |
| | 100 | 116 | 55.38 | 21.09 | |
| | 150 | 230 | 151.22 | 15.24 | |
| | 200 | 844 | 278.99 | 30.28 | |

Table 6.9 Applying SIPA method

| Image format | dpi | Original image size (in KB) | Array size (in KB) | Memory saved using SIPA | Percentage of memory saved using SIPA (%) | Average percentage of memory saved using SIPA (%) |
|--------------|-----|-----------------------------|--------------------|-------------------------|---|---|
| jpeg | 75 | 80.00 | 58.25 | 21.75 | 27.19 | -22.95 |
| | 100 | 112.00 | 120.60 | -8.60 | -7.68 | |
| | 150 | 240.00 | 341.38 | -101.38 | -42.24 | |
| | 200 | 384.00 | 649.18 | -265.18 | -69.06 | |
| bmp | 75 | 1600.00 | 55.77 | 1544.23 | 96.51 | 95.40 |
| | 100 | 2880.00 | 117.98 | 2762.02 | 95.90 | |
| | 150 | 6480.00 | 338.02 | 6141.98 | 94.78 | |
| | 200 | 11504.00 | 643.97 | 10860.03 | 94.40 | |
| gif | 75 | 144.00 | 92.62 | 51.38 | 35.68 | 16.00 |
| | 100 | 240.00 | 185.51 | 54.49 | 22.71 | |
| | 150 | 496.00 | 470.18 | 25.82 | 5.21 | |
| | 200 | 864.00 | 860.40 | 3.60 | 0.42 | |
| tif | 75 | 528.00 | 55.77 | 472.23 | 89.44 | 82.19 |
| | 100 | 704.00 | 117.98 | 586.02 | 83.24 | |
| | 150 | 1616.00 | 338.02 | 1277.98 | 79.08 | |
| | 200 | 2800.00 | 643.97 | 2156.03 | 77.00 | |
| .exe | 75 | 496 | 57.01 | 438.99 | 88.51 | 62.98 |
| | 100 | 544 | 119.29 | 424.71 | 78.07 | |
| | 150 | 736 | 339.70 | 396.30 | 53.85 | |
| | 200 | 944 | 646.57 | 297.43 | 31.51 | |
| .fpx | 75 | 160 | 57.01 | 102.99 | 64.37 | 28.05 |
| | 100 | 208 | 119.29 | 88.71 | 42.65 | |
| | 150 | 384 | 339.70 | 44.30 | 11.54 | |
| | 200 | 608 | 646.57 | -38.57 | -6.34 | |
| .htm | 75 | 192 | 57.01 | 134.99 | 70.31 | 44.64 |
| | 100 | 256 | 119.29 | 136.71 | 53.40 | |
| | 150 | 528 | 339.70 | 188.30 | 35.66 | |
| | 200 | 800 | 646.57 | 153.43 | 19.18 | |
| .max | 75 | 144 | 57.01 | 86.99 | 60.41 | 19.58 |
| | 100 | 192 | 119.29 | 72.71 | 37.87 | |
| | 150 | 336 | 339.70 | -3.70 | -1.10 | |
| | 200 | 544 | 646.57 | -102.57 | -18.86 | |
| .pdf | 75 | 96 | 57.01 | 38.99 | 40.61 | 0.12 |
| | 100 | 144 | 119.29 | 24.71 | 17.16 | |
| | 150 | 288 | 339.70 | -51.70 | -17.95 | |
| | 200 | 464 | 646.57 | -182.57 | -39.35 | |
| .png | 75 | 512 | 57.01 | 454.99 | 88.86 | 80.85 |
| | 100 | 640 | 119.29 | 520.71 | 81.36 | |
| | 150 | 1520 | 339.70 | 1180.30 | 77.65 | |
| | 200 | 2640 | 646.57 | 1993.43 | 75.51 | |
| tif class f | 75 | 32 | 57.01 | -25.01 | -78.16 | -338.81 |
| | 100 | 32 | 119.29 | -87.29 | -272.78 | |
| | 150 | 64 | 339.70 | -275.70 | -430.78 | |
| | 200 | 96 | 646.57 | -550.57 | -573.51 | |

(continued)

Table 6.9 (continued)

| Image format | dpi | Original image size (in KB) | Array size (in KB) | Memory saved using SIPA | Percentage of memory saved using SIPA (%) | Average percentage of memory saved using SIPA (%) |
|-----------------------------|-----|-----------------------------|--------------------|-------------------------|---|---|
| tif group 4 | 75 | 32 | 57.01 | −25.01 | −78.16 | −338.81 |
| | 100 | 32 | 119.29 | −87.29 | −272.78 | |
| | 150 | 64 | 339.70 | −275.70 | −430.78 | |
| | 200 | 96 | 646.57 | −550.57 | −573.51 | |
| tif lzw | 75 | 528 | 57.01 | 470.99 | 89.20 | 82.04 |
| | 100 | 704 | 119.29 | 584.71 | 83.06 | |
| | 150 | 1616 | 339.70 | 1276.30 | 78.98 | |
| | 200 | 2800 | 646.57 | 2153.43 | 76.91 | |
| tif _{uncompressed} | 75 | 1600 | 57.01 | 1542.99 | 96.44 | 95.36 |
| | 100 | 2880 | 119.29 | 2760.71 | 95.86 | |
| | 150 | 6480 | 339.70 | 6140.30 | 94.76 | |
| | 200 | 11536 | 646.57 | 10889.43 | 94.40 | |
| dcx | 75 | 1312 | 57.01 | 1254.99 | 95.65 | 92.41 |
| | 100 | 1504 | 119.29 | 1384.71 | 92.07 | |
| | 150 | 3872 | 339.70 | 3532.30 | 91.23 | |
| | 200 | 6944 | 646.57 | 6297.43 | 90.69 | |
| pcx | 75 | 1312 | 57.01 | 1254.99 | 95.65 | 92.41 |
| | 100 | 1504 | 119.29 | 1384.71 | 92.07 | |
| | 150 | 3872 | 339.70 | 3532.30 | 91.23 | |
| | 200 | 6944 | 646.57 | 6297.43 | 90.69 | |

Applying SIPA method on different image formats with different dpi proved to be efficient for most formats and yielded a considerable memory savings (if we ignored the execution time of the storage process)⁵ which is an important issue to us.

Table 6.9 shows the amount and percentage of memory saved by using SIPA model on different Image formats of the Whole KSU A4 Document scanned Images.

In Table 6.9 whenever the value of the memory saved is negative it implies that applying SIPA method is inefficient and whenever it is positive applying SIPA method is efficient. Although the memory saving percentage turned out to be negative for certain dpi the average saving could be positive which means applying SIPA method can be used for those other dpi where there is a considerable memory saving.

⁵ Even though time of execution is calculated and shown in paper tables, we do not focus much on it since time is affected by other factors like processor speed, RAM size, network reliability, etc.

Table 6.10 The amount and percentage of memory saved by using SIPA method

| Image format | dpi | Original image size (in KB) | Array size (in KB) | Memory saved using SIPA | Percentage of memory saved using SIPA (%) | Average percentage of memory saved using SIPA (%) |
|--------------|-----|-----------------------------------|-----------------------|----------------------------|---|--|
| jpeg | 75 | 60 | 53.96 | 6.04 | 10.07 | -46.33 |
| | 100 | 84 | 111.79 | -27.79 | -33.08 | |
| | 150 | 180 | 303.24 | -123.24 | -68.47 | |
| | 200 | 288 | 558.22 | -270.22 | -93.83 | |
| bmp | 75 | 992 | 51.77 | 940.23 | 94.78 | 93.18 |
| | 100 | 1744 | 109.72 | 1634.28 | 93.71 | |
| | 150 | 3904 | 301.63 | 3602.37 | 92.27 | |
| | 200 | 6944 | 557.76 | 6386.24 | 91.97 | |
| gif | 75 | 96 | 61.38 | 34.62 | 36.06 | 11.97 |
| | 100 | 144 | 124.80 | 19.20 | 13.33 | |
| | 150 | 304 | 318.13 | -14.13 | -4.65 | |
| | 200 | 592 | 573.37 | 18.63 | 3.15 | |
| tif | 75 | 304 | 51.77 | 252.23 | 82.97 | 77.82 |
| | 100 | 496 | 109.72 | 386.28 | 77.88 | |
| | 150 | 992 | 301.63 | 690.37 | 69.59 | |
| | 200 | 2912 | 557.76 | 2354.24 | 80.85 | |
| .exe | 75 | 464 | 26.43 | 437.57 | 94.30 | 81.89 |
| | 100 | 512 | 55.38 | 456.62 | 89.18 | |
| | 150 | 640 | 151.22 | 488.78 | 76.37 | |
| | 200 | 864 | 278.99 | 585.01 | 67.71 | |
| .fpx | 75 | 120 | 26.43 | 93.57 | 77.97 | 57.20 |
| | 100 | 156 | 55.38 | 100.62 | 64.50 | |
| | 150 | 288 | 151.22 | 136.78 | 47.49 | |
| | 200 | 456 | 278.99 | 177.01 | 38.82 | |
| .htm | 75 | 160 | 26.43 | 133.57 | 83.48 | 69.62 |
| | 100 | 208 | 55.38 | 152.62 | 73.38 | |
| | 150 | 400 | 151.22 | 248.78 | 62.20 | |
| | 200 | 688 | 278.99 | 409.01 | 59.45 | |
| .max | 75 | 112 | 26.43 | 85.57 | 76.40 | 54.13 |
| | 100 | 160 | 55.38 | 104.62 | 65.39 | |
| | 150 | 240 | 151.22 | 88.78 | 36.99 | |
| | 200 | 448 | 278.99 | 169.01 | 37.72 | |
| .pdf | 75 | 72 | 26.43 | 45.57 | 63.29 | 40.46 |
| | 100 | 108 | 55.38 | 52.62 | 48.73 | |
| | 150 | 216 | 151.22 | 64.78 | 29.9 | |
| | 200 | 348 | 278.99 | 69.01 | 19.83 | |
| .png | 75 | 336 | 26.43 | 309.57 | 92.13 | 89.02 |
| | 100 | 512 | 55.38 | 456.62 | 89.18 | |
| | 150 | 976 | 151.22 | 824.78 | 84.51 | |
| | 200 | 2864 | 278.99 | 2585.01 | 90.26 | |
| tif class f | 75 | 16 | 26.43 | -10.43 | -65.21 | -135.98 |
| | 100 | 32 | 55.38 | -23.38 | -73.05 | |
| | 150 | 48 | 151.22 | -103.22 | -215.04 | |
| | 200 | 96 | 278.99 | -182.99 | -190.62 | |

(continued)

Table 6.10 (continued)

| Image format | dpi | Original image size (in KB) | Array size (in KB) | Memory saved using SIPA | Percentage of memory saved using SIPA (%) | Average percentage of memory saved using SIPA (%) |
|---------------------|-----|-----------------------------|--------------------|-------------------------|---|---|
| tif group 4 | 75 | 16 | 26.43 | −10.43 | −65.21 | −125.60 |
| | 100 | 32 | 55.38 | −23.38 | −73.05 | |
| | 150 | 48 | 151.22 | −103.22 | −215.04 | |
| | 200 | 112 | 278.99 | −166.99 | −149.10 | |
| Tif lzw | 75 | 304 | 26.43 | 277.57 | 91.30 | 88.83 |
| | 100 | 496 | 55.38 | 440.62 | 88.84 | |
| | 150 | 992 | 151.22 | 840.78 | 84.76 | |
| | 200 | 2912 | 278.99 | 2633.01 | 90.42 | |
| tif uncompressed | 75 | 992 | 26.43 | 965.57 | 97.34 | 96.57 |
| | 100 | 1744 | 55.38 | 1688.62 | 96.82 | |
| | 150 | 3920 | 151.22 | 3768.78 | 96.14 | |
| | 200 | 6960 | 278.99 | 6681.01 | 95.99 | |
| dcx | 75 | 752 | 26.43 | 725.57 | 96.48 | 95.47 |
| | 100 | 1168 | 55.38 | 1112.62 | 95.26 | |
| | 150 | 2304 | 151.22 | 2152.78 | 93.44 | |
| | 200 | 8448 | 278.99 | 8169.01 | 96.70 | |
| pcx | 75 | 752 | 26.43 | 725.57 | 96.48 | 95.47 |
| | 100 | 1168 | 55.38 | 1112.62 | 95.26 | |
| | 150 | 2304 | 151.22 | 2152.78 | 93.44 | |
| | 200 | 8448 | 278.99 | 8169.01 | 96.70 | |

In general we can see that with most image formats applying SIPA method turned out to be highly efficient with the Whole KSU Document page as well as with the Full Body Segment as shown in Table 6.10.

Table 6.10 shows the amount and percentage of memory saved by using SIPA method on different Image formats of the Full Body Segment scanned Images.

6.9 Conclusion

Our goal in this research paper was to introduce Information Pixels and the concept of Storing Information Pixels Addresses (SIPA) and show practically that it is an efficient model for document storage for most image formats.

We observed that the stored images for most image formats are better in size if they are compared to their original sizes, which yielded in a considerable amount of memory savings.

- The time taken to store an image in a specific format with lesser dpi is less than the time taken to store an image in the same format with larger dpi. Hence, this is a time/complexity-quality tradeoff.

- Applying SIPA method on Full Body Segment images of a specific format and dpi take lesser time to be executed than its A4 size image counterpart.

Considering the above two factors, the complexity and time of image storage increases as we move from 75 to 200 dpi. It takes lesser time to store a Full Body Segment scanned image than the Whole A4 Document scanned image.

We can therefore conclude that using 150 dpi Full Body Segment scanned images produces an optimum result because:

- Storage of images with 150 dpi do not take as much time as images with 200 dpi;
- Storage of an image with 150 dpi is not as complex as an image with 200 dpi because the number of information pixels are lesser in images with 150 dpi and hence lesser the size of the array of addresses.
- The quality of images is certainly better than images with 75, 100.

References

1. ALGhalayini M, Shah A (2006) Introducing the (POSSDI) process: the process of optimizing the selection of the scanned document images. In: International joint conferences on computer, information, systems sciences, and engineering (CISSE 06) Dec 4–14, 2006
2. ALGhalayini MA, ALNemah E (2007) An efficient storage and retrieval technique for documents using symantec document segmentation (SDS) Approach. In: The third international joint conferences on computer, information, and systems sciences, and engineering (CISSE 07)
3. Bunke H, Wang PSP (1997) Handbook of character recognition and document image analysis. Available via <http://www.worldscibooks.com/comp/2757.html>
4. Document image analysis. Available via <http://elib.cs.berkeley.edu/dia.html>
5. Basic images processing. Some general commands related to handling Matlab graphics and printing “Simple image processing operations that you can do with Matlab”. Available via http://noodle.med.yale.edu/~papad/ta/handouts/matlab_image.html
6. Getting started with MATLAB. Available via <http://www.stewart.cs.sdsu.edu/cs205/module7/getting6.html>
7. Horiuchi T (2006) Grayscale image segmentation using color space. IEICE Trans Inf Syst E89-D(3):1231–1237
8. Matlab image processing toolbox. Available via <http://homepages.inf.ed.ac.uk/rbf/HIPR2/impmatl.htm>
9. Matlab resources. Available via <http://www.cse.uiuc.edu/heath/sci/comp/matlab.htm>
10. Ozden M, Polat E (2007) A color image segmentation approach for content-based image retrieval. Pattern Recognit 40(4):1318–1325
11. Richard Casey Document Image Analysis. Available via <http://cslu.cse.ogi.edu/HLTSurvey/ch2node4.html>
12. Simone M Document image analysis and recognition. Available via <http://www.dsi.unifi.it/~simone/DIAR/>
13. Simone M Document image analysis and recognition. Available via <http://www.dsi.unifi.it/~simone/DIAR/>

14. Read and display an image. Available via <http://www.mathworks.com/access/helpdesk/help/toolbox/images/getting8.html>
15. Matlab Tutorial (2005) Reading and displaying images. Available via <http://ai.ucsd.edu/Tutorial/matlab.html#images>

Chapter 7

Introducing the Concept of Back-Inking as an Efficient Model for Document Retrieval (Image Reconstruction)

Mohammad A. ALGhalayini

Abstract Today, many institutions and organizations are facing serious problem due to the tremendously increasing size of documents, and this problem is further triggering the storage and retrieval problems due to the continuously growing space and efficiency requirements. This problem is becoming more complex with time and the increase in the size and number of documents in an organization. Therefore, there is a growing demand to address this problem. This demand and challenge can be met by developing a technique to enable specialized document imaging people to use when there is a need for storing and retrieving documents images. Various techniques were developed and reported in the literature by different investigators. These techniques attempt to solve this problem to some extent but, most of the existing techniques still face the efficiency problem, in the case, the number and size of documents increase rapidly. The efficiency is further affected in the existing techniques when documents in a system are reorganized and then stored again. To handle these problems, we need special and efficient retrieval techniques for this type of information retrieval (IR) systems. In this paper, we present an efficient retrieval technique for electronic documents. The proposed technique uses the Back-Inking concept to make the technique more efficient for certain image formats. The use of this approach is a continuation of the SIPA approach which was presented in an earlier paper as an efficient method for

The method of applying SIPA (Storing Information Pixels Addresses) is introduced in: Mohammad A. ALGhalayini “*Introducing The Concept Of Information Pixels and the SIPA (Storing Information Pixels Addresses) Method As an Efficient Model for Document Storage*”; CISSE 2011, The Seventh International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 11).

M. A. ALGhalayini (✉)

Developmental Programs Supervisor Assistant Vice Rectorate for Graduate Studies and Scientific Research, King Saud University, Riyadh, Saudi Arabia
e-mail: malghalayini@gmail.com

document storage of the documents and as a result makes the image retrieval relatively efficient.

7.1 Introduction

By using SIPA technique, we learned how we can store the addresses of information pixels and how it helps in minimizing the storage size of the image components. We also analyzed the time consumed to convert an image from its respective format into an array of addresses [1–6].

It is important to understand the practical significance of this analysis. By converting an image into our format, we are saving the image not as an array with pixel intensity values but as an array with the addresses of information pixels. The time taken to convert the image into such an array is in reality the time taken to save the image in our format. Analogously, if retrieving the image saved in our format is the purpose, in reality we will be converting the array of addresses of information pixels into an image. The time taken to open the image saved in our format must be equivalent to converting the array into an image. In this research paper, we retrieve the image from the array of addresses by applying a technique that we shall call “**Back-Inking**” [7–15].

7.2 Back-Inking Algorithm

As the name implies, in this technique, we work backwards and replace an information pixel to where the address in the array points to. The addresses of information pixels are stored in the array. Every row in the array is made up of two elements. The first element is a pointer that points to the row and the second element is a pointer that points to a column. Together, they work as a 2-dimensional pointer pointing to the original location of the information pixel in an image [16, 17].

The algorithm places a black pixel at the address pointed by the array and the remaining pixels are white.

The Back-Inking Algorithm does the following:

- An array of size $(m \times n)$ is formed with all values in it being 256. This represents an image with all pixels white. Let us call it image R. Here m and n are the numbers of rows and columns (the coordinates) in the original image respectively.
- The algorithm reads the first row in the array of addresses. It contains two elements.
- It points to the row in image R corresponding to the first element among the two values read.
- It then points to the column in image R corresponding to the second element among the two values read.

- This pixel formed by the row-column intersection of addresses is replaced with the value 0.
- The algorithm next reads the second row in the array of addresses and points to the pixel corresponding to that address in image R. This pixel value is replaced by 0.
- The process repeats till the algorithm reads all rows in the array of addresses and replaces the pixels in image R corresponding to those addresses by 0.

Example: Let us simulate this algorithm for a BMP image with 150 pixels per inch.

Step 1: Create an array of size (1327×1004) with each value in it being 256. This image is called R.

This represents an image with intensity 256 all through. Hence the image is white. This is an initialization process. We replace white pixels with black ones according to the address stored in the array.

Step 2: The first element in the array of addresses corresponding to this image is read.

We have seen in previous research paper¹ that the array of addresses that corresponds to a BMP image of 150 pixels per inch has a size of $(43,266 \times 2)$. Let us call this array Array1. The first row of the array is read, that is, Array1 (1,1) and Array1 (1,2).

Array1 (1,1) = 1; Array (1,2) = 241;

Step 3: The algorithm points to the first row in image R.

This is because it points to the row in image R corresponding to the first element read, that is, Array1 (1,1) = 1.

Step 4: It then points to the 241st column in image R.

This is because it points to the column in image R corresponding to the second element read, that is, Array1 (1,2) = 241.

Step 5: The pixel formed by the row-column intersection in step 3 and in step 4 is replaced by a 0.

Intensity 0 corresponds to black. We have now replaced a black pixel at location (1,241).

Step 6: The same process repeats for each row of Array1 and hence all addresses present in Array1 are replaced by a 0 in image R, that is, a black pixel in image R. We shall call this the “**Back-Inking Process**”.

7.3 MATLAB Simulation of Back-Inking Algorithm:

Previously, we stored the image in the form of an array of addresses using MATLAB. We will now simulate the conversion of array of addresses into an image in MATLAB.

¹ ALGhalayini [18].

A snippet of the code that is used for reconstruction of the image is shown below.

```
% Start counter to keep a tab of the time
tic;
% Reconstructed array
Rearray = uint8 (255 * ones (m, n));
for l = 1:lgth
    rearray (Array (l, 1), Array (l, 2)) = 0;
end
figure 2
imshow (rearray)
% End of time counter
toc;
```

- (1) Here *tic* and *toc* commands are used to keep a tab of the time. They return the time taken to convert the array of addresses into an image.
- (2) *Rearray = uint8 (255 * ones (m, n));*

We initialize the reconstructed image (called *rearray*) with all intensity values 255.

Here

M—Number of rows in the image;

N—Number of columns in the image;

ones (m, n) will create an array of size ($m \times n$) with all values in it being 1.

*255*ones (m, n)* will multiply each value in the array *ones (m, n)* by 255. Hence each value in the array is now initialized to 255.

uint 8 is a function used to convert numbers from any data type to the type-*uint 8*. This is because images are supposed to be in the form *uint 8* in MATLAB. *uint8* has a range (0–255) which matches the intensity range in 8-bit grayscale images.

- (3) *for l = 1 : lgth*

```
    rearray (Array (l, 1), Array(l,2)) = 0;
end
```

This part of the program is used to replace each pixel in the image *rearray* with an intensity value 0. The address of the pixel is pointed by the array of addresses (called *Array*).

Here *lgth* = number of rows in the array.

The *for* loop:

```
for l = 1 : lgth
    .....
end
```

executes the statement within the loop *#lgth* number of times and in each iteration, the value of *l* is incremented by 1. The start value of *l* is 1 and its stop value is *#lgth*.

The statement within the *for* loop is:

Fig. 7.1 Shows the reconstructed Image after applying the Back-Inking method the whole A4 document (.gif) format image scanned with (150) dpi



$Rearray (Array (l, 1), Array (l, 2)) = 0$

As explained in the example above, the initial value in the *for* loop is *Array (l, 1) = 1* and *Array (l, 2) = 241*.

Here, $rearray (Array (l, 1), Array (l, 2))$, i.e., $rearray (1, 241) = 0$;

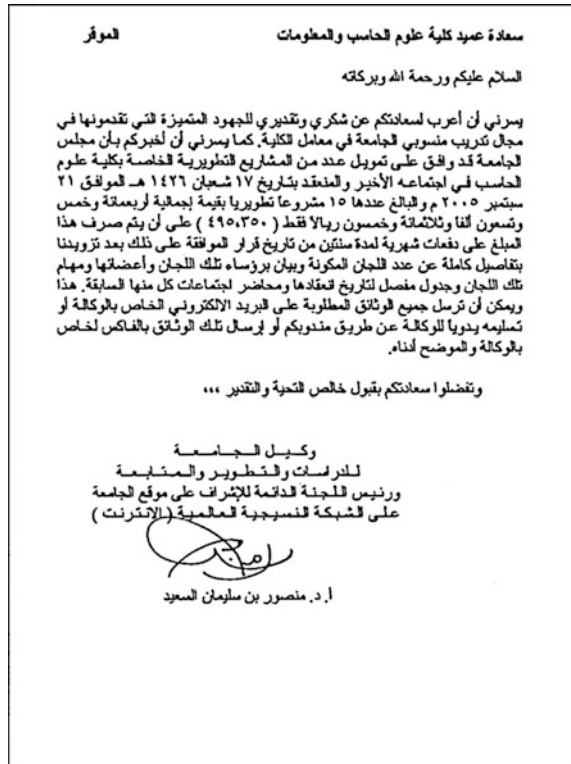
For each iteration in the *for* loop, successive values of addresses stored in *Array* are read and the corresponding values in *rearray* are replaced by 0. Since the *for* loop ends at $l = lgth$, the last iteration is executed till the last element in *Array* is read and a 0 is replaced in its corresponding location in the image.

(4) *Imshow (rearray)* is used to display the reconstructed image.

The same snippet of code can be used for all image formats with different dpi and for both full body segments of the scanned document image and the whole A4 scanned document image types.

For the same dpi, since the size of the full body image is smaller than that of an A4 size image; the number of information pixels in a full body image is lesser, the size of its corresponding array of addresses is smaller and hence it will take lesser time to be reconstructed into an image when compared to its A4 counterpart.

Fig. 7.2 Shows the reconstructed Image after applying the Back-Inking method the full body segment (.tif) format image scanned with (150) dpi



7.4 MATLAB Codes and Reconstructed Images

Now we are ready to see the MATLAB codes and reconstructed images for the examined image types for different image resolutions for both the whole KSU A4 document and for the Full Body Segment images.

The code below is the code for reconstructing (Back-Inking) the whole A4 Document (.gif) format image scanned with (150) dpi.

```
clear all; close all; clc;
tic;
RGBImage = imread ('C:\Users\ALGhalayini_rn\
a4size\gif\full page 256 colors 8bit 150.gif, 'gif');
GRAYImage = RGBImage;
figure 1;
imshow (GRAYImage);
m = length (GRAYImage);
n = length (GRAYImage (1, :));
total_pixels = m * n;
k = 1;
for i = 1 : m
```

Table 7.1 Shows the reconstruction time and the average viewable mark^a of the reconstructed images for different resolutions of the whole KSU document images

| No | Image format type | Image resolution (DPI) | Image reconstruction time (s) | Image evaluation points (10) |
|----|-------------------|------------------------|-------------------------------|------------------------------|
| 1 | GIF | 75 | 1.8223 | 6.2 |
| 2 | | 100 | 1.6245 | 6.9 |
| 3 | | 150 | 7.8764 | 8.9 |
| 4 | | 200 | 23.4180 | 9.7 |
| 5 | BMP | 75 | 0.8154 | 7.8 |
| 6 | | 100 | 1.7831 | 8.6 |
| 7 | | 150 | 8.4214 | 9.2 |
| 8 | | 200 | 24.2599 | 9.6 |
| 9 | JPG | 75 | 1.4210 | 5.9 |
| 10 | | 100 | 3.6006 | 7.1 |
| 11 | | 150 | 14.6970 | 8.6 |
| 12 | | 200 | 59.8907 | 9.1 |
| 13 | TIF | 75 | 0.7510 | 6.3 |
| 14 | | 100 | 1.5989 | 7.9 |
| 15 | | 150 | 7.7328 | 8.9 |
| 16 | | 200 | 19.2055 | 9.6 |

^a The reconstructed images were reviewed and evaluated by 10 different viewers for a mark out of 10

```

for j = 1 : n
    if GRAYimage (i, j) < 128
        Array1 (1, k) = i; Array2 (1, k) = j;
        k = k + 1;
    end
end
end
Array = [Array1' Array2'];
Toc

```

The code below is the code for reconstructing (Back-Inking) the Full Body (.tif) format image scanned with (150) dpi (Figs. 7.1 and 7.2; Tables 7.1 and 7.2).

```

clear all; close all; clc;
tic;
RGBimage = imread ('C:\Users\ALGhalayini_rn\
Fulltif\full page 256 colors 8bit 150.tif', 'tif');
GRAYimage = rgb2gray (RGBimage);
figure 1;
imshow (GRAYimage (454 : 1459, 118 : 1122));
m = length (GRAYimage);
n = length (GRAYimage (1, :));
total_pixels = m * n;
k = 1;

```

Table 7.2 Shows the reconstruction time and the average viewable mark of the reconstructed images for different resolutions of the full body segment images

| No | Image format (type) | Image resolution (DPI) | Image reconstruction time (Seconds) | Image evaluation points(10) |
|----|------------------------|---------------------------|--|--------------------------------|
| 1 | GIF | 75 | 0.6415 | 5.1 |
| 2 | | 100 | 1.4724 | 6.8 |
| 3 | | 150 | 6.1677 | 9.0 |
| 4 | | 200 | 12.8715 | 9.7 |
| 5 | BMP | 75 | 0.7326 | 4.1 |
| 6 | | 100 | 1.4580 | 5.9 |
| 7 | | 150 | 6.1659 | 8.1 |
| 8 | | 200 | 12.6825 | 9.1 |
| 9 | JPG | 75 | 0.7860 | 6.4 |
| 10 | | 100 | 1.7775 | 7.1 |
| 11 | | 150 | 6.8940 | 8.9 |
| 12 | | 200 | 12.8280 | 9.8 |
| 13 | TIF | 75 | 0.6282 | 6.3 |
| 14 | | 100 | 1.4001 | 7.1 |
| 15 | | 150 | 6.1385 | 8.5 |
| 16 | | 200 | 12.9299 | 9.2 |

```

for i = 454 : 1459
for j = 118 : 1122
if GRAYimage (i, j) < 128
Array1 (I, k) = i; Array2 (I, k) = j;
k = k + 1;
end
end
end
Array = [Array1' Array2'];
Toc

```

7.5 Conclusion

Our goal in this research paper was to introduce the concept of **Back-Inking** which is the procedure to retrieve the stored addresses of the Information pixels from the array and plot black pixels back into the white page to reconstruct the whole image back, then display it to the user in a relatively short time.

We observe that the reconstructed images for all formats are better in quality, in that, as the dpi increases, the sharpness of the image increases. But as the dpi increases the time to reconstruct the image also increases.

Through applying this method, we came up with the following conclusions:

- The time taken to retrieve an image in a specific format with lesser dpi is less than the time taken to retrieve an image in the same format with larger dpi. Hence, this is a time/complexity-quality tradeoff.
- Full body images of a specific format and dpi take lesser time to be executed than its A4 size image counterpart.

Considering the above 2 factors, the complexity and time of image retrieval increases as we move from 75 to 200 dpi. It takes lesser time to retrieve a full body segment scanned image than an A4 scanned image.

We can therefore conclude that using 150 dpi full body segment scanned images produces an optimum result because:

- Retrieval of images with 150 dpi do not take as much time as images with 200 dpi;
- Retrieval of an image with 150 dpi is not as complex as an image with 200 dpi because the number of information pixels are lesser in images with 150 dpi and hence lesser the size of the array of addresses.
- The quality of images is certainly better than images with 75,100

References

1. Horiuchi T (2006) Grayscale image segmentation using color space. IEICE Trans Inform Sys E89-D(3):1231–1237
2. Panda S, Nanda P, Mohapatra P (2007) Multiresolution approach for color image segmentation using MRF model. In: Proceedings of the national conference on smart communication technologies and industrial informatics, SCTII, 03–04, Rourkela, Feb 2007, pp 34–42
3. Kang S, Park S, Shin Y, Yoo H, Jang D (2008) Image segmentation using statistical approach via perception-based color information. Int J Comp Sci Netw Secur 8(4):4
4. Bunke H, Wang PSP Handbook of character recognition and document image analysis. Available at <http://www.worldscibooks.com/compsci/2757.html>
5. Simone M Document image analysis and recognition. Available at :<http://www.dsi.unifi.it/~simone/DIAR/>
6. Document image analysis. Available at <http://elib.cs.berkeley.edu/dia.html>
7. Casey R Document image analysis. Available at <http://cslu.cse.ogi.edu/HLTsurvey/ch2node4.html>
8. Basic images processing: some general commands related to handling Matlab graphics and printing. Simple image processing operations that you can do with Matlab. Available at: http://noodle.med.yale.edu/~papad/ta/handouts/matlab_image.html
9. Document image analysis. Available at: <http://elib.cs.berkeley.edu/dia.html>
10. Getting started with MATLAB. Available at: <http://www.stewart.cs.sdsu.edu/cs205/module7/getting6.html>
11. Matlab image processing toolbox. Available at: <http://homepages.inf.ed.ac.uk/rbf/HIPR2/impmatl.htm>
12. Matlab resources. Available at: <http://www.cse.uiuc.edu/heath/scicomp/matlab.htm>

13. Read and display an image. Available at: <http://www.mathworks.com/access/helpdesk/help/toolbox/images/getting8.html>
14. Reading and displaying images. Matlab tutorial, 03/25/2005. Available at: <http://ai.ucsd.edu/Tutorial/matlab.html#images>
15. ALGhalayini M, Shah A (2006) Introducing The (POSSDI) process: the process of optimizing the selection of the scanned document images. In: International joint conferences on computer, information, systems sciences, and engineering (CISSE 06) 4–14 Dec 2006
16. ALGhalayini MA, ALNemah E (2007) An efficient storage and retrieval technique for documents using symantec document segmentation (SDS) approach. CISSE 2007, In: The third international joint conferences on computer, information, and systems sciences, and engineering (CISSE 07)
17. ALGhalayini MA (2011) Introducing the concept of information pixels and the SIPA (storing information pixels addresses) method as an efficient model for document storage, CISSE 2011. In: The seventh international joint conferences on computer, information, and systems sciences, and engineering (CISSE 11)

Chapter 8

Automating the Transformation From a Prototype to a Method of Assembly

Yuval Cohen, Gonen Singer, Maya Golan and Dina Goren-Bar

Abstract This paper describes a new technique that utilizes the typical documentation of complex products to automate the development of the assembly method to be used for production. The technique describes a structured process that gets (as its input) the standard bill of materials (BOM) with specified additional data, and develops a detailed sequential method of assembly operation as its output. This sequential assembly method could be then further automated. The paper also discusses the gap between typical assembly instructions and structured sequential specifications necessary for automating the planning of the assembly method.

8.1 Introduction

This paper deals with an environment of complex assembly of small to medium lots or batches. It strives to bridge the gap between the standard documentation generated at the product development stage, and the sequence of operations used for executing the assembly. The standard documentation includes the bill of materials (BOM) and the assembly instructions. After reviewing hundreds of assembly instructions the authors realized that they usually leave large flexibility to the assembly worker as to their execution. A typical documentation includes series of illustrated assembly instruction cards called route cards (see Fig. 8.1) [1, 2]. One or

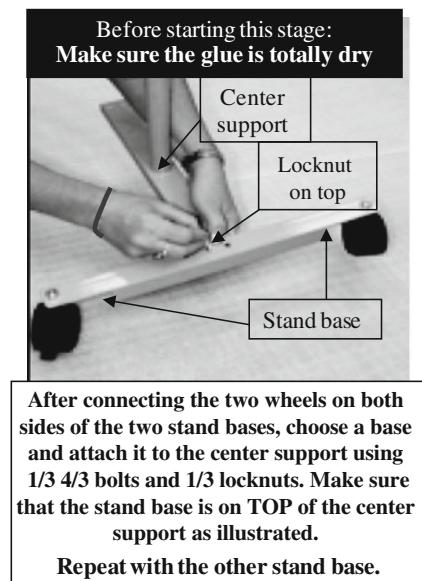
Y. Cohen (✉)

Department of Industrial Engineering, The Open University of Israel, Raanana, Israel
e-mail: yuvalco@openu.ac.il

G. Singer · M. Golan · D. Goren-Bar

Department of Industrial Engineering,
Tel-Aviv Afeka College of Engineering, Tel-Aviv, Israel

Fig. 8.1 Typical assembly route card



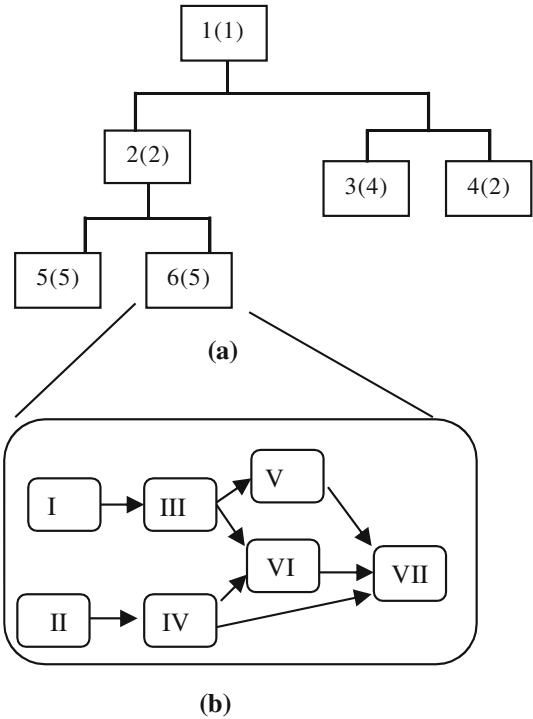
more route cards are generated for each node of the BOM tree [3]. These cards usually include the required tools, materials and fastener items (such as the types of nuts and bolts).

Traditionally, there are two very different ways to describe the assembly structure: (1) BOM tree [4, 5] (often called product structure tree) and (2) precedence diagram [6, 7]. Typical examples of these two models are illustrated in Fig. 8.2. In Fig. 8.2a the numbers are sub-assembly numbers and the numbers in brackets are quantities. In Fig. 8.2b the numbers denote tasks.

Cohen and Goren-Bar [3] had examined hundreds of instances and found consistent matching between Route cards and the nodes of the BOM tree. The rule found was that a route card can always be associated with a BOM tree node, while the node could be associated with one to several route cards. On the other hand, the activities described within each route card could be easily arranged as a precedence diagram. The upper level describes the product structure, and the detailed level describes the activities using a precedence diagram. Thus, our model is a development on top of this hierarchy of two levels and requires a straight sequence. For example, it requires the tree-nodes to be re-arranged in a line. For example, re-arranging the tree of Fig. 8.2a in a line may give a sequence of: 6-5-4-3-2-1, or 5-6-2-4-3-1, or 4-3-5-6-2-1, etc.

The rest of the paper is built as follows: Sect. 8.2 describes the relational tables required to generate the method prototype, Sect. 8.3 describes the BOM extension for assembly sequences, Sect. 8.4 discusses the method prototype, Sect. 8.5 describes how time standards could be generated from the assembly method prototype, and Sect. 8.6 concludes the paper.

Fig. 8.2 **a** An example of product structure tree (BOM tree) **b** An example of a precedence diagram



8.2 The Relational Tables and Sequencing

While assembly processes are done in a sequence the information in Fig. 8.2 does not specify a sequence. It also does not provide the tooling required for each assembly operation, and does not specify in detail the method used for the assembly. In short, there is certain additional information that could benefit the user if it would be added to the BOM.

Tables usually help in organizing information and minimizing its repetition. For minimizing repetitions tables in data bases. Therefore we identify the following necessary tables that would assist in extending the BOM tree:

Tables 8.1, 8.2, 8.3, 8.4, 8.5, 8.6 do not require the sequence of operations for their construction. Therefore Tables 8.1 through 8.6 could be constructed in the given order, but there are many other orders that may work as well. Table 8.7 on the other hand, requires not only the sequence, but also the information from the other six tables. So it will always be the last table to be built. Table 8.7 is already a sequence of assembly operations: so that each stage contains only two assembled parts: the initial sub-assembly and the adjoined part or subassembly. Effective sequencing could be done in one of several ways mentioned briefly in Sect. 8.4. However, we may point out that part of the rough-cut sequence could be found by

Table 8.1 Tool type table

| Tool type number | Tool name | Setup time | Volume | Weight | Operation type |
|------------------|-----------|------------|--------|--------|----------------|
|------------------|-----------|------------|--------|--------|----------------|

simply adopting the sequence of route cards or the general steps found in the assembly instructions.

The Tables 8.1, 8.2, 8.3, 8.4, 8.5, 8.6, 8.7 contain the information required to generate standard time estimates for the assembly process. For example, Table 8.1 lists the set-up time, and the volume and weight that enable assessment of tool handling time. Table 8.2 adds the information regarding the tool location for assessing the time it takes to approach the tool.

8.3 The Bill of Materials (BOM) Extension

This sub-section defines a new annotation scheme for an extended BOM. First we introduce the brackets form to describe the assembly operation. Brackets will be used to define the assembly of two parts. For example, (1, 2) describes the assembly of part 1 and part 2. Now, suppose this sub-assembly of parts 1 and 2 is assembled with part 3. This would be written as [3, (1, 2)]. Since one part is joined at each step, for n parts there are always n-1 brackets. This is true regardless of the assembly order. In addition to the order, the type of operation and the type of tool are of interest for each assembly operation. Therefore we suggest adding operation number and tool number for each bracket. The suggested concept is depicted in Fig. 8.3.

In Fig. 8.3 sub-assembly 8 starts its construction by joining parts 5 and 7 using operation 12 and tool 5. The assembly continues by joining part 6 to the subassembly using operation 8 and tool 3. There is a one to one correspondence between each pair of brackets in Fig. 8.3 and the rows of Table 8.6 (assembly operations). However, the scheme of brackets order also reflects the sequence of assembly.

8.4 The Method Prototype

The same product could be assembled in many different sequences of operations. However, determining a single sequence is necessary for organizing and standardizing the assembly process, for training the workers, and for adjusting the facility to the assembly process. Even assembly instructions are too general and could be carried out in many different ways. This becomes more apparent once we realize that assembly instructions are just a sequence of precedence diagrams of the type portrayed in Fig. 8.2b.

A sequence of mini tasks that appear in Fig. 8.2b generates a “method prototype” and is necessary for further automation, standardization and generating

Table 8.2 Tools table

| Tool number | Tool type number | Location |
|-------------|------------------|----------|
| | | |

Table 8.3 Fastening parts table

| Fastener type # | Tool name | Setup time | Volume | Weight | Fasten. time |
|-----------------|-----------|------------|--------|--------|--------------|
| | | | | | |

Table 8.4 Part type table

| Part type | Volume | Weight | Fastener type 1 number | Fastener type 2 |
|-----------|--------|--------|------------------------|-----------------|
| | | | | |

Table 8.5 Single part table

| Part serial number | Part type number | Location |
|--------------------|------------------|----------|
| | | |

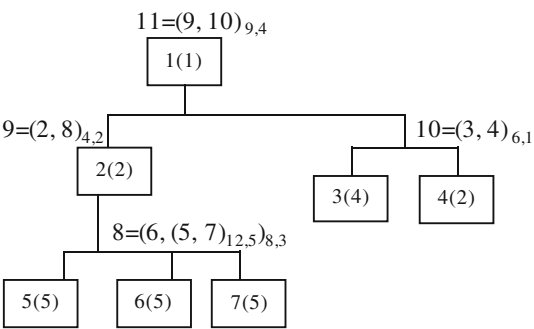
Table 8.6 Operations/tasks table

| Operation number | Operation type | Tool number | Fastening type # | Typical Duration |
|------------------|----------------|-------------|------------------|------------------|
| | | | | |

Table 8.7 Sub assemblies table

| Sub assembly number | Initial Sub assembly | Joining Part type # | Joining Sub assembly | Operation number |
|---------------------|----------------------|---------------------|----------------------|------------------|
| | | | | |

Fig. 8.3 Example of extended BOM tree



standard times. A simplified example for such sequence is Ranked Positional Wait (RPW) in which the tasks are sorted the sum of times of the tasks and all its successors [8–10].

For a complex product assembly, sequencing the operations of the assembly process is closely related to assembly line balancing [10, 11]. Several methods have been suggested for sequencing the assembly operations [11–13]. These methods or other methods of future research could be used for developing the method prototype. However, the “method prototype” must be more detailed specifying the tools, and fasteners used in addition to the parts assembled. The “method prototype” also includes the location of the different parts tools and fasteners and requires the knowledge of their attributes. Due to space limitations we shall not discuss sequences further here.

While optimal sequencing is not the subject of this paper, the next section presents how to generate time-standards from a given sequence of assembly operations.

8.5 Time Standards Generation

In order to generate time standards the sequence of operations must be translated to a sequence of required motions. For example, the sequence in Fig. 8.1 could be:

1. Get one “stand base” in one hand and hold it
2. Get the “center support” in the second hand and align it to the hole in the middle of the “stand base”.
3. Hold the “stand base” and “center support” in one hand
4. Get the bolt and push it through the hole of the two parts.
5. Hold the bolt with the two parts in one hand and get the locknut in the other
6. Align the locknut to the bolt
7. Turn the locknut clockwise three revolutions

Generating the sequence of motions is facilitated by Tables 8.1, 8.2, 8.3, 8.4, 8.5, 8.6, 8.7. The tables enable to find for each operation:

- The parts assembled and their attributes
- The fasteners used and their attributes
- The tools used and their attributes

Each of these motions could be described by a *Predetermined Time and Motion Study (PMTS)* method such as MTM or MOST [14]. These methods are typically used to estimate the standard times of operations before their execution.

For example,

‘Get one “stand base” in one hand and hold it’

Is translated to MOST (assuming available table features relating to the weight and volume of the “stand base”) as follows:

Get one “stand base” in one hand = $A_1B_0G_3$

And hold it – $A_0B_0P_1$

In this example there were no tool and no fastener, and the indexes could be inferred by the part attributes alone. However, as mentioned before, for more complex situations all of the tables of Sect. 8.2 are necessary.

8.6 Conclusion

This paper presents a new technique for modeling the assembly process in a way that facilitates its execution and automation. The technique could also be used for disassembly and maintenance operations. The technique relies on generating a “method prototype” having a strict sequence of operations based on typical documentation. The “method prototype” has a potential to be used also for simulation and validation. The translation of the model to a sequence of tasks and motions is the key for estimating the assembly period based on time standards.

Another advantage of the proposed model is its compatibility with the route card system and the ability to automate only parts of the process while certain parts may remain manual assembly. Future research includes implementation of the proposed technique on an industrial shop floor. This would enable simulation and visualization of operation; it would assist in real-time tracking and failure analysis.

References

1. Adams RJ, Klowden D, Hannaford B (2001) Virtual Training for a Manual Assembly Task. *Haptics-e* 2(2):1–7
2. Agrawala M, Doantam P, Heiser J, Haymaker J, Klingner J, Hanrahan P, Tversky B (2003) Designing effective step-by-step assembly instructions. *ACM Trans Graph (TOG)* 22(3): 828–837
3. Cohen Y, Goren-Bar D (2010) New automated assembly model based on automated route card scheme. In: Proceeding of the international conference on industrial electronics, technology and automation (IETA 10), In IEEE's: Proceedings of the international joint conferences on computer, information, and systems sciences, and engineering (CISSE)
4. Bukchin J, Masin M (2004) Multi-objective design of team oriented assembly systems. *Eur J Oper Res* 156:326–352
5. Tatsiopoulos IP (1996) On the unification of bills of materials and routings. *Comput.Ind* 31(3):293–301
6. Cohen Y, Dar-El E (2010) The sliding frame—extending the concept to various assembly line balancing problems. *IJMTM* 20(1/2/3/4):4–24
7. Cohen Y (2010) A new technique for solving the assembly line design problem. In: Proceedings of the 19th industrial engineering research conference (IERC), Cancun, Mexico
8. Scholl A, Becker C (2006) State-of-the-art exact and heuristic solution procedures for simple assembly line balancing. *Eur J Oper Res* 168:666–693
9. Kim YK, Song WS, Kim JH (2009) A mathematical model and a genetic algorithm for two-sided assembly line balancing. *Comput Oper Res* 36:853–865

10. Eryuruk SH, Kalaoglu F, Baskak M (2008) Assembly line balancing in a clothing company. *Fibres Textiles East Eur* 16(1):93–98
11. Cunha C, Agard B, Kusiak A (2005) Improving manufacturing quality by re-sequencing assembly operations: a data-mining approach. In: *Proceeding of the 18th international conference on production research (ICPR-18)*, Salerno
12. Scholl A, Boysen N, Fließner M (2008) The sequence-dependent assembly line balancing problem. *OR Spectrum* 30:579–609
13. HongGuang L, Cong L (2010) An assembly sequence planning approach with a discrete particle swarm optimization algorithm. *Int J Adv Manuf Technol* 50(5/8):761–770
14. Zandin K (2003) *MOST work measurement systems*, 3rd edn. Marcel Dekker, New York

Chapter 9

Collaborative and Non-Collaborative Dynamic Path Prediction Algorithm for Mobile Agents Collision Detection with Dynamic Obstacles in 3D Space

Elmir Babovic

Abstract In this research the extension of the algorithm for dynamic collaborative path prediction for mobile agents is proposed. This algorithm is inspired by human behavior in group of dynamical obstacles. Mobile agent in collaborative manner uses coordinates of other mobile agents in the same environment to calculate and based on statistical methods predict future path of other objects. For this purpose spatial-temporal variables are decomposed in order to optimize the method and to make it more efficient. This algorithm can be used in mobile robotics, automobile industry and aeronautics. Moreover this method allows full decentralization of collision detection which allows many advantages from minimizing of network traffic to simplifying of inclusion of additional agents in relevant space. Implementation of the algorithm will be low resource consuming allowing mobile agents to free resources for additional tasks.

9.1 Introduction

The inspiration for this work came from human and animals' cognitive activities executed while single unit is moving in group of other moving units. A group of mobile robots is a base for this research but resulting methodology is applicable for any group of mobile agents.

The objective of the research was to propose a functional algorithm for collision detection and avoidance based on collaborative path prediction of dynamic obstacles being other mobile agents from the perspective of one mobile agent.

E. Babovic (✉)

Faculty of Information Technologies, Mostar, Bosnia and Herzegovina
e-mail: babovic.elmir@bih.net.ba

In 3D space group of collaborating mobile agents or autonomous aerial vehicles would exchange their coordinates in order to allow other units to calculate, estimate and predict future paths of other units. The goal of such a scenario is to:

- Minimize network traffic.
- Avoid need for central hierarchical top-level controller.
- Increase autonomy of mobile agents.
- Simplify calculation and decrease of uncertainty in dynamic environment.
- Simplify inclusion of additional mobile agents into system without need to foreseen it, etc.

The algorithm is based on method recently developed which uses statistical methods for calculation of future mobile agent's paths. This is achieved by measuring coordinates every time period t_n which is collaboratively determined in system setup based on average speed and mobile agents physical and technical characteristics and agility. This algorithm used mobile agent's ability to log coordinates and, using statistical methods predict further path of other mobile agents. This way mobile agent is able to detect possible collision and to execute necessary maneuver to avoid collision.

Since one of the goals was to simplify dynamic spatial-temporal analysis of current situation in t_n , it is decided to analyze position and direction of mobile agent in 3D space decomposed. Spatial-temporal state is de-composed into three sub-states. Axis x coordinate with time t is analyzed separately from axis y and y position. Algorithm avoids complex calculation and spatial-temporal analysis. Instead of that kind of calculation, spatial-temporal decomposition is made and statistical tool is used to generate predictive analysis of current path. Since state is de-composed, algorithm re-composes parameters in order to predict full path. This prediction will be sufficient for full collision detection. This research is in line with increased trend of decentralized control of mobile robots [1].

9.2 Recent Developments and Research Background

During last ten years there has been large number of different theoretical-conceptual as well as practical developments in this area. Significant number of researches was done in the area of collision detection in static and dynamical environment. Reference [2] shows experiment with Cross-Coupling controller for mobile robots. Later Forsberg [3] analyzed Range-Weighted Hough transformations and extended Kalman Filter for demonstrating accurate and robust robotic navigation in closed spaces. General concept of autonomous robots navigation consists of three layer architecture [4]. This architecture consists of following elements: sensor systems, planning system and control system.

However cognition and higher level of intelligence were not analyzed. Some elements such as uncertainty and decision making based on incomplete information [5] are analyzed as independent items. However cognition and higher level of

intelligence were not analyzed. Some elements such as uncertainty and decision making based on incomplete information [5] are analyzed as independent items.

One of interesting work relevant for this research is work of Sebastian Thrun [6] who analyses uncertainty and recursive estimation of state Gaussian and Kalman. Relevant research [7] is analysis of collaborative dependency of robots and humans.

Reference [8] shows collaborative encounter of heterogeneous robots in unknown environment with unknown starting location. There are three possible levels of communication:

- No communication,
- Limited communication,
- Full communication.

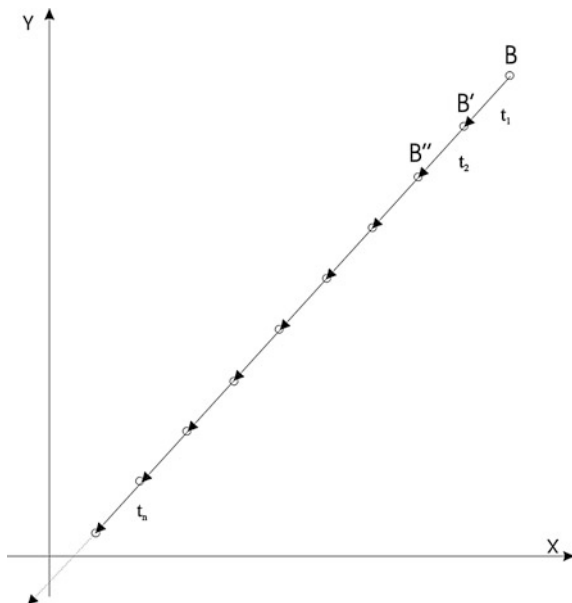
Reference [9] analyses uncertainty in motion planning for reliable robots in health institutions. Breazeal [10] goes one step further in this area and uses Bayesian approach for learning and decision making with level of uncertainty from the human-robot perspective for control of wheeled-chair. Monte Carlo estimation and error lowering optimization of that kind of learning is made by Roy [11]. Trust level using mathematic method is introduced in 2003 and it is based on POMDP [12]. Emma Brunskill [13] analyses reinforced learning as a method in mobile robotics. This research uses third generation of mobile robots [14] which uses cognitive elements of learning and conclusion necessary for collision detection. Collaborative elements are also used in static and dynamic methods for collision detection in 3D polygonal space [15]. Navigation of mobile robot toward predefined target is also done using incremental learning [16]. Significant element in this method is based on reactive space exploration and SOVEREIGN simulation based on neuron models. Collision prevention is often achieved using web cameras and wide angle cameras with Bluetooth and other wireless communication which significantly increases control algorithm and hardware configuration [17, 18].

In case of group of robots with identical technical characteristics, usually individual robots are considered as basic element for research [19]. One of the methods for positioning of static and dynamic obstacles is usage of mobile robots formations [20]. Robots communication and collaboration is also used for path planning based on sensors measurements and algorithms of robots collaboration [21]. Beside that there are many other multi-robots collaboration approaches in the multi-robots networks [22–24].

One of the solution for mentioned problematic is generic algorithm [25] as well as on/line reference generation and control schemes [26]. For the reasons of increased complexity of mobile robots a term *behavior* is used [27]. VSTR [28] (variable single-tracked robot) algorithm for collision avoiding is one of the solutions.

Therefore a new concept of statistical mathematical models and tools are used for mobile agents' collision detection. This algorithm is based on algorithm and concept presented in [29].

Fig. 9.1 Absolute position demonstration of mobile agent motion through 2D space in discrete time frames



9.3 Method

Referring to previous research on proposed method of collision detection based on collaborative path prediction this work is concentrated on proposing and algorithm for this prediction. In order to develop a proposal for algorithm a full concept of mentioned method will be explained as follows.

Considering 2D coordinate system XY , it is assumed that object B moves in equal time frames $t_1 = t_2 = \dots = t_n$ as shown on Fig. 9.1.

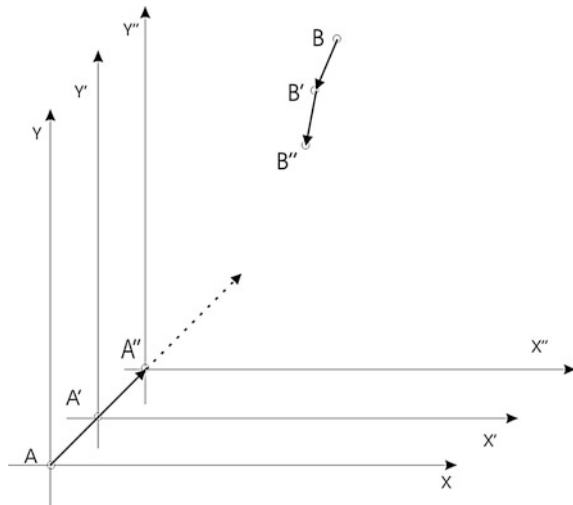
Standard way of considering this scenario is assuming absolute position of center of coordinate system and all mobile agents would be considered with same absolute coordinates.

If we assume object A moves in the same space as object B but only this time we consider object A as reference point. In this case center of the coordinate system moves with object A . This assumption is demonstrated on Fig. 9.2.

Figure 9.2 shows relative position of object B in moving coordinate system in which center is object A . This assumption results in usage of all four quadrants of new A coordinate system. According to assumption illustrated above, following example is based on 3D coordinate system.

In this case relative position of object B (x_B' , y_B' , z_B') in coordinate system of object A is calculated as:

Fig. 9.2 Motion of object A and B and relative positioning of object B from object A



$$\begin{aligned}
 xB' &= xB + \Delta xB + \Delta xA \\
 yB' &= yB + \Delta yB + \Delta yA \\
 zB' &= zB + \Delta zB + \Delta zA
 \end{aligned}
 \tag{9.1}$$

Several positions of object B are logged in order to calculate trend of object B motion. This can be assured by implementing protocol for exchange of absolute coordinates or by implementing long range sensors detection.

For this research case example coordinates x and y are transposed and analyzed against time value. This way there are two independent variables to be analyzed against fixed time variable.

Tracked and logged values are computed using statistical tool regression procedure. Applying regression procedure on case example values shown on Fig. 9.3.

In order to complete prediction it is necessary to analyze current state of objects motion and then to predict motion and path of the object. For regression analysis of current state of motion, simple linear regression is used. Since for this part of motion analysis it is not required to use advanced calculation simple linear regression is enough for describing current state of motion.

$$Y_r = \alpha + \beta X \tag{9.2}$$

In this case, since one coordinate is time and another X and Y separately, for each analysis respectively can be represented as:

$$\begin{aligned}
 (a) \ R_x &= \alpha_1 + \beta_1 t. \\
 (b) \ R_y &= \alpha_2 + \beta_2 t. \\
 (c) \ R_z &= \alpha_3 + \beta_3 t.
 \end{aligned}
 \tag{9.3}$$

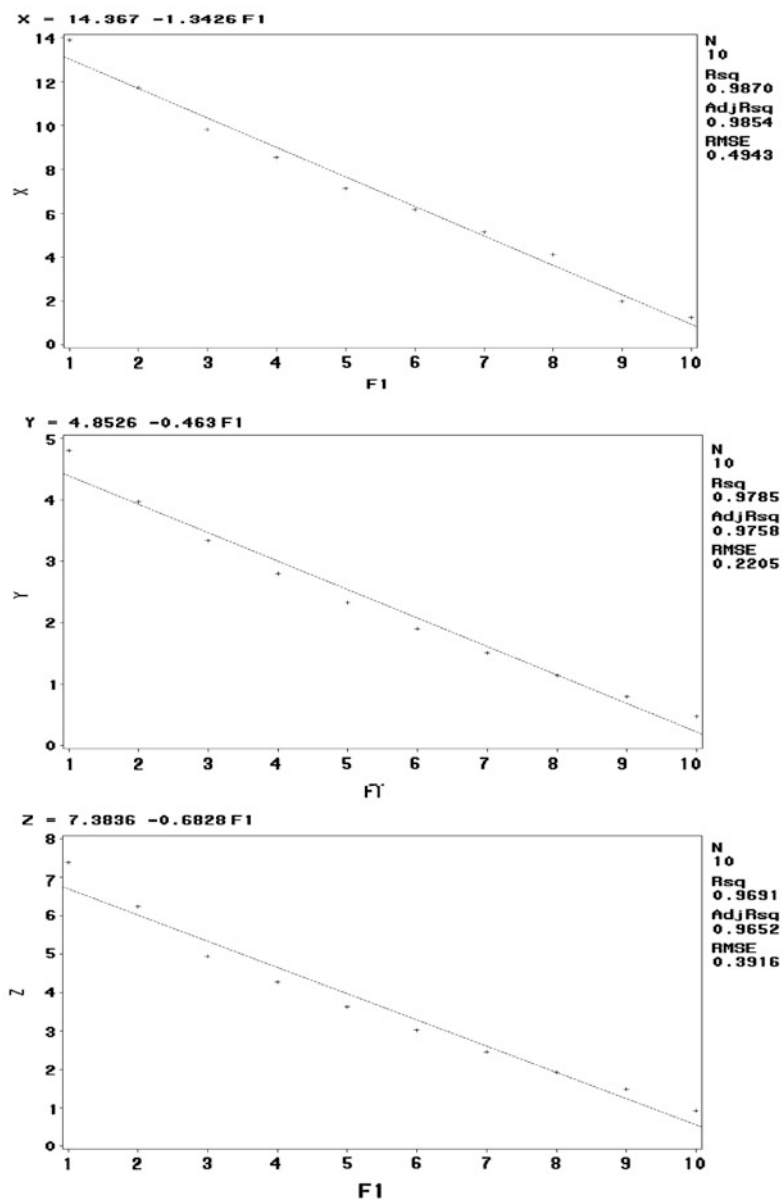


Fig. 9.3 Applied regression analysis on variables x, y and z based on independent variable t

where

$$\begin{aligned}\alpha &= \frac{\Sigma Y - \beta \Sigma X}{n} \\ \beta &= \frac{n \Sigma (XY) - \Sigma X \Sigma Y}{n \Sigma X^2 - (\Sigma X)^2}.\end{aligned}\quad (9.4)$$

Number of sampling $N = 10$. The usage of only ten samplings is also a demonstration that depending of the velocity of the mobile agents and its agility different amounts of sufficient number of samplings can be used. In this simulation relatively slow aerial vehicle with velocity of 0.02 m/s is used. For these range of velocity in real situation ten samples is sufficient amount for accurate and confident calculation of path. From regression analysis of ten samples in this example results are following.

Regression for coordinate X is $R_x = 14.38 - 1.34t$, for coordinate Y is $R_y = 4.85 - 0.46t$ and for Z is $R_z = 7.38 - 0.68t$. In this example X variable of object B in dynamic coordinate system of object A is moving linearly in equal time periods t while Y and Z variables are moving in curved shape.

For X variable calculated R-Square is 0.987 and Adjusted R-Square is 0.985. For Y variable calculated R-Square is 0.9785 and Adjusted R-Square is 0.9758. For Z variable calculated R-Square is 0.969 and Adjusted R-Square is 0.965.

Since variance of all X , Y and Z are explained over 96 % confidence in prediction of path can be considered true.

The subject of one of future works should be evaluation of confidence of calculated values which will be function based on variables such as average relative velocity of the objects, objects mass and agility, objects size etc. For this research it is considered that collision point predicted in $t_{n+m} = t_{\text{collision}}$ is $X_{\text{collision}} = Y_{\text{collision}} = Z_{\text{collision}} = 0$.

In order to continuously predict future path, forecast of values is applied as shown on Fig. 9.4.

The mobile agent has to predict and to avoid collision moment by altering own course. Calculation of path and path prediction of both coordinates is done dynamically through all the time of motion of mobile agent from the start position to the target position. As soon as mobile agent detects collision state $X = Y = Z = 0$ in relevant future it has to modify path. In this work it is called **Relevant predicted collision time** and in this simplified example it is t_{n+1} .

In order to calculate exact point of collision numerical values of prediction variables are used.

Table 9.1 contain measure and forecasted values for variables X and Y with 95 % upper and lower confidence limits. The problem of density of measurements of variables or in another term sampling frequency is solved by calculating the time needed for mobile agent to pass the path half of its own size with constant speed. So minimal sampling frequency is:

$$f = \frac{1}{\frac{l/2}{v}}. \quad (9.5)$$

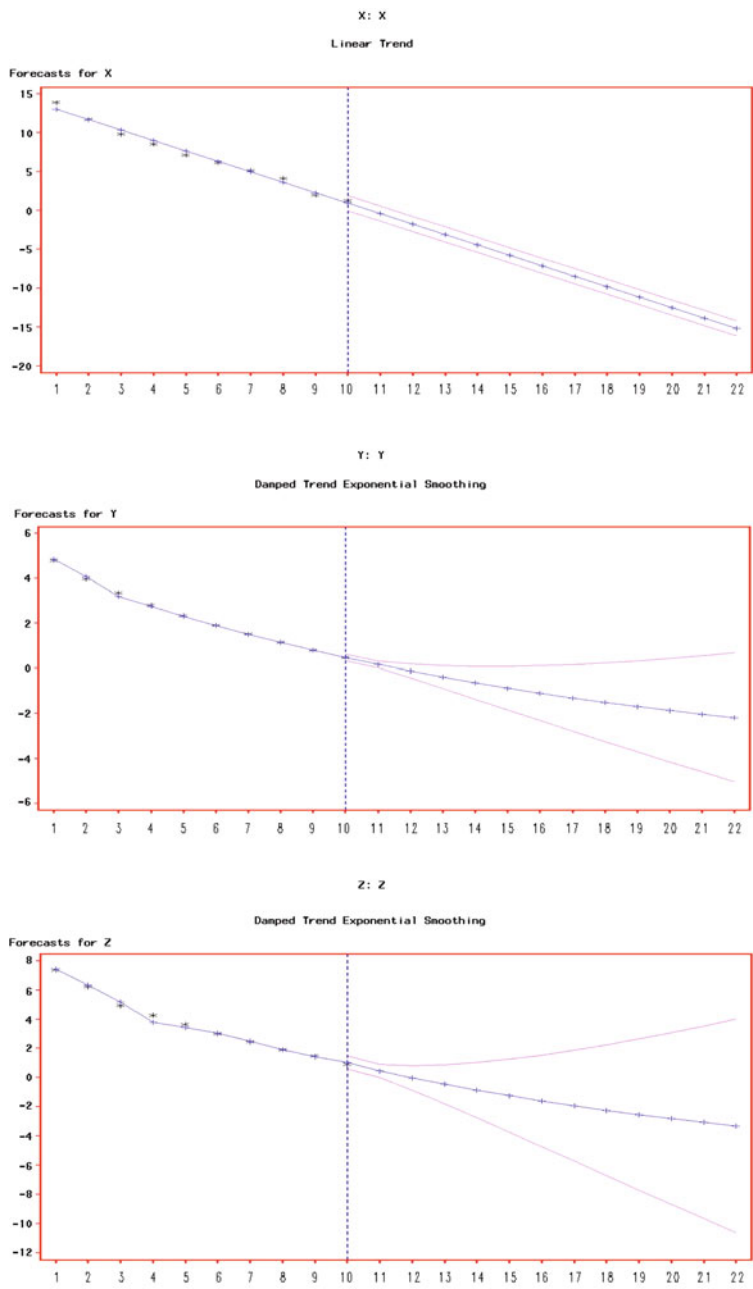


Fig. 9.4 Forecasted values of the variables x, y and z in future time periods based on updated values of the variables

Table 9.1 Values of X, Y and Z predicted variable with 95 % upper and lower confidence

| X | X U95 (%) | X L95 (%) | Y | Y U95 (%) | Y L95 (%) | Z | Z U95 (%) | Z L95 (%) |
|--------|-----------|-----------|-------|-----------|-----------|-------|-----------|-----------|
| 13.89 | 13.99 | 12.06 | 4.80 | 4.99 | 4.70 | 7.39 | 7.89 | 6.99 |
| 11.73 | 12.65 | 10.71 | 3.97 | 4.21 | 3.93 | 6.24 | 6.79 | 5.89 |
| 9.82 | 11.31 | 9.37 | 3.34 | 3.34 | 3.05 | 4.94 | 5.64 | 4.74 |
| 8.55 | 9.97 | 8.03 | 2.80 | 2.88 | 2.60 | 4.27 | 4.24 | 3.34 |
| 7.14 | 8.62 | 6.69 | 2.33 | 2.44 | 2.15 | 3.63 | 3.89 | 2.99 |
| 6.18 | 7.28 | 5.34 | 1.90 | 2.03 | 1.74 | 3.02 | 3.50 | 2.60 |
| 5.15 | 5.94 | 4.00 | 1.51 | 1.64 | 1.36 | 2.46 | 2.94 | 2.04 |
| 4.12 | 4.60 | 2.66 | 1.14 | 1.28 | 1.00 | 1.92 | 2.38 | 1.48 |
| 2.00 | 3.25 | 1.32 | 0.80 | 0.94 | 0.66 | 1.49 | 1.87 | 0.97 |
| 1.25 | 1.91 | -0.03 | 0.48 | 0.62 | 0.33 | 0.92 | 1.50 | 0.60 |
| -0.40 | 0.57 | -1.37 | 0.17 | 0.31 | 0.03 | 0.45 | 0.90 | 0.00 |
| -1.74 | -0.78 | -2.71 | -0.12 | 0.20 | -0.43 | -0.01 | 0.83 | -0.86 |
| -3.09 | -2.12 | -4.06 | -0.39 | 0.12 | -0.89 | -0.45 | 0.89 | -1.78 |
| -4.43 | -3.46 | -5.40 | -0.64 | 0.09 | -1.37 | -0.85 | 1.04 | -2.75 |
| -5.77 | -4.80 | -6.74 | -0.88 | 0.09 | -1.84 | -1.23 | 1.27 | -3.73 |
| -7.11 | -6.15 | -8.08 | -1.10 | 0.12 | -2.31 | -1.59 | 1.55 | -4.72 |
| -8.46 | -7.49 | -9.43 | -1.31 | 0.16 | -2.78 | -1.92 | 1.87 | -5.72 |
| -9.80 | -8.83 | -10.77 | -1.50 | 0.24 | -3.24 | -2.24 | 2.24 | -6.72 |
| -11.14 | -10.17 | -12.11 | -1.69 | 0.32 | -3.70 | -2.53 | 2.65 | -7.71 |
| -12.48 | -11.52 | -13.45 | -1.86 | 0.43 | -4.16 | -2.81 | 3.08 | -8.69 |
| -13.83 | -12.86 | -14.80 | -2.03 | 0.55 | -4.60 | -3.06 | 3.54 | -9.67 |

where l is physical length of mobile object and v average relative speed.

In order to verify collision avoidance in Relevant predicted collision time it is necessary to evaluate 95 % confidence limits values as well as main variables of both coordinates.

In the example used in this research variable X is detected to be in collision with upper 95 % confidence value inside of Relevant predicted collision time at $t = 11$ and variables Y and Z with upper 95 % confidence value inside of Relevant predicted collision time at $t = 12$. This represents marginal case. This intersection situation is illustrated in Fig. 9.5.

At this point mobile agent A performs evading maneuver per any implemented path planning algorithm considering intersection point of collision with object B as static obstacle. In future research algorithm for velocity alteration for collision avoidance will be analyzed.

First part of Fig. 9.6 represents current path of mobile agent in 3D and its respective 2D coordinates. Variables X, Y and Z analyzed in forecasted time are forming 4D hyperspace, representing 3D space changing shape and size in time as 4th dimension. That is represented in Fig. 9.6. This dynamically metamorphous hyperspace finally represents potential collision area for other mobile agents. Output of this process of collision detection will be used in mobile agents' path planning algorithm.

Depending on the relative objects velocity, mass, size and agility this Relevant predicted collision time will be t_{n+cag} , where cag represent the value which in this

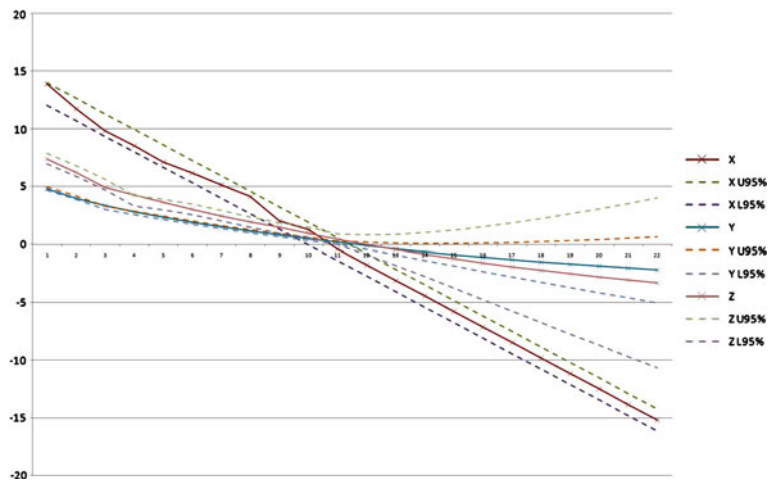
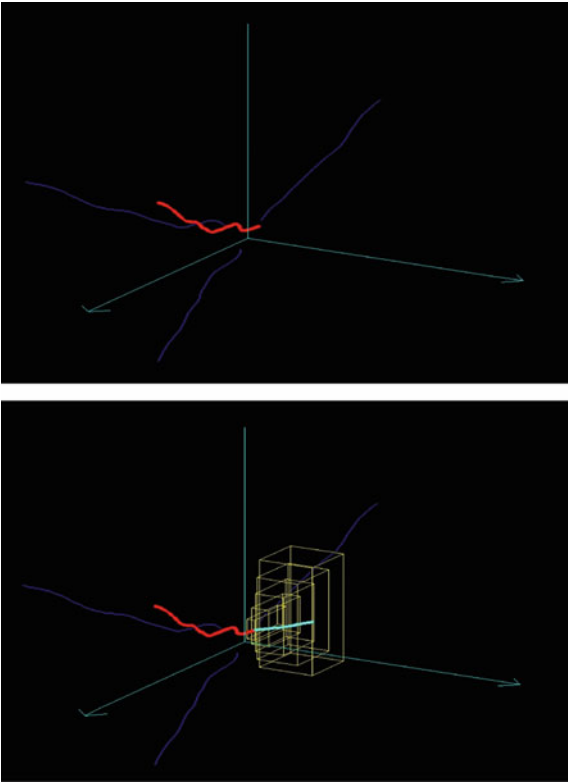


Fig. 9.5 Intersection graph showing X, Y and Z variables, their respective 95 % upper and lower confidence limits and their intersection with 0 value line

Fig. 9.6 Current path in last sampling period in 3D coordinate system with respective 2D coordinate paths. Second image shows hyperspace of future time-space state of mobile agent



research is called **Coefficient of agility**. Coefficient of agility is measured value based of physical characteristics of the mobile agent including minimal and maximal velocity, maximal deflection angle, maximal response time from control mechanism to motors etc. Coefficient of agility has to be measured for each mobile agent and will be expressed in time units necessary for mobile agent to modify path and avoid collision.

In case of lack of agility measurement, velocity adjustment should be applied. However, velocity adjustment(s) will result in revision of calculation and prediction because they lead to modification of measurement time periods. There are two main modes of this method:

- Collaborative mode—based on active coordinate exchange,
- Non-Collaborative mode—based on sensor detection and coordinate mapping.

In this research results are based on collaborative mode. However, algorithm is applicable to non-collaborative mode as well.

Research findings shows that this method can be successfully used for collision detection for variable sized group of mobile agents in unstructured and unpredictable environment. Implications of this research are in the area of mobile robots to automobiles and airborne objects.

9.4 Path Prediction Algorithm

As explained in previous chapter concept is based on mathematical statistical model of path prediction. For successful usage of proposed concept it is necessary to obtain valid information on mobile agents coordinate in equal time periods. Input data for algorithm are ID information of mobile agents and their respective coordinates.

Mentioned variables are inputs for algorithm which flow diagram is shown on Fig. 9.7.

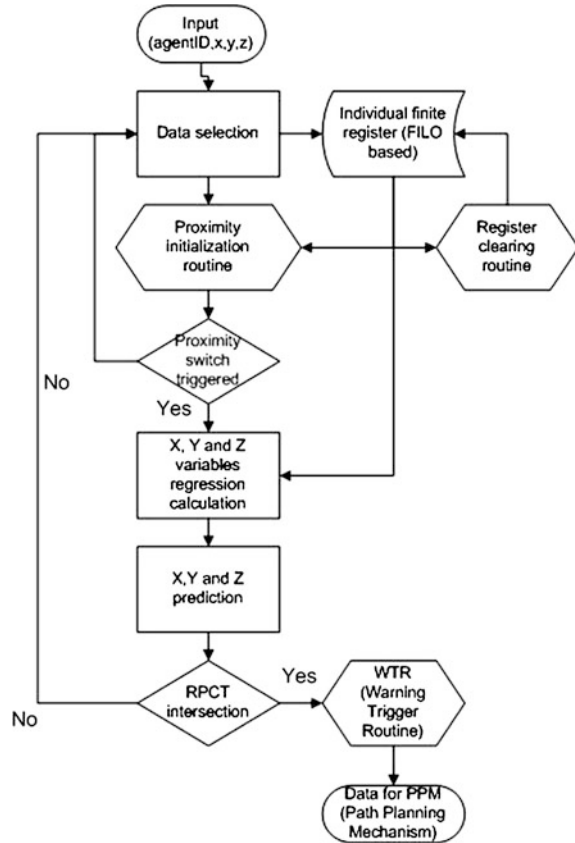
In following lines pseudo code for algorithm displayed on flow diagram on Fig. 9.7 is presented:

```

routine capture_data (agentID, x, y, z)
{read register agentID, x[], y[], z[];
sort agentID, x[], y[], z[]
store agentID, x[], y[], z[]};
routine path_prediction_routine (register. agentID)
{calculate regression x[],y[], z[];
predictRegister. agentID = predict_path(agentID);
if (RPCT = True)
{call external_route_plnning(predictRegister. agentID);}
}
routine regression (x[],y[], z[])
{calculate regression for x[], y[], z[];

```


Fig. 9.7 Flow diagram for collaborative and non-collaborative dynamic path prediction algorithm for mobile agents collision detection with dynamic obstacles



```
return regression analysis;}
```

```
routine predict_path (agentID)
```

```
{read register.agentID;
calculate prediction x[];
calculate prediction y[];
calculate prediction z[];
return prediction x[], y[], z[];}
```

```
routine main(void)
```

```
{call capture_data;
if x < criticalTresholdX and y < criticalTresholdY and z < criticalTresholdZ
{call path_prediction_routine;}
else
{call capture_data;}
if register.agentID.status = Full;
{call shift_register;}
}
```

Implementation of illustrated algorithm allows mobile agent to constantly monitor movements of other mobile agents representing dynamical obstacles. Figures 9.3, 9.4 and 9.5 as well as Table 9.1 indicates experimental results in simulated scenario.

9.5 Future Research

Future research in domain of mentioned method development will be concentrated on measuring Coefficient of agility of the mobile agent and minimal time period for signaling coordinates or sampling frequency, coping with object size and prediction of closest part as well as velocity modification for collision avoidance.

9.6 Conclusion

This research resulted in operative algorithm for collaborative and non-collaborative dynamic path prediction for mobile agent's collision detection with dynamic obstacles which can be used in mobile robotic, automobile industry and aeronautics. Special value of this method and algorithm is because it allows advantages such as:

- Increase of autonomy,
- Decrease of uncertainty,
- Allows non-hierarchical operation,
- Simplify inclusion of additional mobile agents into system.

References

1. Umedachi T, Takeda K, Nakagaki T, Kobayashi R, Ishiguro A (2010) Fully decentralized control of a soft-bodied robot inspired by true slime mold. *Biol Cybern* 102(3):261–269. Available from: academic search complete, Ipswich, MA. Accessed 12 May 2010
2. Feng L et al (1993) Cross-coupling motion controller for mobile robots. *IEEE control systems*, 0272-1708/93
3. Forsberg J et al (1995) Mobile robot navigation using the range-weighted Hough transform. *IEEE Robot Autom Mag* 2:18–25
4. Gat E (1998) Three-layer architectures. *Artificial intelligence and mobile robots*, AAAI/The MIT Press, Cambridge, pp 195–210
5. Luger GF, Stubblefield WA (1998) *Artificial intelligence*. Addison Wesley Longman, Reading, pp 247–292
6. Thrun S et al (2006) *Probabilistic robotics*. Massachusetts Institute of Technology
7. Graf B et al (2004) Mobile robot assistants. *IEEE Robot Autom Mag* 1070-9932/04:68–69

8. Roy N et al (2002) Collaborative robot exploration and rendezvous algorithms, performance bounds and observations. *ARJournal*, Kluwer Academic Publishers, Dordrecht
9. Roy N et al (2003) Planning under uncertainty for reliable health care robotics. In: The 4th international conference on field and service robotics, Pittsburgh, 14–16 July 2003
10. Breazeal C et al (2007) Efficient model learning for dialog management. In: 2nd ACM/IEEE international conference on human-robot interaction, Arlington, 10–12 March 2007
11. Roy N, McCallum A (2001) Toward optimal active learning through monte carlo estimation of error reduction. In: Proceedings of the international conference on machine learning (ICML 2001), Williamstown
12. Roy N, Gordon G (2002) Exponential family PCA for belief compression in POMDPs. In: Advances in neural information processing (15) NIPS, Vancouver, Dec 2002
13. Brunskill E (2008) CORL: a continuous-state off-set-dynamics reinforcement learner, UAI
14. Klapka P (2001) Kybernetika a umělá inteligence
15. van Waveren JMP, Rothkrantz LJM (2008) Automated static and dynamic obstacle avoidance in arbitrary 3D polygonal worlds. Mobile robots motion planning new challenges, pp 455–468
16. Gnadt W, Grossberg S (2008) SOVEREIGN: an autonomous neural system for incrementally learning to navigate towards a rewarded goal. Mobile robots motion planning new challenges, pp 99–119
17. Casini M, Garulli A, Giannitrapani A, Vicino A (2009) A matlab-based remote lab for multi-robot experiments. In: 8th IFAC symposium on advances in control education, Kumamoto, 21–23 Oct 2009
18. Payá L, Reinoso O, Sánchez A, Gil A, Fernández L (2009) An educational tool for mobile robots remote interaction. In: 8th IFAC symposium on advances in control education, Kumamoto, 21–23 Oct 2009
19. Cezayirli A, Kerestecioglu F (2009) On preserving connectivity of autonomous mobile robots. In: IEEE international symposium on intelligent control, Saint Petersburg, 8–10 July 2009
20. Jonathan AR, Aeyels D (2009) Multi-robot coverage to locate fixed and moving targets. In: IEEE international symposium on intelligent control, Saint Petersburg, 8–10 July 2009
21. Parlaktuna O, Sipahioglu A, Kirlik G, Yazici A (2009) Multi-robot sensor-based coverage path planning using capacitated arc routing approach. In: IEEE international symposium on intelligent control, Saint Petersburg, 8–10 July 2009
22. Kurabayashi D et al (1994) Cooperative sweeping by multiple mobile robots. *Proc IEEE Int Conf Robot Automat* 3:1744–1749
23. Latimer D et al (2002) Towards sensor based coverage with robot teams. *Proc IEEE Int Conf Robot Automat* 1:961–967
24. Mei Y, Yung-Hsiang L, Charlie HY, George LCS (2006) Deployment of mobile robots with energy and timing constraints. *IEEE Trans Robot Automat* 22(3):507–522
25. M. Ozkan, A. Yazici, M. Kapanoglu, O. Parlaktuna (2009) A genetic algorithm for task completion time minimization for multi-robot sensor-based coverage. In: IEEE International symposium on intelligent control, Saint Petersburg, July 8–10
26. Ferrara A, Rubagotti M (2009) A dynamic obstacle avoidance strategy for a mobile robot based on sliding mode control. In: 18th IEEE international conference on control applications, Saint Petersburg, July 8–10
27. Teymur C, Temeltaş H (2010) A new behavior combining method for mobile robots. *ITU journal series D: engineering*, 1 Feb 2010
28. Jeong HK, Choi KH, Kim SH, Kwak YK (2008) Driving mode decision in the obstacle negotiation of a variable single-tracked robot. *Adv Robot* 22:1421–1438
29. Babovic E (2011) Collaborative and non-collaborative dynamic path prediction algorithm for mobile agents collision detection with dynamic obstacles in a two-dimensional space, IEEM, Singapore

Chapter 10

Website Analysis of Top 100 Most Valuable Companies in Romania

Lavinia D. Rusu and Liciniu A. Kovács

Abstract There exists today a wide range of sites on the Web, from personal to content oriented, and from academic to purely commercial websites, and certainly others will appear in the near future. Unfortunately, so many websites cannot capture the attention and interest of visitors/customers, many of the existing websites being poorly designed. The main purpose of this study is to identify the best designed website in the top 100 most valuable companies in Romania. Due to their nature, the criteria and the number of awarded points were valued according to a subjective basis, as described in papers of Gálfi et al. [1] and Kovács et al. [2]. On the one hand, as we will see, there is little or no connection between the company's position in the top and the quality of their website. On the other hand, the analyzed websites are far from perfection, and we would recommend that the administrators and/or webmasters pay more attention to pages design and upgrading frequency.

L. D. Rusu · L. A. Kovács (✉)
Department of Business, Babeş-Bolyai University, 7 Horea Street,
400174 Cluj-Napoca, Romania
e-mail: liciniu@yahoo.com
URL: <http://www.liciniu.ro>

L. D. Rusu
e-mail: rusu_lavy@yahoo.com

10.1 Introduction

A. Case study objectives:

- to conduct a comparative analysis in order to find out the development stage of specific webpages and/or webpage elements of top 100 most valuable companies in Romania;
- to classify the companies in each industry sector, based on the total number of points obtained by the *specific* webpages and/or webpage elements taken into consideration;
- to classify the top 100 most valuable companies in Romania, based on the total number of points obtained by the *common* webpages and/or webpage elements taken into consideration.

Note: top 100 most valuable companies in Romania and the evaluation criteria developed by Capital Partners were published in “Ziarul Financiar” (“Financial Newspaper”), and on its dedicated webpage on November 27, 2010.

B. Case study methods:

- classifying all 100 companies according to industry sectors;
- determining what *specific* webpages and/or webpage elements to visit by considering the industry sector;
- establishing the webpages and/or webpage elements that are *common* to all of top 100 most valuable companies in Romania;
- visiting and evaluate the chosen webpages and/or webpage elements of all top 100 most valuable companies in Romania;
- entering data into a Microsoft Excel spreadsheet to determine the values of each page/element of the websites taken into consideration. In the spreadsheet tables we calculated the total values, mean values, and elaborated charts, as shown below.

10.2 Top 100 Most Valuable Companies in Romania According to Industry Sectors

As we can see in Figs. 10.1, 10.2, the industry sectors of top 100 most valuable companies in Romania are as follows:

- 26 (over a quarter of the) companies are in the oil and energy business,
- 17 are banks and insurance companies,
- 17 are industrial and construction companies,
- 13 companies are producers and importers of consumer goods,
- 8 are telecom companies,

Fig. 10.1 Top 100 most valuable companies in Romania according to industry sectors

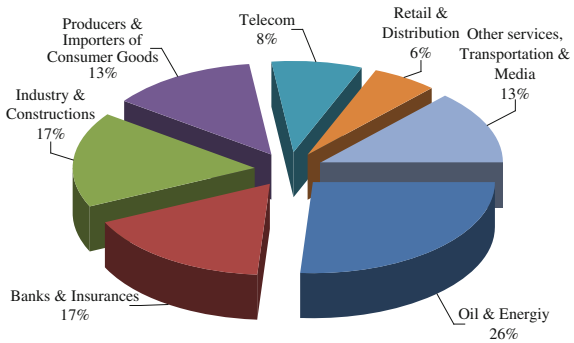
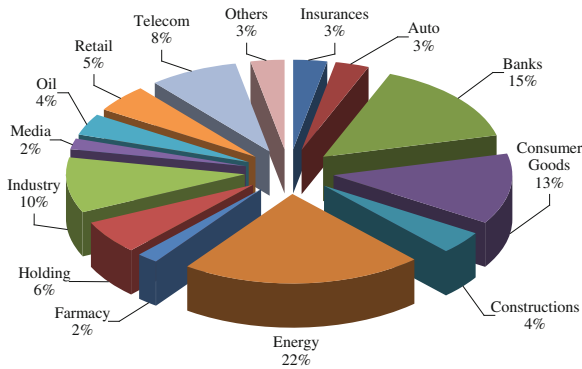


Fig. 10.2 Top 100 most valuable companies in Romania according to industry sectors (detailed)



- 6 companies are in the retail and distribution business,
- 3 are service, transportation and media companies.

Note: We noticed that only 97 of the top 100 most valuable companies in Romania (in 2010) have websites. In other words, the analysis is limited to 97 sets of data.

10.3 Website Pages/Elements to Be Analyzed

At this stage, we determined the *specific* webpages and/or webpage elements to be visited for each industry sector as seen in Table 10.1. The table also reflects the 7 webpages and/or webpage elements *common* to all of top 100 most valuables companies in Romania.

Table 10.1 (continued)

| Pages/elements to be analyzed | News | Search | Home | FAQ | Terms and conditions | Press release | Sustainable development | Location | Reports | Archive | Program | Online broadcast | Site map |
|----------------------------------|---------------------------------|--------|------|-----|-------------------------|---------------|----------------------------|----------|---------|---------|---------|---------------------|----------|
| Activity domains | Specific website pages/elements | | | | | | | | | | | | |
| Industry | | | x | | | x | | | | | | | x |
| Media | | | | x | | | | | | x | | x | |
| Oil | | | x | | | | | | x | | | | |
| Retail | | x | | | | x | | x | | | | | x |
| Telecom | x | | | | x | | | | | | | | x |

Fig. 10.3 Results for
“About Us” pages

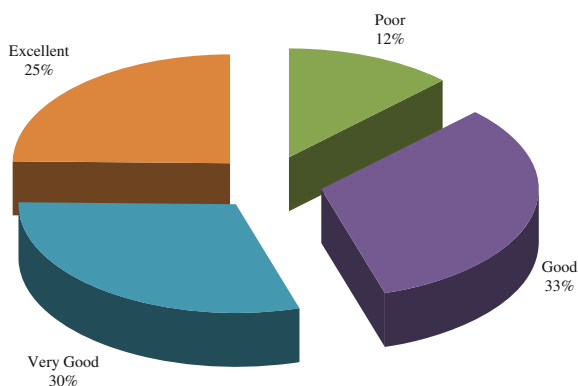
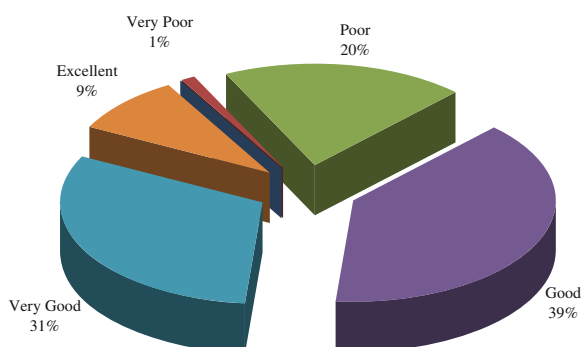


Fig. 10.4 Results for
“Contact Information” pages



10.4 Interim Results

For each industry sector determined earlier and specified below, the following companies achieved the best results:

- Insurance: GROUPAMA ASIGURĂRI, <http://www.groupama.ro>, 37 points;
- Auto: AUTOMOBILE DACIA, <http://www.dacia.ro>, 35 points;
- Banks: BRD-SOCGEN (BRD), <http://www.brd.ro>, 37 points;
- Consumer goods: BRITISH AMERICAN TOBACCO, <http://www.bat.com>, 40 points;
- Constructions: HOLCIM, <http://www.holcim.ro>, 38 points;
- Energy: E.ON MOLDOVA DISTRIBUȚIE, <http://www.eon-energie.ro>, 39 points;
- Pharmacy: TERAPIA, <http://www.terapia.ro>, 33 points;
- Holding: SIF BANAT-CRIȘANA (SIF1), <http://www.sif1.ro>, 37 points;
- Industry: PIRELLI, <http://www.pirelli.com>, 42 points;
- Media: PRO TV SA, <http://www.protv.ro>, 41 points;
- Oil: PETROM (SNP), <http://www.petrom.com>, 41 points;

Fig. 10.5 Results for “Careers” pages

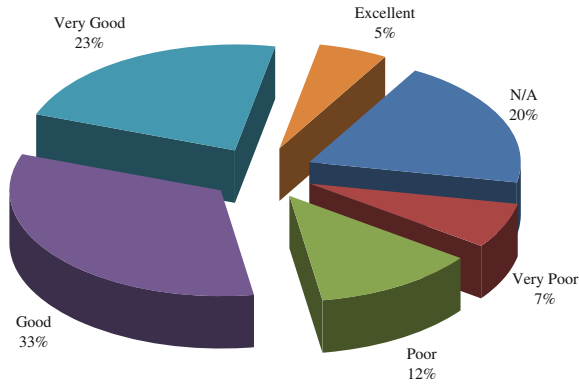
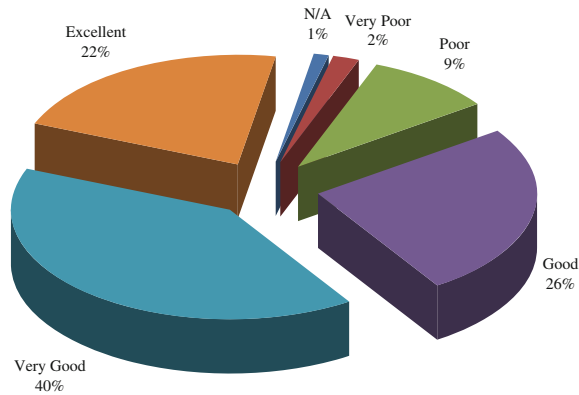


Fig. 10.6 Results for “Products and/or Services” pages



- Retail: SELGROS, <http://www.selgros.ro>, 37 points;
- Telecom: NOKIA, <http://www.nokia.ro>, 38 points.

10.5 Percentages and Mean Values

Based on the total number of awarded points, we can draw the following conclusions about the 97 analyzed websites:

- “About Us” pages are classified as follows: 25 % are excellent, 30 % very good, 33 % good, and 12 % poor (Fig. 10.3);
- “Contact Information” webpages are: 9 %—excellent, 31 %—very good, 39 %—good, 20 %—poor, and 1 %—very poor (Fig. 10.4);

Fig. 10.7 Results for “Foreign Language Versions” pages

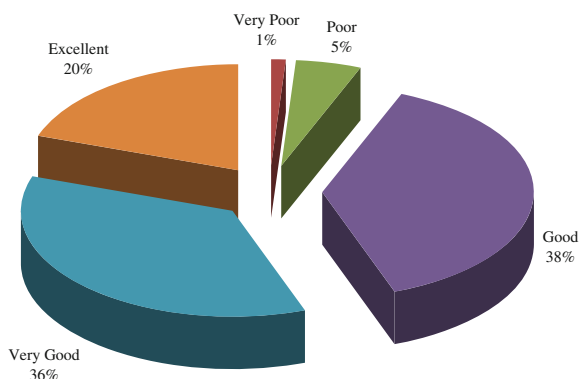
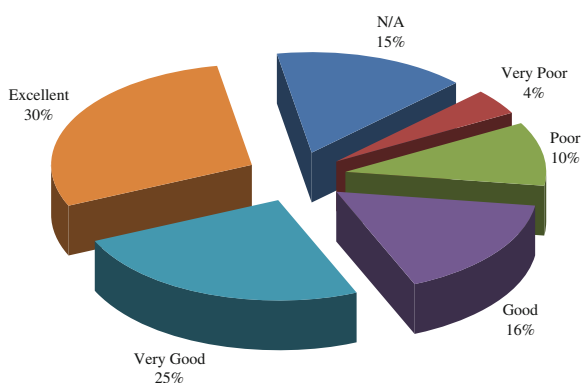


Fig. 10.8 Results for “Attractiveness” criterion



- Pages that refer to “Careers” are nonexistent on 20 % of websites, while the existent ones are classified as follows: 5 % are excellent, 23 % very good, 33 % good, 12 % poor, and 7 % very poor (Fig. 10.5);
- “Products and/or Services” pages are classified as follows: 22 % are excellent, 40 % very good, 26 % good, 9 % poor, 2 % very poor, and 1 % nonexistent (Fig. 10.6);
- Pages that refer to “Foreign Language Versions” are nonexistent on 15 % of websites, while the existent ones are classified as follows: 30 % are excellent, 25 % very good, 16 % good, 10 % poor, and 4 % very poor (Fig. 10.7);
- The “Attractiveness” criterion has yielded the following results: 20 % of the websites are excellent, 36 % very good, 38 % good, 5 % poor, and 1 % very poor (Fig. 10.8);
- The “Navigation easiness” criterion has yielded the following results: 15 % of the websites are excellent, 49 % very good, 29 % good, 6 % poor, and 1 % very poor (Fig. 10.9).

The calculated mean values (for the analyzed website pages/elements) are as follows:

Fig. 10.9 Results for “Navigation easiness” criterion

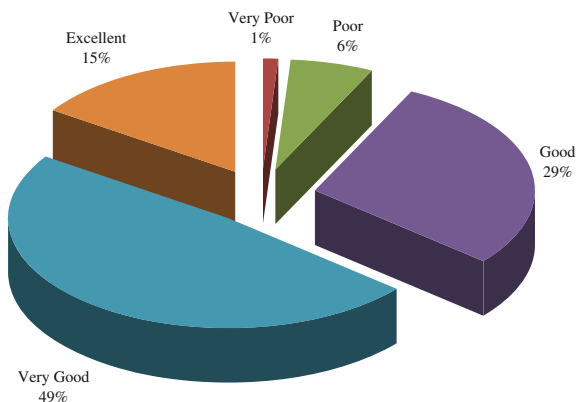


Fig. 10.10 Data that refer to top 100 most valuable companies in Romania (<http://www.zf.ro/zf-english/most-valuable-companies-in-romania-7723482/>)

- 3.67 for “About Us”;
- 3.28 for “Contact Information”;
- 2.47 for “Careers” (the lowest);
- 3.67 for “Products and/or Services”;
- 3.18 for “Foreign Language Versions”;
- 3.68 for “Attractiveness”;
- 3.71 for “Navigation easiness” (the highest).

Table 10.2 Position in top 100 and website position of 100 top 100 most valuable companies in Romania in year 2010

| Position in top 100 | Website position | Company | Activity domain | Website address | Total points |
|---------------------|------------------|----------------------------|-----------------|---|--------------|
| 90 | 1 | PIRELLI TYRES | Industry | http://www.pirelli.com | 34 |
| 1 | 2 | PETROM (SNP) | Oil | http://www.petrom.com | 32 |
| 28 | 3 | BRITISH AMERICAN TOBACCO | Consumer goods | http://www.bat.com | 32 |
| 69 | 4 | LINDE GAZ | Energy | http://www.linde-gas.ro | 32 |
| 57 | 5 | ELECTRICA TRANSILVANIA SUD | Energy | http://www.efts.ro | 31 |
| 11 | 6 | AUTOMOBILE DACIA | Auto | http://www.dacia.ro | 30 |
| 25 | 7 | BANCA TRANSILVANIA (TLV) | Banks | http://www.bancatransilvania.ro | 30 |
| 29 | 8 | NOKIA ROMÂNIA | Telecom | http://www.nokia.ro | 30 |
| 85 | 9 | GROUPAMA ASIGURĂRI | Insurance | http://www.groupama.ro | 30 |
| 7 | 10 | BRD-SOGEN (BRD) | Banks | http://www.brd.ro | 29 |
| 38 | 11 | URSUS BREWERIES | Consumer goods | http://www.ursus-breweries.ro | 29 |
| 51 | 12 | E.ON MOLDOVA DISTRIBUTIE | Energy | http://www.eon-energie.ro | 29 |
| 70 | 13 | QAB (Pepsi) | Consumer goods | http://www.pepsico.com | 29 |
| 86 | 14 | ARCTIC | Industry | http://www.arctic.ro | 29 |
| 43 | 15 | TRANSELECTRICA (TEL) | Energy | http://www.transelectrica.ro | 28 |
| 47 | 16 | SELGROS | Retail | http://www.selgros.ro | 28 |
| 55 | 17 | E.ON GAZ | Energy | http://www.eon-gaz-romania.ro | 28 |
| 89 | 18 | ENEL ENERGIE | Energy | http://www.enel.ro | 28 |
| 93 | 19 | SCHAEFFLER | Industry | http://www.schaeffler.ro | 28 |
| 94 | 20 | HENKEL ROMÂNIA | Consumer goods | http://www.henkel.ro | 28 |
| 16 | 21 | CEZ DISTRIBUTIE | Energy | http://www.cez.ro | 27 |
| 19 | 22 | HOLCIM | Consumer goods | http://www.holcim.ro | 27 |
| 36 | 23 | P&G | Constructions | http://www.pgalkans.com | 27 |
| 39 | 24 | COSMOTE | Consumer goods | http://www.cosmote.ro | 27 |
| 45 | 25 | APA NOVA | Telecom | http://www.apanovabucuresti.ro | 27 |
| 64 | 26 | CONTINENTAL AUTOMOTIVE | Consumer goods | http://www.conti-online.com | 27 |

(continued)

Table 10.2 (continued)

| Position in top 100 | Website position | Company | Activity domain | Website address | Total points |
|---------------------|------------------|----------------------------|-----------------|---|--------------|
| 72 | 27 | JT INTERNATIONAL | Auto | http://www.jti.com | 27 |
| 78 | 28 | SIF BANAT-CRIȘANA (SIFI) | Consumer goods | http://www.sifl.ro | 27 |
| 83 | 29 | UNILEVER | Holding | http://www.unilever.ro | 27 |
| 92 | 30 | COCA-COLA HBC | Consumer goods | http://www.thecoca-colacompany.com | 27 |
| 4 | 31 | ORANGE | Telecom | http://www.orange.ro | 26 |
| 5 | 32 | VODAFONE | Telecom | http://www.vodafone.ro | 26 |
| 13 | 33 | RAIFFEISEN BANK | Banks | http://www.raiffeisen.ro | 26 |
| 46 | 34 | EXIMBANK | Banks | http://www.eximbank.ro | 26 |
| 49 | 35 | PHILIP MORRIS | Consumer goods | http://www.pmi.com | 26 |
| 52 | 36 | ALLIANZ - ȚIRIAC ASIGURĂRI | Insurance | http://www.allianztiriac.ro | 26 |
| 62 | 37 | ALPHA BANK | Banks | http://www.alphabank.ro | 26 |
| 66 | 38 | PIRAEUS BANK | Banks | http://www.piraeusbank.ro | 26 |
| 80 | 39 | TERAPIA | Pharmacy | http://www.terapia.ro | 26 |
| 99 | 40 | KRAFT FOODS | Consumer goods | http://www.kraftfoodscompany.com | 26 |
| 8 | 41 | BANCA COMERCIALĂ ROMÂNĂ | Banks | http://www.bcr.ro | 25 |
| 12 | 42 | TRANSRAZ (TGN) | Energy | http://www.transgaz.ro | 25 |
| 21 | 43 | UNICREDIT ȚIRIAC BANK | Banks | http://www.unicredit-tiriac.ro | 25 |
| 26 | 44 | ALRO (ALR) | Industry | http://www.alro.ro/home.ro.html | 25 |
| 31 | 45 | AEROPORTUL HENRI COANDĂ | Transport | http://www.otp-airport.ro | 25 |
| 56 | 46 | BANCPOST | Banks | http://www.bancpost.ro | 25 |
| 67 | 47 | BANCA ROMÂNNEASCĂ | Banks | http://www.banca-romaneasca.ro | 25 |
| 9 | 48 | NUCLEAR ELECTRICĂ | Energy | http://www.nuclearelectrica.ro | 24 |
| 10 | 49 | ROMTELECOM | Telecom | http://www.romtelecom.ro | 24 |
| 22 | 50 | METRO CASH & CARRY | Retail | http://www.metro.ro | 24 |
| 41 | 51 | HEINEKEN | Consumer goods | http://www.heineken.ro | 24 |
| 44 | 52 | E.ON GAZ DISTRIBUTIE | Energy | http://www.eon-gaz-distributie.ro | 24 |

(continued)

Table 10.2 (continued)

| Position in top 100 | Website position | Company | Activity domain | Website address | Total points |
|---------------------|------------------|------------------------------|-----------------|---|--------------|
| 53 | 53 | LUKOIL | Oil | http://www.lukoil.ro | 24 |
| 71 | 54 | REAL | Retail | http://www.real-hypermart.ro | 24 |
| 96 | 55 | HOLZINDUSTRIE SCHWEIGHOFER | Industry | http://www.schweighofer.at | 24 |
| 15 | 56 | LAFARGE CIMENT | Constructions | http://www.lafarge.ro | 23 |
| 18 | 57 | ARCELMITTAL GALATI | Industry | http://www.arcelormittal.com | 23 |
| 33 | 58 | CARREFOUR | Retail | http://www.carrefour.ro | 23 |
| 58 | 59 | OMNIASIG VIENNA INSURANCE | Insurance | http://www.omniasig.ro | 23 |
| 79 | 60 | GRUP SERVICII PETROLIERE | Industry | http://www.gspoffshore.com | 23 |
| 97 | 61 | ALCATEL - LUCENT | Telecom | http://www.alcatel-lucent.com | 23 |
| 100 | 62 | JOHNSON CONTROLS | Industry | http://www.johnsoncontrols.ro | 23 |
| 2 | 63 | FONDUL PROPRIETATEA | Holding | http://www.fondulproprietatea.ro | 22 |
| 24 | 64 | CEC BANK | Banks | http://www.cec.ro | 22 |
| 30 | 65 | PRO TV SA | Media | http://www.protv.ro | 22 |
| 73 | 66 | A&D PHARMA | Pharmacy | http://www.adpharma.ro | 22 |
| 77 | 67 | MOL | Oil | http://www.molromania.ro | 22 |
| 91 | 68 | SILCOTUB | Industry | http://www.tenaris.com/romania | 22 |
| 34 | 69 | CARPATCEMENT | Constructions | http://www.heidelbergcement.ro | 21 |
| 63 | 70 | ING BANK SUCURSALA BUCUREȘTI | Banks | http://www.ing.ro | 21 |
| 88 | 71 | UPC ROMÂNIA | Telecom | http://www.upc.ro | 21 |
| 98 | 72 | CITIBANK | Banks | http://www.citibank.ro | 21 |
| 3 | 73 | HIDROELECTRICA | Energy | http://www.hidroelectrica.ro | 20 |
| 32 | 74 | KAUFLAND | Retail | http://www.kaufland.ro | 20 |
| 35 | 75 | RCS & RDS | Telecom | http://www.rcs-rds.ro | 20 |
| 40 | 76 | ELECTRICA MUNTENIA NORD | Energy | http://www.mnd.electrica.ro | 20 |
| 68 | 77 | PORSCHE ROMÂNIA | Auto | http://www.porscheromania.ro | 20 |
| 74 | 78 | RBS BANK | Banks | http://www.rbs.ro | 20 |

(continued)

Table 10.2 (continued)

| Position in top 100 | Website position | Company | Activity domain | Website address | Total points |
|---------------------|------------------|-----------------------------|-----------------|---|--------------|
| 75 | 79 | VOLKSBANK | Banks | http://www.volksbank.ro | 20 |
| 87 | 80 | TV ANTENA 1 | Media | http://www.a1.ro | 20 |
| 17 | 81 | GDF SUEZ ENERGY | Energy | http://www.gdfsuez-energy.ro | 19 |
| 42 | 82 | CE CRAIOVA | Energy | http://www.cencraiova.ro | 19 |
| 6 | 83 | ROMGAZ | Energy | http://www.romgaz.ro | 18 |
| 48 | 84 | ROMPETROL RAFINARE (RRC) | Oil | http://www.rompetrol-rafinare.ro | 18 |
| 82 | 85 | ROMSTRADE | Constructions | http://www.romstrade.ro | 18 |
| 23 | 86 | LOTERIA ROMÂNĂ | Services | http://www.loto.ro | 17 |
| 95 | 87 | SIF MUNTENIA (SIF4) | Holding | http://www.sifmuntenia.ro | 17 |
| 50 | 88 | CE ROVINARI | Energy | http://www.cerovinari.ro | 16 |
| 81 | 89 | SIF TRANSILVANIA (SIF3) | Holding | http://www.siftransilvania.ro | 16 |
| 84 | 90 | SIF MOLDOVA (SIF2) | Holding | http://www.sifm.ro | 16 |
| 20 | 91 | CE TURCENI | Energy | http://www.eturceni.ro | 15 |
| 27 | 92 | ELECTROCENTRALE BUCUREȘTI | Energy | http://www.elcen.ro | 15 |
| 59 | 93 | ELECTRICA TRANSILVANIA NORD | Energy | http://www.cj.electrica.ro | 14 |
| 65 | 94 | SIF OLTEANIA (SIF5) | Holding | http://www.sifolt.ro | 14 |
| 76 | 95 | BERGENBIER | Consumer goods | http://www.prieteniiistudece.ro | 14 |
| 60 | 96 | INTERAGRO | Industry | http://www.interagro.ro | 11 |
| 54 | 97 | INTERBRANDS | Distribution | http://shop.interbrands.ro | 10 |
| 14 | 98 | Enel Distribuție Muntenia | Energy | N/A | 0 |
| 37 | 99 | Enel Distribuție Banat | Energy | N/A | 0 |
| 61 | 100 | Enel Distribuție Dobrogea | Energy | N/A | 0 |

Notes

The website analysis of top 100 most valuable companies in Romania was conducted between March and July, 2011

Some data that refer to top 100 most valuable companies in Romania are published on “Ziarul Financiar” (ZF) dedicated webpage at <http://www.zf.ro/zf-english/most-valuable-companies-in-romania-7723482/> (Fig. 10.10)

The highest and the lowest results:

- The highest result was obtained by the website belonging to Pirelli Tyres—<http://www.pirelli.com/> (34 points and a mean value of 4.86);
- The lowest result was obtained by the website belonging to Interbrands—<http://www.shop.interbrands.ro/> (10 points and a mean value of 1.43).

10.6 Conclusions

Based on the final results of this comparative study shown on Table 10.2 we notice that:

- All of the 97 websites examined failed to obtain the maximum score possible, i.e. 35 points in the analysis of the *common* website pages/elements taken into consideration;
- Three companies of the top 100 most valuable companies in Romania (year 2010) do not have websites—therefore we recommend building/creating such sites;
- The analyzed websites are far from being perfect, and we recommend that their administrators/webmasters pay more attention to the design, upgrade frequency, and continuous improvement.

In order to obtain better results, one can increase the number of visitors who evaluate websites, can use several Internet browsers (e.g. Internet Explorer, Google Chrome, Opera, Mozilla, Konqueror, Netscape, Mozilla Firefox, Hot Java Browser, etc.) under different operating systems (e.g. Windows, Linux, Solaris, etc.), and different versions for the same browser and/or operating system.

Acknowledgments I am grateful to Ms. Monica Livia Cormoș for revising this paper.

References

1. Gálfi VM, Kovács LA, Chifu-Oros CI, Moldovan Ș (2005) Web presence of travel agencies from Transylvania-Romania and Hungary. In: SINTES 12 international symposium, vol 3, XII edn, pp 521–526, 20–22 Oct 2005, Craiova, Romania. <http://www.liciniu.ro>
2. Kovács LA, Rus VR, Chifu CI (2006) Case study on French, Greek and Romanian hotel websites—a comparative approach. *Int J Bus Res* VI(1):163–170. <http://www.liciniu.ro>

Chapter 11

Comparison of PI and Fractional PI Controllers on a Hydraulic Canal Using Pareto Fronts

Y. Chang

Abstract In evolutionary computation, a lot of research has been done on multi-objective optimization (MOO). MOO is based on the concept of Pareto-optimal sets, also known as, Pareto-optimal fronts. However, there has been very little research done in comparing controllers using Pareto-optimal fronts. The problem of designing controllers can be viewed as a MOO problem where we try to optimize the performance, robustness and other characteristics of the controller. This paper uses a plant model from a hydraulics application which compares a proportional-integrator (PI) controller and fractional PI controller (FPI). A Pareto-optimal front is generated for each of these controllers. Since each objective is an extra dimension, if we optimize for n objectives then the Pareto-optimal front will have n dimensions. Therefore it is difficult to compare the controllers visually. Firstly, the number of dimensions is reduced using a feature selection technique called population-based incremental learning (PBIL). The Pareto-optimal front points are then classified using a nearest centroid classification. If the classification accuracy is high then the centroid is a good representation of the cluster of points (within the Pareto-optimal front under consideration). The centroids of two fronts were then used to compare the PI and FPI controller.

11.1 Introduction

In the past few decades, there has been an increasing interest in the application fractional-order calculus to engineering problems. In control engineering this calculus has been used for system identification, PID controller design, lead-lag

Y. Chang (✉)

Department of Electrical Engineering, University of Cape Town,
Lovers Walk Street, Cape Town, South Africa
e-mail: Changp89@gmail.com

compensator design, and more [1]. As a particular example Feliu-Batlle et al. [2] designed and compared two PI controllers implemented on a hydraulic canal. One of the controllers was a classical PI controller and the other was a FPI controller. The plant and controllers in the paper by Feliu-Batlle et al. are used as a case study in this paper to further investigate the differences between classical and FPI controllers, with a view to quantifying their relative performance.

The comparison of the controllers is performed by analyzing the Pareto fronts generated by each controller. In essence, controller design simply tries to find the optimal controller for a given plant in terms of a predefined cost function. However since controllers have characteristics that oppose each other (e.g. the response time and robustness), improving the performance of one characteristic may reduce the performance of at least one of the others. Therefore like most multi-objective optimization (MOO) problems there is more than one solution and the main thrust of our research is biased towards a *posteriori* decision making in MOO [3, 4].

By generating the set of optimal solutions for both the PI and FPI controller, a quantitative comparison between the two controllers can be made [5, 6]. This was done by Moore [5] by considering unary and binary hyper-volumes while Ho [6] compared the controllers by using parallel co-ordinates and level diagrams.

11.2 Background

11.2.1 Fractional-Order PI Controller

There are many definitions for fractional-order operators [7]. One of the frequently used definitions is the Riemann–Liouville definition, which is

$${}_a D_t^\alpha f(t) = \frac{1}{\Gamma(m-\alpha)} \left(\frac{d}{dt} \right)^m \int_a^t \frac{f(\tau)}{(t-\tau)^{1-(m-\alpha)}} d\tau \quad (11.1)$$

for $m-1 < \alpha < m$, where $\Gamma(\cdot)$ is the Euler gamma function.

This definition can be used for both integration and differentiation. Using the Laplace transform, it can be shown that for a signal $x(t): D^\alpha x(t) = s^\alpha X(s)$ where $\alpha > 0$ and when initial conditions are set to zero. Therefore, a fractional-order differential equation can be written as a transfer function:

$$G(s) = \frac{a_1 s^{\alpha_1} + a_2 s^{\alpha_2} + \dots + a_n s^{\alpha_m}}{b_1 s^{\beta_1} + b_2 s^{\beta_2} + \dots + b_n s^{\beta_n}} \quad (11.2)$$

where $a_i, b_j \in \mathbb{R}$ and $\alpha_i, \beta_j > 0$ for $1 < i < m$, $1 < j < n$, $m \leq n$ and once again where the initial conditions are zero.

Fractional-order PID controllers are typically of the following forms

$$G_0(s) = K_p + \frac{K_I}{s^\lambda} + K_D s^\mu \quad \text{where } \lambda, \mu > 0 \quad (11.3)$$

Clearly, when $\lambda = 1$ and $\mu = 1$ the controller is a classical PID controller.

The FPI controller, where $\mu = 0$, has also been a popular study because of its simplicity since there are only three parameters that require tuning [8].

11.2.2 Pareto-Optimal Fronts

Controller design can be considered a MOO problem. The general MOO problem can be defined as follows: to optimize

$$F(\vec{x}) = \{f_1(\vec{x}), \dots, f_k(\vec{x})\} \in Z \quad (11.4)$$

where $\vec{x} = (x_1, \dots, x_n) \in D$. Optimization is taken as meaning to either minimize or maximize the objective functions depending on how the problem is defined.

A *Pareto-optimal* set of solutions to this problem is one where an objective function cannot be improved by reducing the value on another objective function [9]. Pareto optimality is based on the concept of Pareto dominance, where a vector a is said to dominate vector b (denoted $a > b$) if and only if

$$\begin{aligned} \forall i \in \{1, \dots, n\} : f_i(a) &\geq f_i(b) \wedge \\ \exists j \in \{1, \dots, n\} : f_j(a) &> f_j(b) \end{aligned} \quad (11.5)$$

For a given set, any decision variables that are not dominated by any other decision variable in the set are called *non-dominated*. A set of non-dominated variables is called a *Pareto-optimal set* or a *Pareto-optimal front*.

When searching for solutions to MOO problems, we try to find a set of solutions that can approximate the Pareto-optimal front well [10]. The popular approaches to solving MOO problems are evolutionary algorithms and more recently particle swarm optimization [11, 12].

The two most important factors that are considered when searching for the front are the convergence and diversity. Many indicators have been suggested to optimize for both of these factors, however one of the more popular indicators in the literature is the hyper-volume indicator that was first suggested by Zitzler and Thiele [9]. This indicator has the advantage of being sensitive to any improvement in the front [13] however calculating the indicator is computationally intensive [14]. Figure 11.1 shows the hyper-volumes for a simple two-dimensional example.

11.3 Decision and Objective Space

This paper uses the plant and controllers from the paper by Feliu-Battle et al. [2] as a case study. The plant transfer function is

$$g(s) = \frac{K}{(T_1 s + 1)(T_2 + 1)} e^{-s\tau} \quad (11.6)$$

Fig. 11.1 Hyper-volumes for two example Pareto-optimal fronts

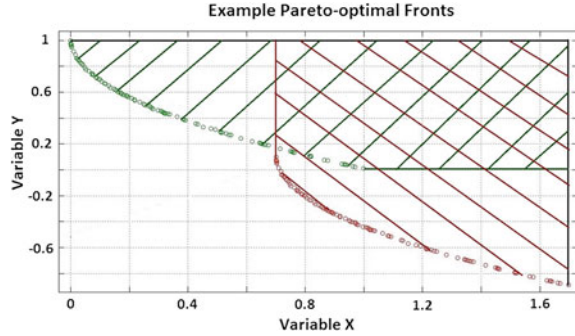


Table 11.1 Plant parameter bounds

| | K | T_1 | T_2 | τ |
|---------|-------|--------|-------|--------|
| Minimum | 0.234 | 7.921 | 0.383 | 2.6 |
| Maximum | 1.1 | 12.419 | 1.566 | 2.6 |

The parameters vary between the bounds shown in Table 11.1.

Feliu-Battle et al. designed a classical PI controller (which will be referred to as the PI controller) and a FPI controller. The controller transfer functions are respectively:

$$C_{PI}(s) = \frac{b_o + b_1 s}{s} = \frac{0.2 + 3.62}{s}, \quad (11.7)$$

and

$$C_{FPI}(s) = \frac{b_o + b_1 s}{s^\lambda} = \frac{0.2 + 3.2s}{s^{0.37}}, \quad (0 < \lambda < 1) \quad (11.8)$$

Two Pareto-optimal fronts were generated, one for each controller. For the PI controller, the decision variables were $\vec{x} = \{b_0, b_1\}$ and for the FPI controller: $\vec{x} = \{b_0, b_1, \lambda\}$. Based on the values chosen for the controllers, the decision space was restricted to $0 < b_0 < 5$ and $0 < b_1 < 3$.

Ideally, the objective functions will be relatively anti-correlated and will have many conflicts [15]. Having too many objectives can increase computation time and make the data difficult to visualize. Also more objectives increase the size of the objective space and therefore require more individuals to be generated to sufficiently cover the space. However, not including an important objective may result in a Pareto-optimal front that may not otherwise have been Pareto-optimal. It is not always possible to know *a priori* what objectives are important. Therefore an over-defined list of objective functions was preferable to an under-defined list.

In this study the following objectives were chosen to be minimized:

1. Maximum percentage overshoot
2. Damping factor

3. Settling time
4. Rise time
5. Bandwidth
6. Integral error squared
7. Integral error time squared
8. Integral input squared

These were chosen because they can be calculated (or approximately) from a time domain analysis. Other objectives that may have been useful are the high frequency gain, the phase and gain margins and more [4, 16].

However, these objectives are correlated with at least one of the other objectives in the list. Also, there are obvious correlations between some of the chosen objectives, such as between settling time and rise time.

11.4 Pareto-Optimal Front

11.4.1 *Generating the Front*

The “Objective-wise Multi-objective Optimization using Self-adaptive Differential Evolution” (OWMOSaDE) algorithm by Huang et al. [17] was slightly modified and used to generate the Pareto-optimal set. In the original algorithm when the maximum size of the archive is reached, less crowded individuals are selected to spread out the solutions. Instead of using the harmonic average distance as a crowding measure, an entropy-based diversity measure by Wang et al. [18] was used. Wang et al. showed that the entropy-based diversity measure generated a more diverse set of solutions than the harmonic average distance.

When calculating the values for the objectives of the FPI controller, the fractional operator was approximated using the Oustaloup algorithm [19]. This is a well-established method for approximating fractional operators using linear transfer functions in the frequency domain [20]. A modified Oustaloup algorithm has been proposed by Xue et al. [20] which is able to approximate over a larger range of frequencies. However, this algorithm was not used because it was found that after a number of trial simulations the algorithm occasionally generated transfer functions with unstable poles.

Once the Pareto-optimal front had been generated, the set of decision variables (two for the PI controller and three for the FPI) was used to calculate four more objectives of the system:

1. open loop phase margin
2. open loop gain margin
3. closed loop oscillations
4. high frequency gain.

This increased the dimension of the front to twelve.

11.4.2 Objective Reduction

There has already been some work done on reducing the number of objectives [15, 21, 22]. Brockhoff and Zitzler [15] use the definition of δ -conflict to determine which objectives to remove. The idea is to remove the objective functions that do not make an error larger than δ in the omitted objectives. However, this means that the omission of objectives is dependent on the values. Therefore an objective that is scaled in a small range is more likely to be removed than an objective with a large range.

Therefore a feature selection approach was adopted. This was done using standard population-based incremental learning [23]. The subsets of the features were evaluated according to the number of points that were Pareto non-dominating after the unselected features had been removed. Thus a subset of features X is better than a subset Y if there are more non-dominating points. In this way, a subset of features that preserved the Pareto-optimal front as faithfully as possible was found.

The Pareto-optimal fronts could be reduced to four dimensions while retaining 89 % of the original fronts. The four final features that were used were:

1. maximum percentage overshoot
2. damping factor
3. rise time
4. closed loop bandwidth.

Removing one of these four features can preserve at most 75 % of the original fronts.

11.5 Controller Comparison

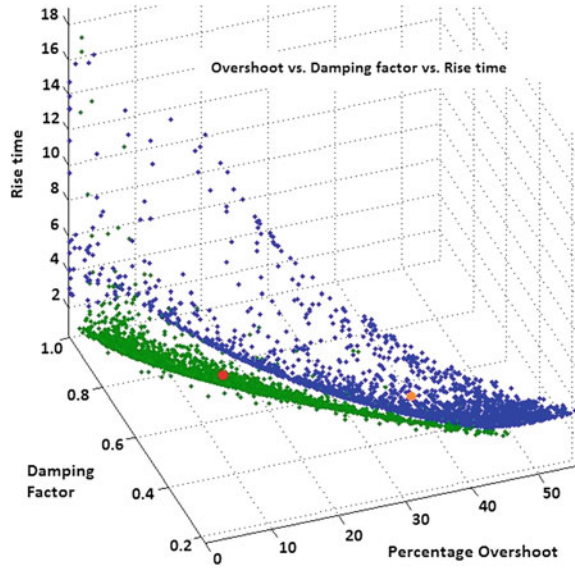
When comparing two Pareto-optimal fronts, one of three situations will occur:

1. one front dominates all the points of the other front;
2. some points of front A are dominated by point from front B while other points in A dominate points in B; and
3. no points of either front dominate each other because they are disjoint.

In the first case, the conclusion is simple—the one Pareto-optimal front is better than the other and therefore one controller is better than the other controller. In the second case, it is of interest to find in which situations A dominates B and vice versa. The third case simply means that certain areas of the objective space can be reached by only one of the controllers and other areas can only be reached with the other controller. It is also interesting to characterize these areas.

The Pareto-optimal fronts for the PI and FPI controller were combined and only 0.436 % of the total number of points was dominated (i.e. 26 points out of 5968). Therefore the two fronts are largely disjoint. Since the two fronts are almost

Fig. 11.2 Centroids and 3D scatter plots of overshoot, damping factor and rise time for PI and FPI controllers



disjoint, classification techniques were used to reduce the number of points to a tractable number. The aim was to be able to classify a point as either a PI or FPI controller by comparing the distances to the centroids of each cluster (or Pareto-optimal front). Therefore if most of the points can be correctly classified then the centroids would be good representative points of their respective Pareto-optimal fronts. So by examining the centroids, we are effectively comparing the two fronts.

A nearest centroid classification [24] was done on the raw data but the percentage of misclassifications was 43.45 %. Figure 11.2 shows a plot of the overshoot, damping factor and rise time for each controller. The figure shows that the Pareto-optimal fronts are mostly disjoint. It can also be seen visually why the misclassification rate is so high. Since the classification uses a Euclidean metric, we can imagine a plane midway between the two centroids. Many of the points on the side closer to the one centroid should be classified with the other centroid, and vice versa.

To improve the classification, the following was performed. The two Pareto-optimal fronts were combined into a single data set and normalized to a mean of 0.0 and standard deviation of 1.0. The data were then orthogonalized and whitened [25] using singular value decomposition (SVD), i.e. if X is the normalized data then

$$X = U * S * V^T \quad (11.9)$$

and the matrix U was used in place of X . A nearest centroid classification was then performed on these new data and the percentage of misclassifications was 2.88 %. This shows a large improvement in the classification accuracy.

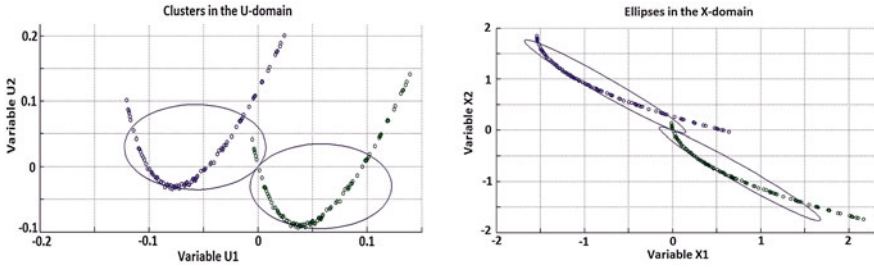


Fig. 11.3 Nearest centroid classification of the original Pareto-optimal front (*left*) and a nearest centroid classification after using SVD to orthogonalize and whiten the data (*right*)

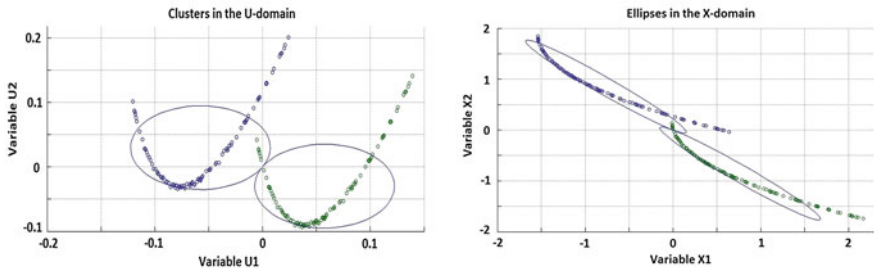


Fig. 11.4 Spherical clusters in the U-domain (*left*) map to elliptical clusters in the X-domain (*right*)

Figure 11.3 illustrates the principle of the classification using SVD. The figure on the left shows how the points would be classified if a nearest centroid was used on the raw data. The figure on the right shows the classification after normalization and SVD using nearest centroid.

The centroids of each cluster were then remapped back into the objective space. The values for each of the controllers are given in the first columns of Tables 11.2 and 11.3 respectively.

Another advantage to note about the use of SVD is that the map from U to X is a matrix multiplication. Since matrix multiplication can be geometrically interpreted as rotations and scaling we can get an idea of approximate shape of the Pareto-optimal front. For example if we consider clusters in the domain U as hyper-spheres then these spheres will be hyper-ellipsoids in the X domain (Fig. 11.4).

Lastly we can use the hyper-volume indicator to obtain an approximate measure of how much one controller is better than the other, i.e. if the controller has a higher hyper-volume then it either covers more of the objective space or is better at minimizing the objective functions (or both) than the other controller.

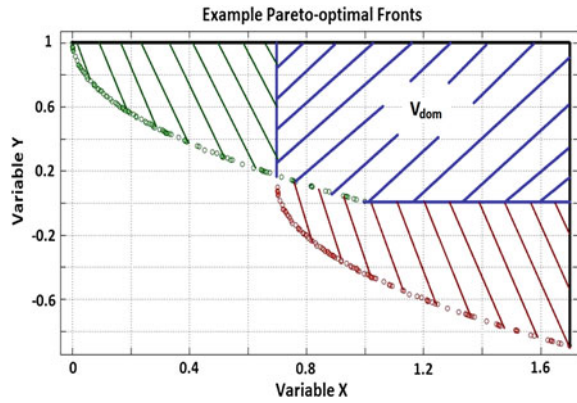
Table 11.2 PI controller centroid and the selected Pareto-optimal front points

| | Centroid | A | B | C |
|-----------|----------|------|-------|------|
| Overshoot | 35.6 | 34.2 | 35.1 | 32.1 |
| Damping | 0.39 | 0.33 | 0.37 | 0.38 |
| Rise time | 4.76 | 4.73 | 2.48 | 3.47 |
| Bandwidth | 0.726 | 0.40 | 0.726 | 0.53 |

Table 11.3 FPI controller centroid and the selected Pareto-optimal front points

| | Centroid | A | B |
|-----------|----------|-------|-------|
| Overshoot | 13.45 | 13.38 | 0.00 |
| Damping | 0.634 | 0.631 | 0.776 |
| Rise time | 48.5 | 2.01 | 48.2 |
| Bandwidth | 0.946 | 0.957 | 0.072 |

Fig. 11.5 Hyper-volume of the dominated points and the hyper-volumes of the original Pareto-optimal fronts subtract V_{dom}



The hyper-volume V_C was calculated for the PI controller, the hyper-volume V_F was calculated for the FPI controller and the points that were dominated when the two Pareto-optimal fronts were combined V_{dom} (see Fig. 11.5 for a 2D example). So,

$$V_{\text{dom}} = V_C \cap V_F \quad (11.10)$$

The HypE algorithm by Bader and Zitzler [15] was used on the normalized data to calculate the indicators. It was found that the hyper-volume of the PI and FPI controllers were approximately the same:

$$V_C - V_{\text{dom}} = 3010$$

$$V_F - V_{\text{dom}} = 3068$$

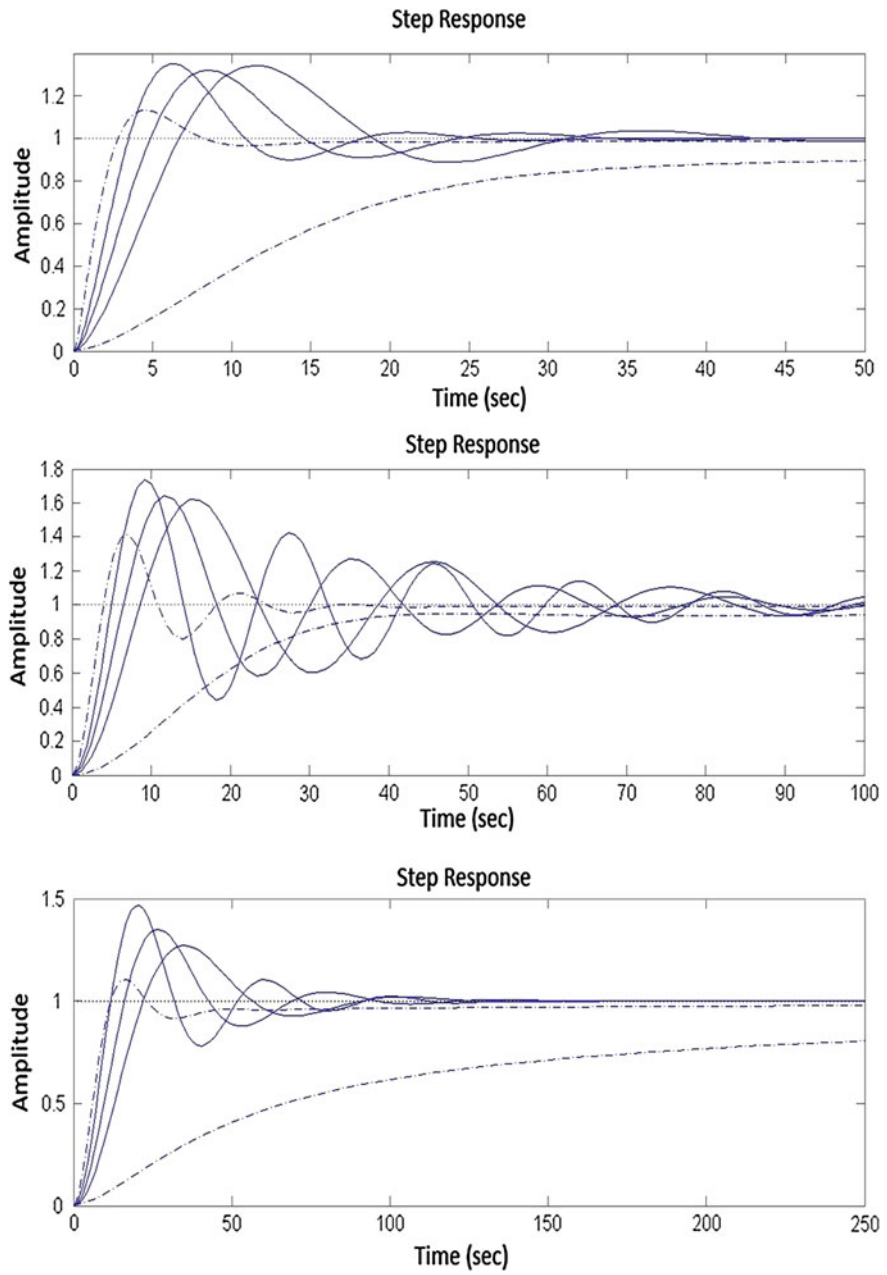


Fig. 11.6 Closed loop step responses for the selected PI controllers (*solid lines*) and selected FPI controllers (*dashed line*) for the worst case parameters for the plant

11.6 Results and Discussion

The centroid for the PI controller cannot be practically implemented; therefore three points were chosen that approximate the centroid as closely possible. The first point (A) was chosen by ignoring the value of the bandwidth, the second (B) was chosen by ignoring the rise time and the third (C) was a compromise between the bandwidth and rise time. The value of the parameters is shown in Table 11.2.

The centroid for the FPI controller also cannot be implemented because of the unreasonable rise time (see Table 11.3). Therefore two points were chosen: the first (A) was as close to the centroid as possible (when ignoring the rise time) and the other (B) was chosen because the rise time was the most similar to the centroid's rise time.

The closed loop step response for the PI and FPI controllers was then simulated for the worst-case values of the plant.

Comparing the centroids we can see that the FPI controller gives less overshoot, more damping and higher bandwidth but a slower rise time. Figure 11.6 shows the closed loop step responses of the controllers given in Tables 11.2 and 11.3. These simulations correspond to the predicted responses of the controllers.

A comparison of the Pareto-optimal fronts revealed that the fronts are almost disjoint and therefore each controller covers a different region in the objective space. Furthermore, using a hyper-volume indicator we can see that the space covered by each of the controllers is very similar. However, it should be noted that the hyper-volume indicator is dependent on the scale of each objective function. Therefore the data were normalized to reduce this effect.

11.7 Conclusion

A comparison between PI and FPI controllers has been made using Pareto-optimal fronts. This analysis was based on the paper by Feliu-Batlle et al. [2].

The original Pareto-optimal fronts consisted of twelve dimensions. However, using population-based incremental learning, almost 90 % of the fronts could be retained when considering only four factors: maximum percentage overshoot, damping factor, closed loop rise time and the closed loop bandwidth.

Using the 4D Pareto-optimal fronts, a classification method was used to find the centroids of the fronts. The centroids showed that, in general, FPI controllers have less overshoot, more damping and less bandwidth, but a higher rise time.

The controllers cover different regions in the objective space and using the hyper-volume indicator, it was shown that the size of these regions were approximately equal.

Therefore, 90 % of PI controllers are neither inferior nor superior to FPI controllers but rather the decision to use a PI or FPI controller depends on the problem.

References

1. Monje CA, Chen YQ, Vinagre BM, Xue D, Feliu V (2010) Fractional-order systems and controls: fundamentals and applications, *Advances in Industrial Control*. Springer, Berlin
2. Feliu-Batlle V, Rivas-Perez R, Sanchez-Rodriguez L, Riuz-Torija MA (2009) Robust fractional-order PI controller implemented on a laboratory hydraulic canal. *J Hydraul Eng* 135(4):271–282
3. Andersson J (2000) A survey of multiobjective optimization in engineering design. Technical Report LiTH-IKP-R-1097, Dept. of Mech. Eng. Linköping University, Sweden
4. Liu GP, Yang JB, Whidborne JF (2002) Multiobjective optimisation and control. Research Studies Press, Baldock
5. Moore D (2009) Optimal controller comparison. In: *Proceedings of IETA*, pp. 209–214
6. Ho KW (2010) Pareto front investigation of multivariable control systems. In: *International Joint Conferences on Computer, Information, and Systems Sciences and Engineering (CISSE) conference 2010*
7. Zhao C, Xue D, Chen YQ (2005) A fractional order PID tuning algorithm for a class of fractional order plants. In: *Proceedings of the IEEE International Conference on Mechatronics and Automation*. 2005
8. Miloš S, Martin C (2006) The fractional-order PID controller outperforms the classical one. In: *7th International Scientific Technical Conference, Process Control*, 2006
9. Zitzler E, Thiele L (1999) Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE Transac Evol Comput* 3(4):257–271
10. Zitzler E, Thiele L, Bader J (2010) On set-based multiobjective optimization. *IEEE Transac Evol Comput* 14(1):58–79
11. Coello CAC (2000) An updated survey of GA-based multiobjective optimization techniques. *ACM Comput Surv* 32(2):109–143
12. Huang VL, Suganthan PN, Liang JJ (2006) Comprehensive learning particle swarm optimizer for solving multiobjective optimization problems. *Int J Intell Syst* 21:209–226
13. Zitzler E, Brockhoff D, Thiele L (2007) The hypervolume indicator revisited: on the design of Pareto-compliant indicators via weighted integration. In: *Proceedings of the 4th International Conference on Evolutionary Multi-Criterion Optimization*, March 05–08, 2007, Matsushima, Japan
14. Bader J, Zitzler E (2011) HypE: An algorithm for fast hypervolume-based many-objective optimization. *Evol Comput* 19(1):45–76
15. Brockhoff D, Zitzler E (2007) Offline and online objective reduction in evolutionary multiobjective optimization based on objective conflicts. TIK Report 269, Institut für Technische Informatik und Kommunikationsnetze, ETH Zürich, April 2007
16. Seborg DE, Edgar TF, Mellichamp DA, Doyle FJ (2010) *III Process dynamics and control*. Wiley 3rd edn
17. Huang VL, Zhao SZ, Mallipeddi R, Suganthan PN (2009) Multi-objective optimization using self-adaptive differential evolution algorithm. In: *Proceedings of the 11th conference on Congress on Evolutionary Computation*, pp. 190–194, Trondheim, Norway
18. Wang Y, Wu L, Yuan X (2009) Multi-objective self-adaptive differential evolution with elitist archive and crowding entropy-based diversity measure. *Soft Computing—A Fusion of Foundations, Methodologies and Applications* 14(3):193–209, October 2009
19. Oustaloup A, Levron F, Mathieu B, Nanot FM (2000) Frequency-band complex noninteger differentiator: Characterization and synthesis. *IEEE Transac Circ Syst—I: Fundamental Theory and Applications*, 47(1):25–39
20. Xue D, Zhao C, Chen Y (2006) A modified approximation method of fractional order system. In: *Proceedings of the 2006 IEEE International Conference on Mechatronics and Automation*, 2006

21. Brockhoff D, Zitzler E (2007) Objective reduction in multiobjective optimization: The minimum objective subset problem. In: Waldmann KH, Stocker UM (eds). In: Operations Research Proceedings 2006, pp. 423–429, Springer, Berlin
22. Brockhoff D, Zitzler E (2009) Objective reduction in evolutionary multiobjective optimization: Theory and applications. *Evol Comput* 17(2):135–166
23. Cano JR, Herrera F, Lozano M (2003) Using evolutionary algorithms as instance selection for data reduction in KDD: an experimental study. *IEEE Transac Evol Comput* 7(6):561–575
24. Duda RO, Hart PE (2000) *Pattern Classification*, 2nd ed., Wiley-Interscience, Hoboken
25. Comon P, Jutten C (2010) *Handbook of blind source separation: independent component analysis and applications*, Academic Press, Oxford

Chapter 12

Remote Sensing Investigation of Red Mud Catastrophe and Results of Image Processing Assessment

J. Berke, V. Kozma-Bognár, P. Burai, L. D. Kováts, T. Tomor
and T. Németh

Abstract Data collection with the help of remote sensing is significant in the field of gathering information about the environment quickly and effectively. In case of the red sludge disaster in Ajka, Hungary Oct. 4, 2010 our group of researchers carried out an extensive remote sensing data collection with the co-ordination of Science Council of the Committee of the Government Coordination Commission, applying the most up-to-date technologies of remote sensing and data processing. In this publication besides the main points in planning and executing air shots we also summarize the results of data processing with image analysis during the evaluation of the catastrophe and the compensation period.

12.1 Introduction

Hungary's largest ecological disaster took place on October 4, 2010 at 1:30 p.m. when the western dam of cassette X of the sludge reservoir, belonging to a privately owned company, Magyar Alumínium Zrt (Hungarian Aluminum Co.),

J. Berke (✉) · V. Kozma-Bognár · P. Burai · T. Tomor
Róbert Károly College, Gyöngyös, Mátrai út. 36, 3200 Hungary
e-mail: berke64@gmail.com

L. D. Kováts
Iridium Measurement Technology Ltd, Budapest, Váci út 51/b, 1134 Hungary

T. Németh
Hungarian Academy of Sciences, Budapest, Széchenyi István tér 9, 1051 Hungary

J. Berke
Dennis Gabor College, Budapest, Mézők út. 39, 1139 Hungary

V. Kozma-Bognár
Pannon University, Georgikon Faculty, Keszthely, Festetics út 7, 8360 Hungary

had ruptured. Due to the ruptured dam, a mixture of 1,600,000 m³ (our calculated value) of red sludge and water inundated the lower sections of the settlements of Kolontár, Devecser and Somlóvásárhely via the Torna creek. The spilling red sludge flooded 800 ha of surrounding areas. The most extreme devastation was caused in the villages of Devecser and Kolontár, which are located near the reservoir [9].

Our group of researchers carried out an extensive remote sensing data collection with the co-ordination of Science Council of the Committee of the Government Coordination Commission, applying the most up-to-date technologies of remote sensing and data processing.

Besides remote sensing with the help of visible (VIS) devices, studies applying Near Infrared (NIR), Far Infrared (FIR) cameras, and hyperspectral (HYS) devices are getting more and more common [5, 6]. Applying FIR cameras (with a band of 8–14 μm) has become a general routine in the remote measurement of temperature in case of objects, based on thermal radiation [1, 7]. There are frequent inspections for identification of faulty spots with the help of devices of different manufacturers, mainly in the fields of electric network, machines and health care. The use of FIR cameras for martial and provost purposes is also remarkable. In Hungary there are investigations based on not only VIS, NIR and FIR recordings but in the common fields. For aeronautic and on-the-spot measurements we have developed, tested and applied devices with high resolution and an ability to record multi-spectral/hyper-spectral image data. We have worked out unique diagnostic and data processing methods to make the optimal integration of images deriving from different methods and sensors possible. This complex knowledge allows us to evaluate and reconstruct events having happened genuinely and also plays a great role in prevention. With the help of control studies disasters could be avoided.

12.2 Purpose of Data Collection

The main purpose of remote sensing recording was to monitor the investigation of environmental damage, to locate the area of contamination precisely, to estimate the concentration of substances in the mud and to estimate the status of the flooded area. We also aimed to provide geodetic data acquired with the help of remote sensing, necessary for high-resolution visual information gained from specific spectral ranges and for the realistic post-modelling of the event. Our further goal was to provide information helping to reveal the causes leading to the tear of the dyke and to make proposals on further studies to be carried out.

Table 12.1 Main parameters of data collection with the help of air-shot

| | Visible data (VIS) | | Near Infrared data (NIR) | Far Infrared data (FIR) | Hyper-spectral data | LiDAR data |
|------------------------|----------------------------|----------------------------|------------------------------------|----------------------------|----------------------------|------------|
| Type of sensor | Canon CMOS | Canon CMOS IR | Hexium Infra Diagnostic System 110 | AISA Eagle | Leica ALS 60 | |
| Time of flight | 2010.10.11. 14:00–18:00 | 2010.10.11. 14:00–18:00 | 2010.10.11. 14:00–18:00 | 2010.10.10. 11:00–14:30 | 2010.10.11. 12:00–14:00 | |
| Height of flight (m) | 350, 1,000 | 350, 1,000 | 350, 1,000 | 1,650 | 800 | |
| Spectral band | 400–700 nm | 720–1,150 nm | 8,000–14,000 nm | 400–970 nm (253 bands) | – | |
| Geometrical resolution | 0, 2 m | 0, 2 m | 0, 6 m | 1, 1 m | 4 Pts/m ² | |
| Data recording | 14/16 bit/pixel | 14/16 bit/pixel | 14/16 bit/pixel | 14/16 bit/pixel | 16 bit | |

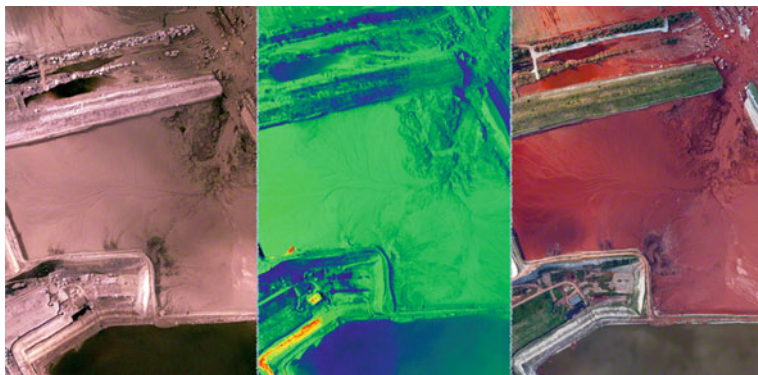


Fig. 12.1 Main type of images of data collection with the help of air-shot—Near Infrared (*left*), Far Infrared (*middle*) and Visible (*right*)

12.3 Data Recording

The peculiarity of this event needed unique planning and implementation. Data collection with the help of remote sensing—according to the complex aims—was carried out by a number of experts participating the project. Besides, we paid special attention to on-the-spot data collection at the time of air-shots, to the opportunity of processing the different remote-sensing technologies and to the collection of data corresponding both short and long term processing goals. Table 12.1 and Fig. 12.1 shows basic data about the images gathered during our flights and air-shots.

12.4 Data Processing

Before the assessment of images the necessary pre-processing tasks had been carried out: synchronization in time and place, filtration of optical and sensor-made noise, radiometric and geometric correction. During the processing of data we partly used our innovative programs (Spectral Fractal Dimension (SFD) based processing [2, 3] processing of FIR images [7]), and also special GIS and image processing software (ITT ENVI, Specim CaliGeo, Erdas Imagine, ESRI ArcGIS).

12.5 Results

The evaluation of data provided by different remote sensing technologies was carried out in parallel, in consideration of the seriousness of the situation, applying unique and integrated data processing methods. With the help of our own method based on the fractal system we identified the noisy bands, the optimal image bands

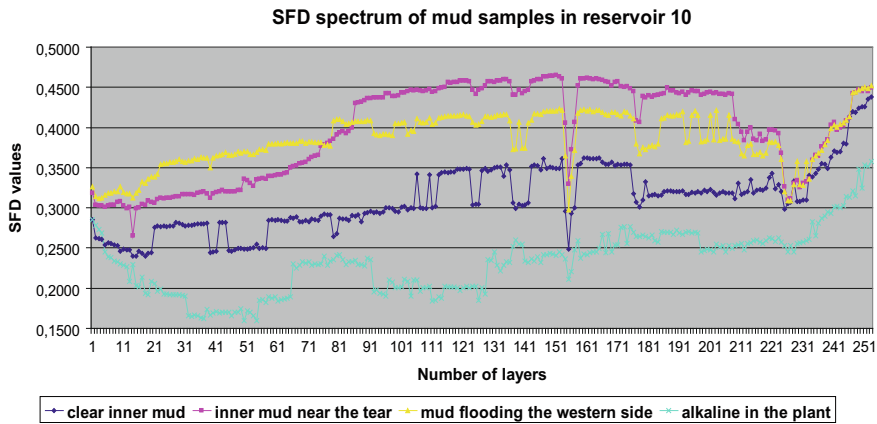


Fig. 12.2 ‘Spectral fractal fingerprints’ of mud samples in the plant/Spectral Fractal Dimension values based on hyperspectral image data

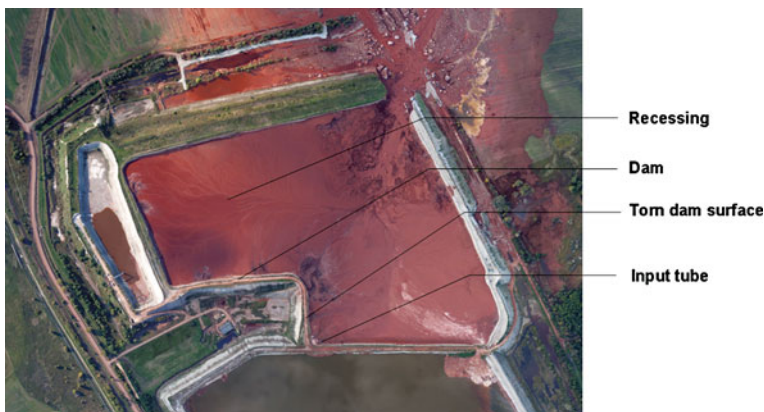


Fig. 12.3 Localization of cracks, breaks and slivering

from among the hyperspectral, VIS, NIR and FIR images [8] and also the muddy areas based on Spectral Fractal Dimension curves [3, 4], (Fig. 12.2).

In order to compute Spectral Fractal Dimension (more than two image layers or bands and equal to spectral resolution), the definition of spectral fractal dimension can be applied to the measured data like a function (number of valuable spectral boxes in proportion to the whole number of boxes), computing with simple mathematical average as follows [3, 4]:

$$SFD_{measured} = \frac{n \times \sum_{j=1}^{S-1} \frac{\log(BM_j)}{\log(BT_j)}}{S - 1} \quad (12.1)$$

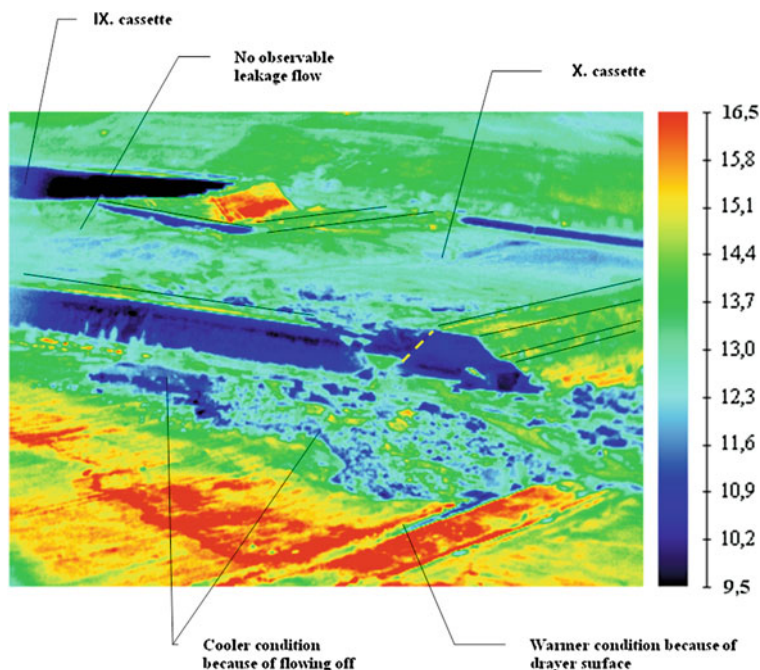


Fig. 12.4 Identification of wet and dry areas near the reservoir in lateral's thermo-shot

where

- n —number of image layers or bands.
- S —spectral resolution of the layer, in bits.
- BM_j —number of spectral boxes containing valuable pixels in case of j -bits.
- BT_j —total number of possible spectral boxes in case of j -bits.

The number of possible spectral boxes (BT_j) in case of j -bits as follows:

$$BT_j = (2^S)^n \quad (12.2)$$

Based on the air-shots we have defined the movements involving the Northern wall of the dyke together with their causes and we have localized the cut-off points and the slivering of the wall (Fig. 12.3).

We have identified the wet, leaking areas in the images shot near the reservoir. During the lateral examination of the wall of the dyke we studied clues alluding to cracks and leaks (Fig. 12.4).

Considering the reference surfaces we have managed to define the elevation of mud in the reservoir before the tear of the dyke and also the amount of mud flooding out of the reservoir. The reference level of the sludge in the reservoir before the tear was defined with the help of data gained with stereo evaluation of air-shots in September 2010. The area after the tear of the dyke was given by a

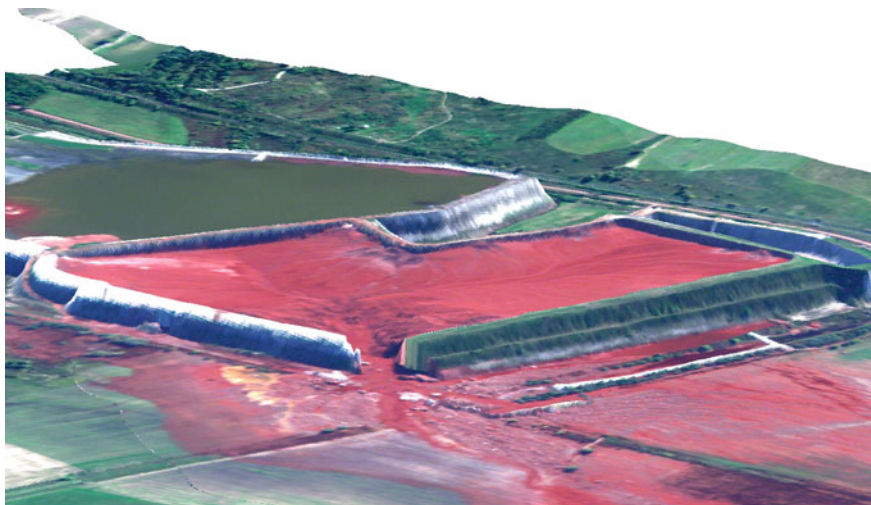


Fig. 12.5 The LIDAR image used for the definition of the amount of mud having flooded out of the reservoir

measurement with LIDAR technology (Fig. 12.5). The volume between the two areas was defined by their intersection. The 'reconstruction' of the broken piece of wall was also made based on the shots above. We have also defined the cross-section and longitudinal section of the reservoir.

After the geometric and radiometric correction of hyperspectral images we classified the area. We applied the Spectral Angle Mapper (SAM) method for the impoundment of the flooded area based on spectra on-the-spot.

We applied different classes (end members) according to the surface cover. The wet phase, with dry matter content less than 30 %, was defined separately. When applying the SAM, the angle value was optimized based on the control areas. After the area impoundment we defined the spectra correlating with the thickness of mud (in case of a mud with 30–70 % by mass) with the help of regression analysis, based on on-the-spot samples. Based on the Red Mud Layer Index (RMLI) calculation from the 550 nm and the 682 nm band cassette.

$$RMLI = \frac{B682\text{ nm} - B549\text{ nm}}{B682\text{ nm} + B549\text{ nm}} \quad (12.3)$$

we defined the threshold values for four thickness categories of previously marked areas (Fig. 12.6).

Based on data from the assessment further analysis and modelling processes have been carried out: a simulation of the tear of the dyke, suffusion intensity calculation and spread modelling. Based on our analysis operational steps have been taken by the authorities: planning and building of embankment, planning of prevention and preparation of compensation.

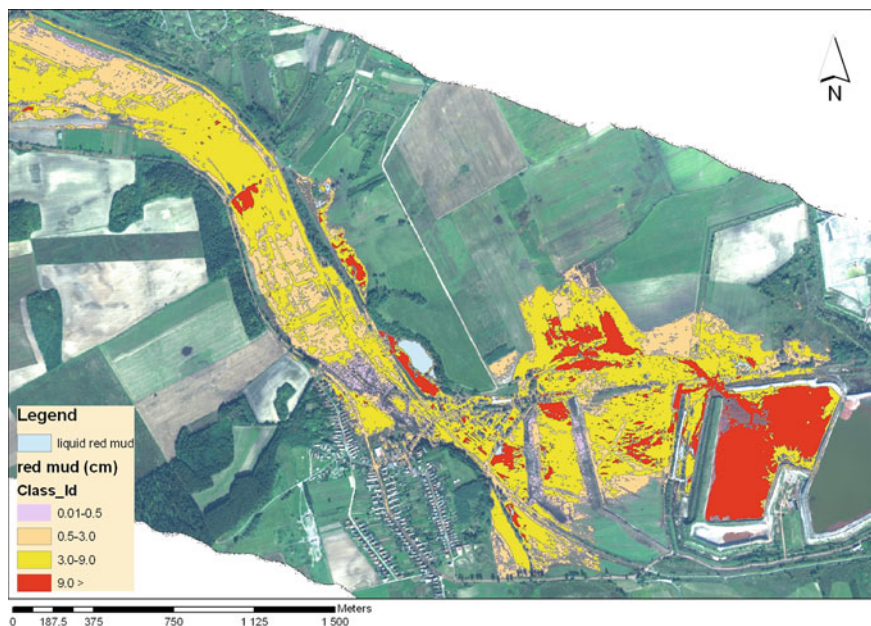


Fig. 12.6 Hyperspectral mosaic (RGB) and areas covered by layers of different thickness

Scientific duties have been assigned in case of a catastrophe and methodological protocols have been defined.

References

1. Anda A (1993) Surface Temperature as an Important Parameter of Plant Stand. *Időjárás*. 97(4):259–269
2. Authors internet site of parameter spectral fractal dimension. <http://www.digkep.hu/sfd/index.htm>
3. Berke J (2007) Measuring of spectral fractal dimension. *J New Math Nat Comput* 3/3:409–418. ISSN: 1793-0057
4. Berke J (2010) Using spectral fractal dimension in image classification Innovations and advances in computer sciences and engineering. Springer Science + Business Media B.V. 2010. DOI: [10.1007/978-90-481-3658-2_41](https://doi.org/10.1007/978-90-481-3658-2_41)
5. Chi M, Feng R, Bruzzone L (2008) Classification of hyperspectral remote-sensing data with primal SVM for small-sized training dataset problem. *Adv Space Res* 41(11):1793–1799
6. Frank M, Pan Z, Raber B, Lenart CS (2010) Vegetation management of utility corridors using high-resolution hyperspectral imaging and lidar. 2nd IEEE GRSS workshop on hyperspectral image and signal processing-WHISPERS'2010. Reykjavik, 2010
7. Kováts LD (1998) Infra television aided maintenance diagnostic. In: Proceedings of the condition monitoring and diagnostic engineering management (comadem) international congress. Monash University, Melbourne, 1998. pp 527–532
8. Kozma-Bognár V, Berke J (2010) New evaluation techniques of hyperspectral data. *J Syst Cybern Inform* 8(5):1690–4524 ISSN: 1690-4524, <http://www.iiisci.org/journal/SCI/>
9. Official website of the Hungarian government “Redsludge” tragedy. <http://redsludge.bm.hu>

Chapter 13

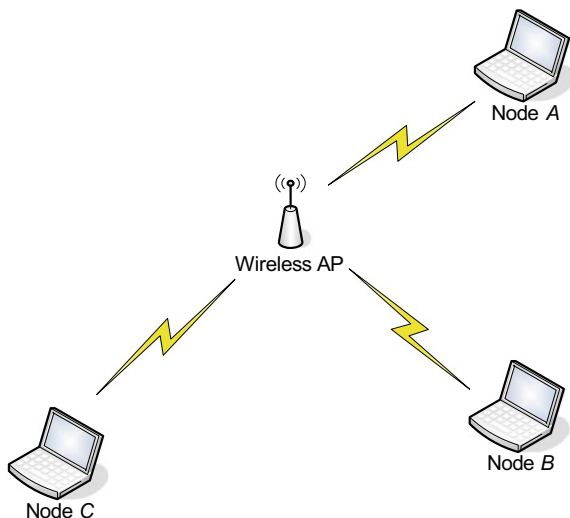
802.11e QoS Performance Evaluation

Yunus Simsek and Hetal Jasani

Abstract Wireless networks have become a tremendous network solution for home and enterprise users, without worrying about wire, and mobility. With IEEE 802.11e Wireless Local Area Networks (WLAN), we have implemented Quality of Service (QoS) to our wireless network that provides significant improvements for high-priority QoS traffic. However, these improvements have a negative performance impact to lower-priority traffic, such as HTTP, FTP etc. In this paper, we examine two IEEE 802.11e QoS functions which are Point Coordination Function (PCF), and Hybrid Coordination Function (HCF). We also examine their negative effects on performance of lower-priority traffic in two different WLAN infrastructures which are Basic Service Set (BSS) and Extended Service Set (ESS), using OPNET Modeler software. We evaluate the impact of high-priority traffic (with QoS enabled) on low-priority-traffic when they use the same access point. All of the simulation results proved that it has a significant detrimental impact to low-priority-traffic. Performance of HTTP, FTP, and database traffic drops when VoIP and video applications are using same access point.

Y. Simsek (✉) · H. Jasani
Department of Computer Science, Northern Kentucky University,
Highland Heights, KY 41099, USA
e-mail: info@yunussimsek.net

H. Jasani
e-mail: Jasanih1@nku.edu

Fig. 13.1 Basic service set

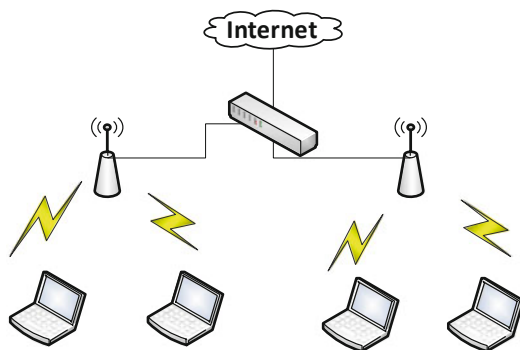
13.1 Introduction

Nowadays, with the growing needs of mobility and demands of accessing information anywhere, anytime with the simplest setup, a significant decrease in price of wireless equipment, and no-cabling need, brought Wireless Local Area Network (WLAN) one step closer to becoming a dominant real-world solution for either business or home users. The high demand of mobility brought a quality service problem with it, which can cause a lower performance on critical delay sensitive applications, such as Voice over Wireless LAN and streaming videos. The IEEE 802.11e is a validated improvement to the IEEE 802.11 standard that defines a set of Quality of Service amendment for WLAN applications through modifications to the Media Access Control (MAC) layer. This standard is mainly considered for significantly important delay-sensitive applications, such as Voice over Wireless LAN and streaming multimedia [1]. In this section, we introduce the BSS, ESS, DCF, PCF, and HCF.

13.1.1 Basic Service Set

Basic Service Set (BSS) is the basic function of the IEEE 802.11 WLAN defined in the IEEE 802.11-1999 standard. The BSS refers to the group of networking stations communicating with one to another by using single access point (AP). In BSS structure, an AP acts as a master controller to the networking stations [2] (Fig. 13.1).

Fig. 13.2 Extended service set



13.1.2 Extended Service Set

BSS cannot support mobility and roaming. Extended Service Set (ESS) is a set of two or more APs that works in same network. It is a combination of Basic Service Set (BSS) that form a single network [3]. Since ESS is using multiple APs, it can be used for many users over a wider area. APs are positioned such a way that it facilitates roaming (Fig. 13.2).

13.1.3 Distributed Coordination Function

To share the medium between multiple stations, some channel access method is needed. The basic 802.11 MAC layer uses the distributed coordination function (DCF) as channel access method. DCF relies on Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) in order to perform its duty [4].

13.1.4 Point Coordination Function

Point Coordination Function (PCF) is a Media Access Control (MAC) procedure that is used in IEEE 802.11 WLANs [5]. The PCF option has been implemented in IEEE 802.11 in order to provide contention-free data transmission and to support time-bounded services as well as data, voice or mixed [6]. In PCF, the AP takes point coordinator role, and it controls the medium access in a poll-and-response manner. The time period during the PCF operates is called the contention-free period (CFP). Before CFP starts the AP works under DCF, but it causes to use of the priority inter-frame space (PIFS) to seize medium, and then sends out beacon packet in duration of the CFP to each station one at a time to give them a turn to send data. The AP is the coordinator. Although this allows for a better management of QoS, PCF does not define classes of traffic as is common with other QoS systems [7].

13.1.5 The IEEE 802.11e MAC Protocol

DCF works well enough for data transmission. However, it is not efficient for real time traffic to provide better QoS. PCF is better, but many wireless manufacturers have decided not to facilitate the optional PCF service in their equipment [8]. The IEEE 802.11e draft specifies some improvement to provide QoS over WLANs through a new coordination function: the hybrid coordination function (HCF). According to draft, there are two methods of channel access: HCF Controlled Channel Access (HCCA) and Enhanced Distributed Channel Access (EDCA). EDCA defines four different streams or traffic categories [9]. HCCA is a new form of PCF and is based on polling. EDCA is good for enterprise business settings and while HCCA is more appropriate for home environment [10].

13.2 Performance Evaluation

13.2.1 Quality of Service Impact on WLAN Performance

Even though QoS has significant improvement for high-priority-traffic, it also has high cost for low-priority-Traffic, which can cause a significant negative impact on HTTP/FTP based traffic. We prove this by running several simulations on two different infrastructures to see these impacts.

Before implementation of the new wireless network, it is crucially important to evaluate network infrastructure based on: the number of client, structure of building, and external factors, etc. There are few softwares available on the market that can offer real-time network simulation without highly cost of equipment and waste of time. OPNET is a networks modeler tool that was designed by OPNET Technologies Inc, and can provide a powerful network simulation to user [11]. In this paper, we use OPNET Modeler version 16.0. OPNET Modeler incorporates a detailed and accurate model of the physical channel and of the IEEE 802.11 MAC layer [8].

13.2.2 Performance Measurement Units

In this paper, we collect Network Delay, Network Load, Throughput and Media Access Delay.

- **Network Delay**—represent end to end delay of all the packets that received by the WLAN nodes and it forwards all the packets to the higher layer. When the access point enabled this delay includes medium access delay at the source MAC, reception of all the fragments individually, and transfers of the frames via access point.

- Load—the load indicates total bits submitted to wireless LAN layers.
- Network Throughput—the throughput is an average rate of successful message delivery over a physical or logical link or passes through a certain network node. It is typically measured in bits per second [12].
- Media Access Delay—Represent the global statistic for the total of queuing and contention delays of the data [13].
- Object Response Time (sec)—represent response time for each inlined object from the HTML (web) page.
- Page Response Time (sec)—represent time required to retrieve the entire page with all the contained inline objects [13].
- FTP Download Response Time (sec)—represent time elapsed between sending a request and receiving the response. It measured from the time a client sends a request to the server (FTP server) to the time it receives a response packet [10].
- Database Entry Response Time (sec)—represent time elapsed between sending a request and receiving the response packet [13].
- Database Query Response Time (sec)—represent time elapsed between sending a request and receiving the response packet [13].

13.3 Simulation Results

13.3.1 First Scenario (HCF/PCF Enabled on BSS Infrastructure) vs. (HCF/PCF Disable on BSS Infrastructure)

In order to perform our first scenario we have created 2 voices, 4 FTP, and 2 web nodes with 1 access point on 3 floors in the building. In this scenario, we examine two different QoS on the same infrastructure (BSS) by running 15 min simulation in OPNET Modeler. We collected the statistics of network delay, network load, throughput, and media access delay. In order to make appropriate conclusion we had to run simulation over 10 min, since longer traffic always has a big impact to network traffic. The results have been obtained in Overlaid Statistics and average mode and shown Figs. 13.3, 13.4, 13.5, 13.6 and 13.7.

As we can see when HCF/PCF are not enable (Blue Line) network delay is low, but when we enable HCF/PCF then it immediately starts to increase network delay. As we can see, the same affect occurs here also, when HCF/PCF disable (Blue Line) Network load is higher, when it is not Network Load is low. The same affect is almost seen here as well, HCF/PCF enable mode extremely affects delay. Even though we get really close results in this graph, HCF/PCF Enable mode (Red Line) still has impact to Throughput.

Fig. 13.3 BSS Network deployment map in OPNET

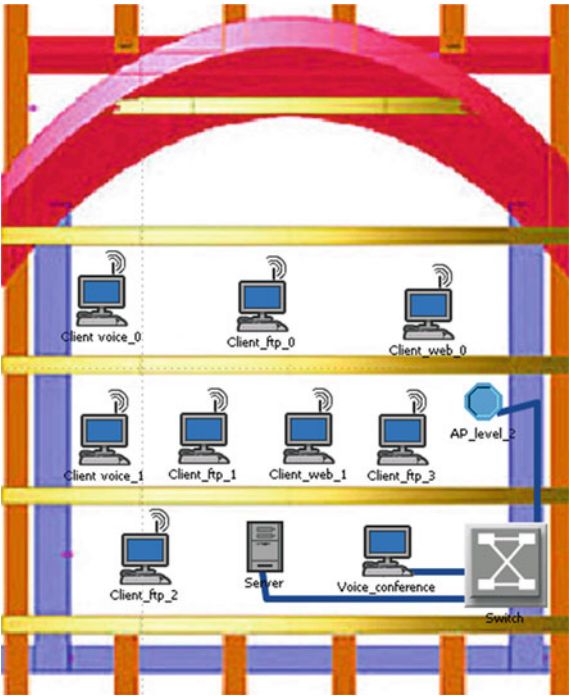


Fig. 13.4 Network delay

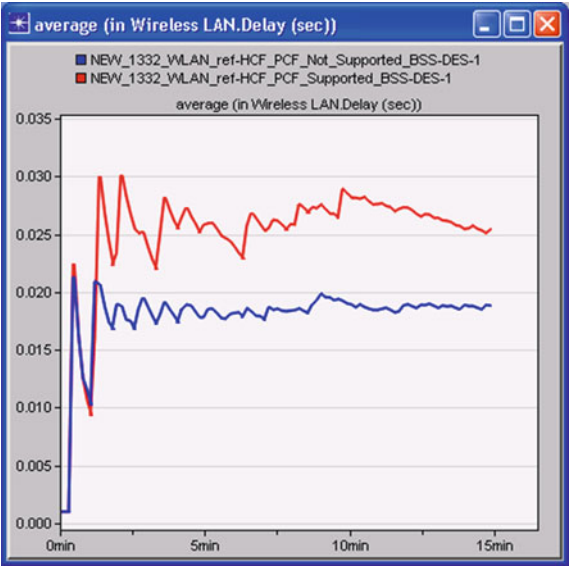


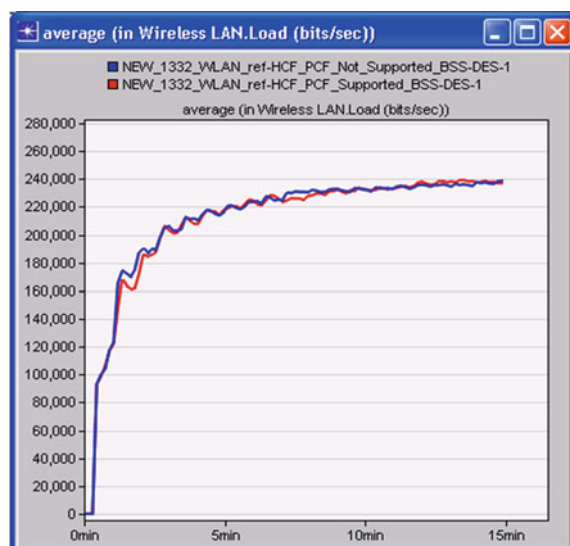
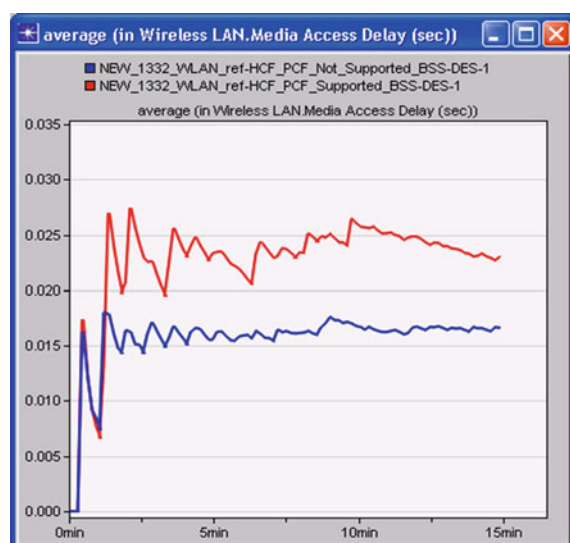
Fig. 13.5 Network load**Fig. 13.6** Media access delay

Fig. 13.7 Throughput

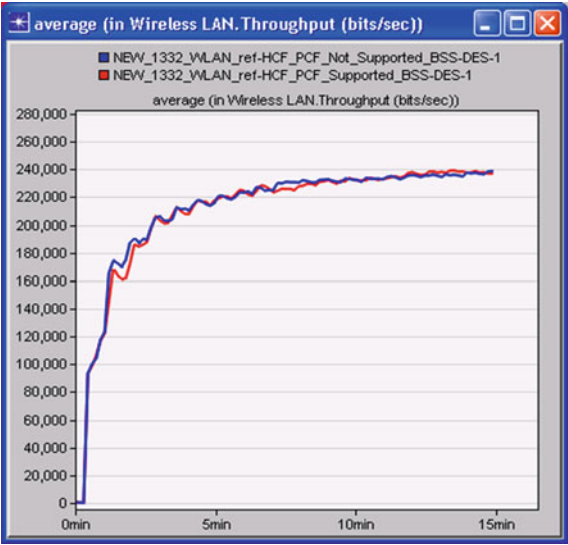
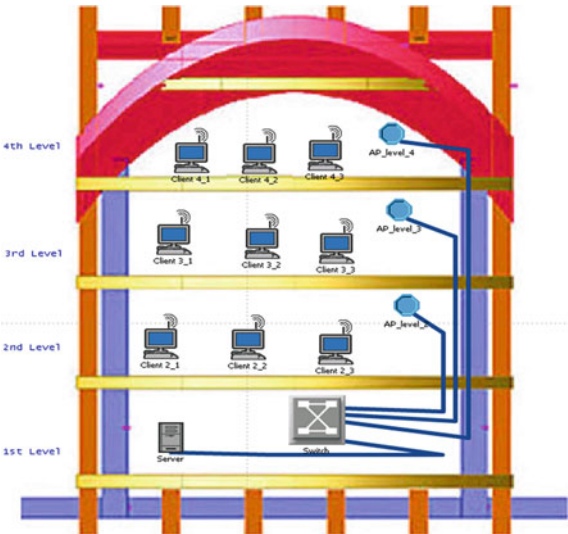


Fig. 13.8 ESS Network deployment map in OPNET



13.3.2 Second Scenario (HCF/PCF Enabled on ESS Infrastructure) vs (HCF/PCF Disable on ESS Infrastructure)

In order to perform our second scenario we have created 9 wireless nodes with 3 access points on 4 floors in the building as shown in Fig. 13.8. In this scenario, we

Fig. 13.9 Network delay

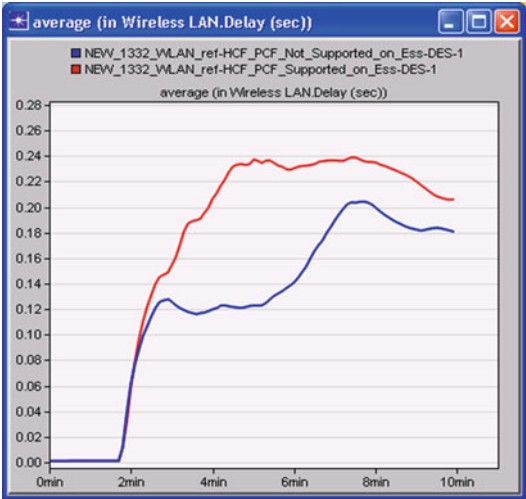
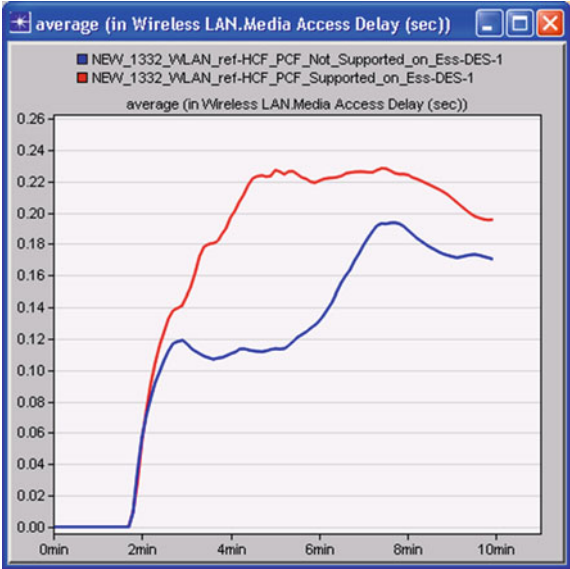


Fig. 13.10 Media access delay



examine two different QoS on the same infrastructure (ESS) by running 10 min simulation in OPNET Modeler.

After running 10 min simulation on OPNET Modeler, we have successfully collected Network Delay, Network Load, Throughput, and Media Access Delay statistics. The results have been obtained in Overlaid Statistics and average mode and shown Figs. 13.9, 13.10, 13.11, 13.12, and 13.13.

It seems in Fig. 13.8, when we enable QoS on Extended Set Service infrastructure it significantly started to impact Performance by increasing delay. It seems

Fig. 13.11 Network load

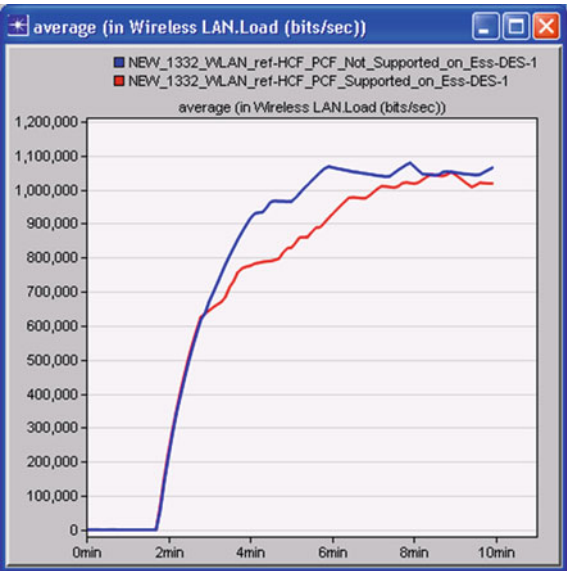


Fig. 13.12 Throughput

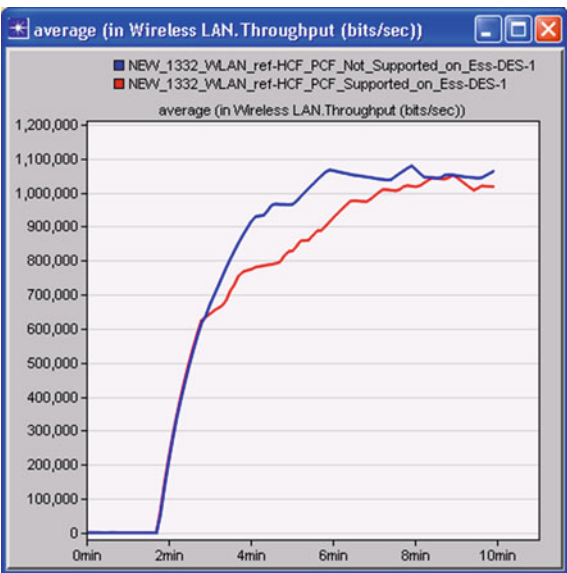
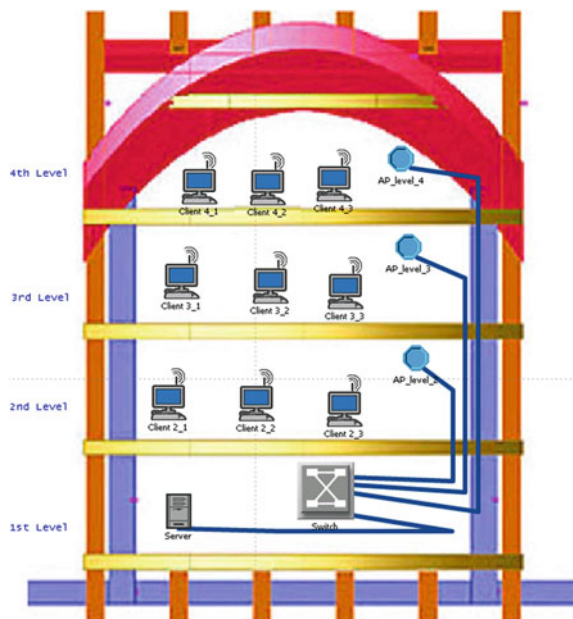


Fig. 13.9, enabling QoS also has negative impact on load by decreasing load performance on low-priority-traffic in Extended Set Service infrastructure. It is obvious that the QoS has a big impact on Media Access Load as we see in Fig. 13.10 We can ensure that the QoS also has a negative impact on Throughput.

Fig. 13.13 ESS Network deployment map in OPNET



13.3.3 Third Scenario Evaluating HTTP/FTP Services on PCF/HCF Enable Mode in ESS Infrastructure

In our third scenario, we examine how QoS service can affect low-priority traffic by running simulation on HTTP and FTP services in Extended Set Service (ESS) Infrastructures. In order to perform our third scenario we have created 9 wireless nodes with 3 access points on 4 floors in the building as shown in Fig. 13.12 and we set the OPNET simulator to run 10 min simulation in OPNET Modeler.

After running 10 min simulation on OPNET Modeler, we have successfully collected FTP Download Response time (sec), FTP Upload Response time (sec), HTTP object response time (sec) and HTTP page response time (sec) statistics. The results have been obtained in Overlaid Statistics and average ode and shown Figs. 13.14, 13.15, 13.16 and 13.17

Since FTP is low-priority-traffic, QoS has significant negative impact to FTP download response time as seen in Fig. 13.14 We have seen the same result FTP Upload Response time as well. QoS also has a negative impact on HTTP traffic too, as it is seen in Fig. 13.17 We can see the same impact HTTP page Response as well.

Fig. 13.14 FTP download response time

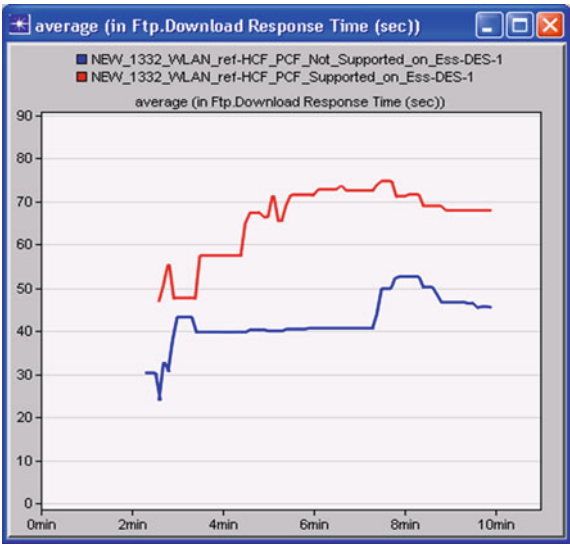
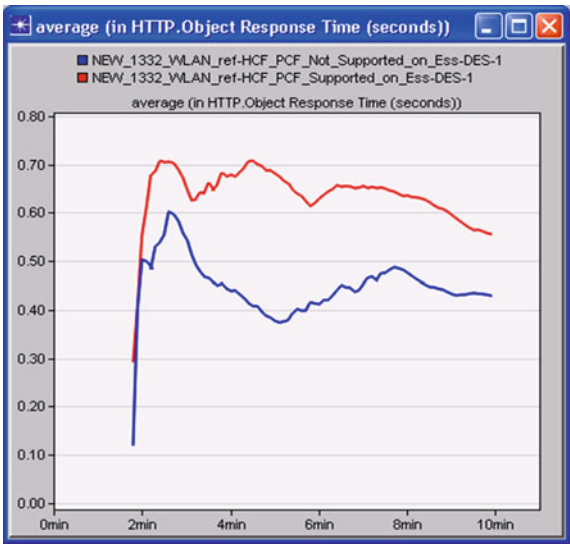


Fig. 13.15 HTTP object response time



13.3.4 Fourth Scenario Evaluating Database Performance on PCF/HCF Enable Mode in ESS Infrastructure

The growing demand of Mobility is the main reason that explains why Wireless technologies are everywhere in our life. The same reason is why a growing number of businesses are implementing 802.11 Wireless to their work environment. As we know, database is the main concern for either performance or security, and there

Fig. 13.16 FTP upload response time

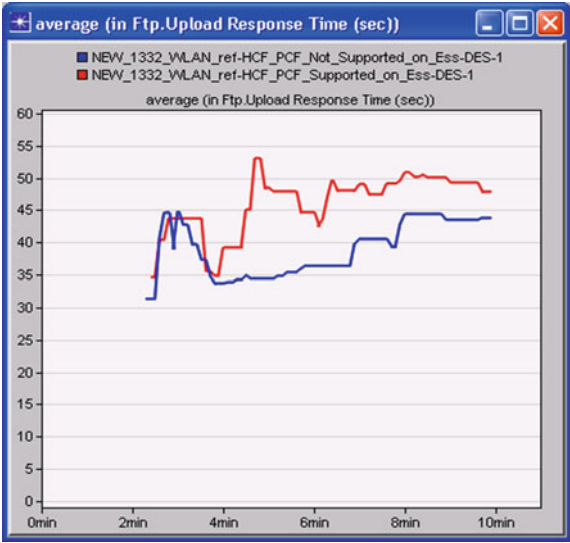
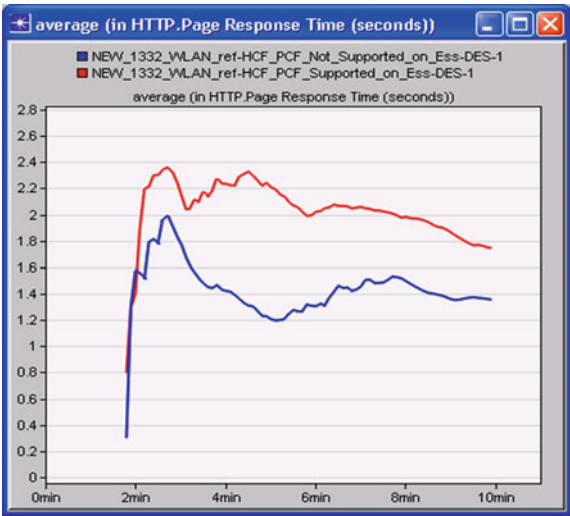


Fig. 13.17 HTTP page response time



are a variety of database softwares being used by most of the companies. In this scenario, we examine Database performance on 802.11e QoS enabled Wireless Local Area Network in Extended Set Service infrastructure.

After running 10 min simulation on OPNET Modeler, we have successfully collected Database Entry Response time (sec) and Database Query Response time (sec) statistics. Results have been obtained in Overlaid Statistics and average mode and shown Figs. 13.18 and 13.19.

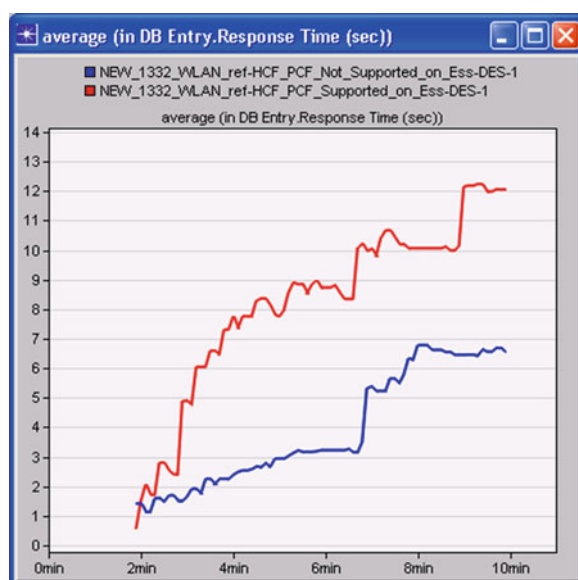


Fig. 13.18 Database entry response time

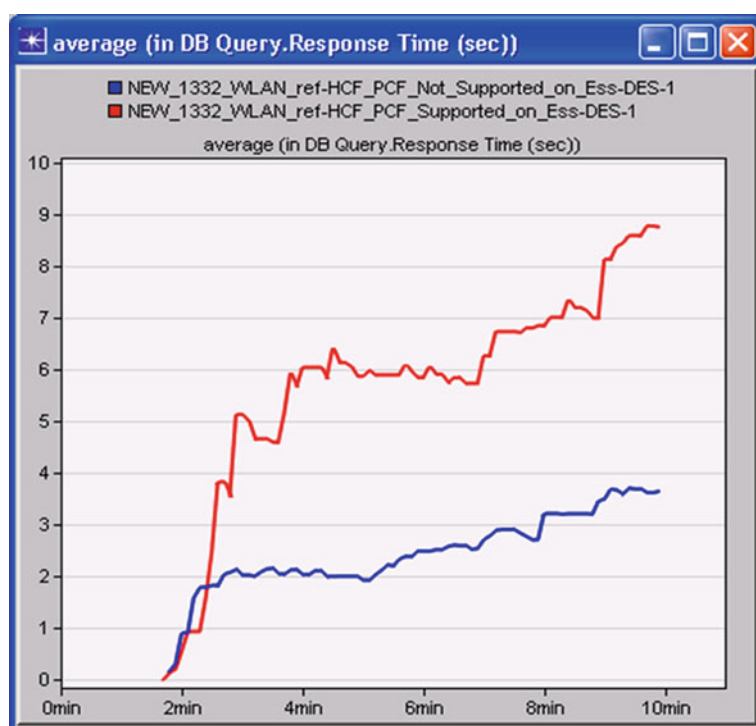


Fig. 13.19 Database query response time

As we see in Fig. 13.18, the QoS has the worst impact to Database Performance. QoS has a significant worse impact on Database query response time as well. If we consider business environments, there are thousands of queries executing every minute which can cause significant delay for them to get a response from the queries when we enable QoS on the same Wireless LAN AP.

13.4 Conclusion

We have evaluated Quality of Service (QoS) impact on low-priority-traffic when they use the same access point. With IEEE 802.11e, we have implemented QoS to our wireless network that provides significant improvements for high-priority QoS traffic. However, these improvements have a negative performance impact to lower-priority traffic, such as HTTP, FTP etc. We examined two IEEE 802.11e QoS functions which are Point Coordination Function (PCF), and Hybrid Coordination Function (HCF). We also examined their negative effects on performance of lower-priority traffic in two different Wireless Local Area Networks (WLAN) infrastructures which are Basic Service Set (BSS) and Extended Service Set (ESS), using OPNET Modeler. We evaluate the impact of high-priority traffic (with QoS enabled) on low-priority-traffic when they use the same access point. All of the simulation results proved that it has a significant detrimental impact to low-priority-traffic. Performance of HTTP, FTP, and database traffic drops when VoIP and video applications are using same access point. When we need to use VoIP or video streaming services, we should move them to another access point and let them work in Voice and Video streaming client with their access point in order to get better performance.

Acknowledgments The authors would like to acknowledge the National Science Foundation for offering a grant to apply toward the research. The authors would also like to give an extended acknowledgment to the OPNET team for allowing the use of their OPNET Modeler software for educational advancement at Northern Kentucky University.

References

1. Medium Access Control (MAC) Quality of Service Enhancements IEEE Computer Society (2005) Retrieved from <http://standards.ieee.org/getieee802/download/802.11e-2005.pdf>. Accessed 11 Nov 2005
2. Pham P (2005) Comprehensive analysis of the IEEE 802.11. *Mob Netw Appl* 10(5):691–703
3. Zhao D (2006) Distributed schemes for fair throughput in infrastructure-based IEEE 802.11 mesh networks. In: *Proceedings of the 3rd International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QShine '06)*, ACM, New York
4. Jacobson V, Smetters DK, Thornton JD, Plass MF, Briggs NH, Braynard RL (2009) Networking named content. In: *Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies (CoNEXT '09)*, ACM, New York, pp 1–12

5. Cecchetti G, Lina Ruscetti A (2008) Performance evaluation of real-time schedulers for HCCA function in IEEE 802.11e wireless networks. In: Proceedings of the 4th ACM Symposium on QoS and Security for Wireless and Mobile Networks (Q2SWinet '08), ACM, New York, 1–8
6. Anastasi G, Lenzini L (2000) QoS provided by the IEEE 802.11 wireless LAN to advanced data applications: a simulation analysis. *Wirel Netw* 6(2):99–108
7. Zhu H, Cao G (2004) On improving the performance of IEEE 802.11 with relay-enabled PCF. *Mob Netw Appl* 9(4):423–434
8. IEEE Computer Society (2007) IEEE standard for information technology—telecommunications and information exchange between systems—local and metropolitan area networks—specific requirements [Revision of IEEE Std 802.11-1999]. (PDF Version), Retrieved from <http://standards.ieee.org/getieee802/download/802.11-2007.pdf>
9. Xiao Y, Rosdahl J (2003) Performance analysis and enhancement for the current and future IEEE 802.11 MAC protocols. *SIGMOBILE Mob Comput Commun Rev* 7(2):6–19
10. Ciampa M (2006) CWNA guide to wireless LANs, 2nd edn
11. OPNET Technologies, Inc., Initials (n.d.) Network modeling | network simulation. Retrieved from http://www.opnet.com/solutions/network_rd/modeler.html
12. Kulkarni S, Prasad PS, Agrawal P (2008) Performance enhancement of mobile ad hoc networks using nodal cooperation. In: Proceedings of the 4th Annual International Conference on Wireless Internet (WICON '08), ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium 54:8
13. OPNET Technologies, Inc., Initials (2010) OPNET modeler documentation set [Version: 16.0] (Online Help Files)

Chapter 14

Evaluation of Different Designs to Represent Missing Information in SQL Databases

Erki Eessaar and Elari Saal

Abstract It is possible to use different designs to deal with the missing information problem in SQL databases. In this paper, we use a multi-criteria decision support method, which is based on the Analytic Hierarchy Process, to evaluate a set of possible designs for dealing with missing information. One of the designs uses SQL NULL-marks, another design uses tables that are in the sixth normal form, and one of the designs uses special values to represent missing information. We evaluate the designs, which are implemented in a PostgreSQLTM database, in terms of two hypothetical contexts. We create a test database, perform measurements, and use the results to compare the designs. The results of the evaluation provide new insights into the advantages and disadvantages of the different designs.

14.1 Introduction

It is possible that some information, which should be recorded in a database, is missing (is unknown). There are several reasons why the information could be missing. Therefore, one of the problems in the context of databases is how to deal with the missing information and capture the reason why the information is missing. In this paper, we concentrate our attention to SQL databases. SQL database language is *an implementation* of the relational data model. Entire tuples of a relation could be missing in case of relational databases. It is also possible that one knows only some, but not all, attribute values that are needed to form a tuple in order to record it in a relational database.

E. Eessaar (✉) · E. Saal
Department of Informatics, Tallinn University of Technology, Raja 15,
12618 Tallinn, Estonia
e-mail: eessaar@staff.ttu.ee

E. Saal
e-mail: elari.saal@mail.ee

SQL provides NULL-mark (NULL-flag) that can be used to denote missing information [1]. Date [2] stresses that NULL is not a value, although it is common to speak as if it is. If one compares a value with NULL or compares two NULLs, then the result would be the truth value “unknown” (SQL uses three-valued logic). The Third Manifesto [3], which is a formal proposal for a foundation of relational database management systems (DBMSs) stresses that NULLs should not be used to represent missing information. According to their definition tuples do not contain NULLs because NULLs are not values. Date [4] points to the practical problems that occur due to the use of NULLs.

How is it possible to reduce the need of using NULLs in SQL databases? Date and Darwen [5] propose to use decomposition approach, according to which one has to create tables that are in the sixth normal form in order to deal with the missing information problem. They suggest to create different tables to represent different kinds of missing information. Similarly, the Anchor Modeling modeling technique [6] suggests generation of specifications of tables that are in the sixth normal form and cite the absence of NULLs as one of the advantages of the approach. On the other hand, Ref. [6] does not suggest the creation of separate tables for representing missing information. Let us assume that an entity type ET (“anchor” in [6]) has an attribute a . Based on [6], if the value of a is missing in case of some entity with the type ET, then it means that the corresponding row is missing from the table that is created based on a .

Karwin [7] describes a practice according to which developers select ordinary values (that belong to the type of a column c) in order to represent unknown or inapplicable values in c . It makes it more difficult to write queries. Karwin [7] lists this practice as an antipattern. PredictiveDBTM [8] is a DBMS, which built-in features allow its users to predict missing values based on the existing values in the database. The system also calculates risk of being wrong.

Next, we shortly describe additional proposals for dealing with missing information in the relational databases. Codd [9] suggests to use A-marks (missing-but-applicable value mark) and I-marks (inapplicable-value mark) to represent missing information and distinguish between different reasons why information is missing. If an attribute a of a tuple t has A-mark, then it means that a can have value but the value is currently unknown to the users. If an attribute a of a tuple t has I-mark, then it means that a cannot have value in case of the object that is represented by the tuple t . Liu and Sunderraman [10] propose to use I-tables to represent indefinite and maybe facts. Date and Darwen [11] propose to determine special values in the type definitions in order to use these values for representing missing information. Date and Darwen [5] introduce approaches that use multirelations, inheritance of types, or relation valued attributes to deal with the missing information.

The goal of the paper is to evaluate three designs of SQL databases that use different means to deal with missing information. We use a multi-criteria decision support method [12], which is based on the Analytic Hierarchy Process (AHP) [13], to evaluate the designs.

The rest of the paper is organized as follows. Firstly, we explain the setup of the experiment by shortly describing the method that we use to evaluate the designs.

In addition, we introduce contexts, alternatives (designs), criteria, software measures that correspond to the criteria and will be used to evaluate the alternatives, and implementation of a test database. Secondly, we present the results of the evaluation of the designs and discuss the results. Finally, we conclude and point to the future work with the current topic.

14.2 Setup of the Experiment

14.2.1 Evaluation Method

The method, which we more thoroughly explain in [12], is based on the AHP and uses pairwise comparisons. One has to use a nine-point scale in case of the comparisons. AHP allows us to make decisions by using a hierarchical model. At the highest level of the model there is a goal. The goal might be to find the best database design from a set of given designs. At the second level there are objectives, which correspond to the criteria that one has to take into account while comparing alternatives (database designs). Each criterion can have sub-criteria. At the lowest level there are alternatives, between which the choice will be made. Suitability of a particular database design depends on the context where it will be used. Therefore, the method requires specification of the context where the designs will be used. Each context specifies a set of requirements and optionally a DBMS that will be used to implement the designs. Information about the context is used to compare criteria pairwise to find their relative importance (weights). The method suggests the use of criteria that have corresponding software measures. If there is a criterion c that has an associated software measure m , then one has to find the value of m in case of all the different alternatives. This information will be used to compare alternatives pairwise in terms of c . One has to perform the same task in case of all the alternatives in order to calculate the value of m in case of all the alternatives. Hence, one may have to implement one or more test database and generate test data to perform measurements in case of different designs. It is possible to find criteria based on quality models like the ISO 9126 quality model [14]. In addition, one could search suitable criteria by considering three database levels (external, conceptual, and internal) that are specified in the ANSI/SPARC database architecture [2]. One should use a set of criteria that cover all the database levels in order to thoroughly evaluate database designs. The use of explicitly specified contexts and software measures should improve the objectivity of the evaluation results.

14.2.2 Contexts

In this study, we evaluated designs (alternatives) in terms of two *hypothetical* contexts.

Context 1. It describes an online transaction processing system, which has quite stable set of requirements that will not change much over time. The system must answer queries within seconds and allow quick registration of data. The database is small in terms of the number of entities, the data of which will be recorded in it. Hence, the database size is not an important issue. The database will be implemented by using PostgreSQLTM DBMS [15].

Context 2. It describes a system for recording and maintaining the results of measurements. The system will evolve significantly over time and in the future there may be additional types of measurements, the resulting data of which the system has to manage. The system must answer queries within minutes. The database contains a lot of data and will grow considerably over the coming years. Hence, the database size is an important issue. The database will be implemented by using PostgreSQLTM DBMS [15].

14.2.3 Alternatives

For this experiment, we created a part, a simplified database of a radio station as the test database and considered three database designs as alternatives. The test database is used to record information about artists as well as their songs and albums. In addition, the test database is used to record information about the playlist of a radio station. We could use databases from different spheres of knowledge, influence, or activity as the test databases of this evaluation as long as we can use the selected software measures in case of these databases. We present the designs by using diagrams that are created by using the Rational Rose data modeling profile [16]. If a column of a base table (table in short) is annotated with “NN”, then it means that it is a mandatory column (it has NOT NULL constraint). In addition, all the columns that belong to the primary keys (annotated with “PK”) are also mandatory. If a column is not annotated with “NN” or “PK”, then it means that it is an optional column (it allows NULLs).

Figure 14.1 presents the design of the test database that uses SQL NULLs (*SQL NULL design*).

Figure 14.2 illustrates the design of the test database, which uses sixth normal form (6NF) tables (*6NF design*). Table T is in 6NF if and only if it cannot be nonloss decomposed at all (other than the identity projection of T) [17]. Date [17] also notes that the identity projection of a table T is the projection over all of its columns. For each non-key attribute in a conceptual data model, we created two different 6NF tables. For instance, in case of attribute *name* of entity type *Artist*, we created table *Artist_name*, which is used to record artist names that are known.

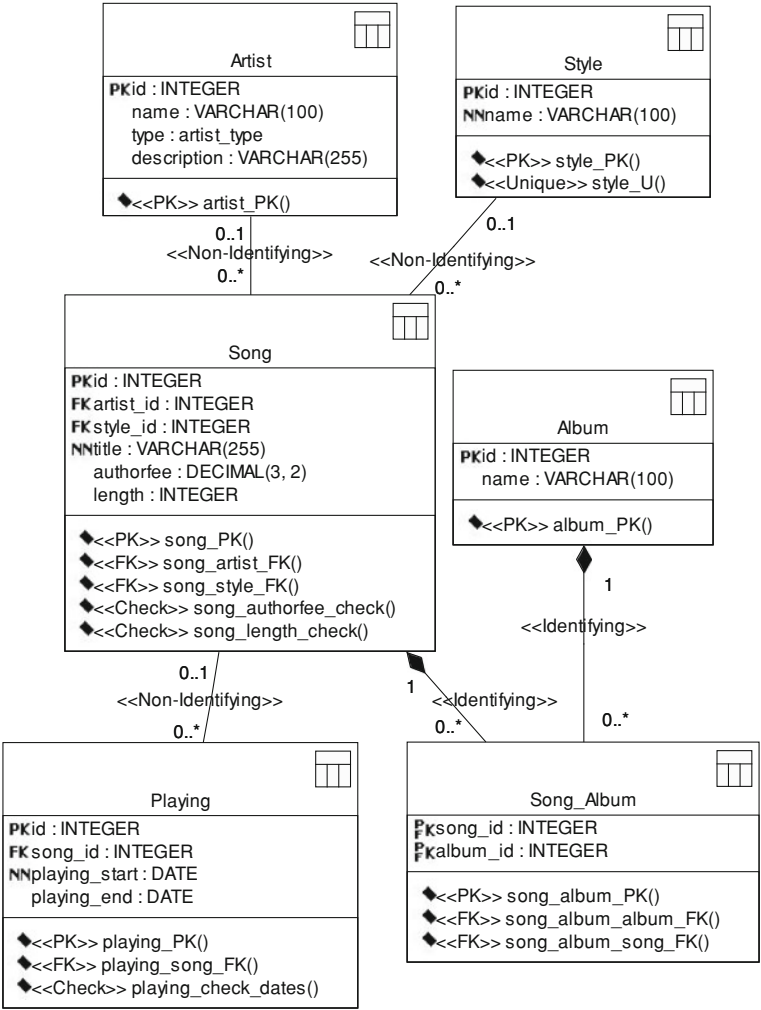


Fig. 14.1 Design of tables in case of the SQL NULL design

Table *Artist_name_unknown* is used to record the identifiers of artists, whose name is unknown.

Figure 14.3 presents the design of the test database that uses special values to represent missing information (*special values design*). Let us assume that a column *c* has the type TP. In case of the special values design one uses a set of values that belong to TP to represent missing information. One has to select such values from TP that are otherwise not used in *c*. For instance, we used special value “-1” in case of columns with the type INTEGER or VARCHAR to denote that the value is unknown. We used the special value “1970-01-01 00:00:01” in case of columns with the type TIMESTAMP to denote that the value is unknown. The selection of

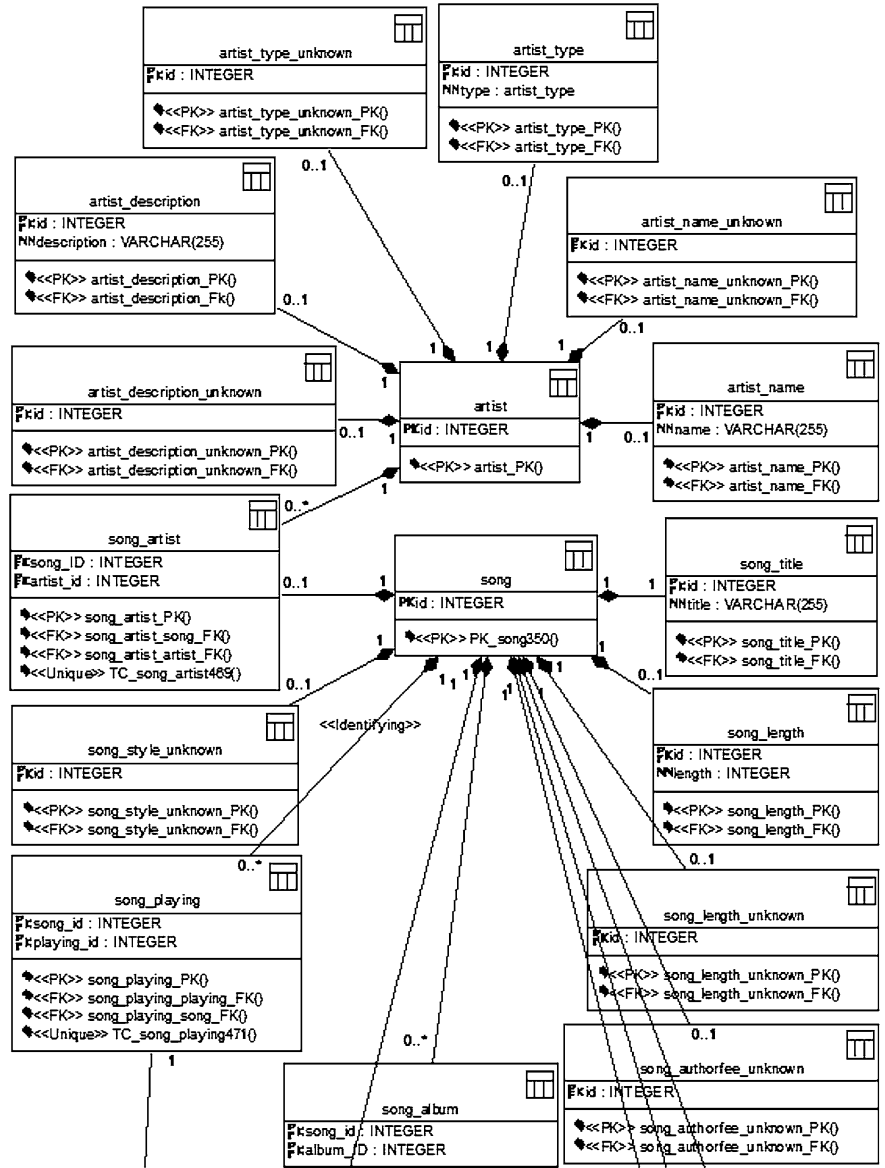
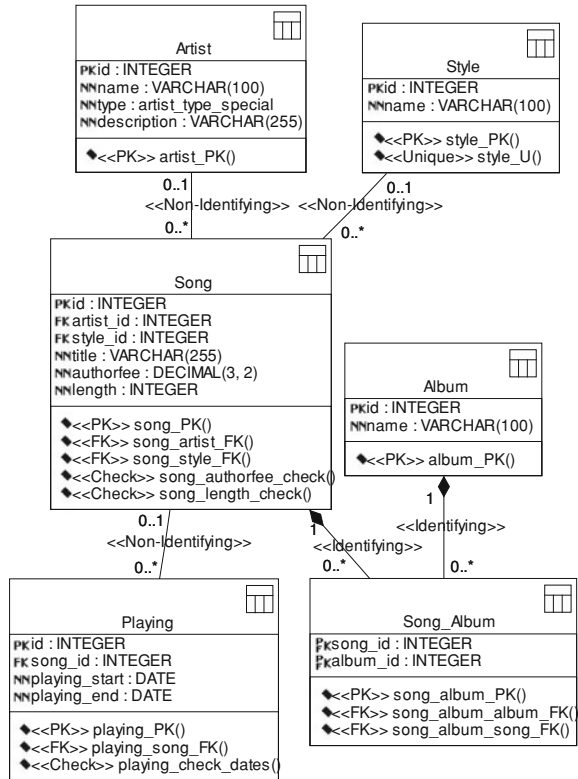


Fig. 14.2 Design of some of the tables in case of the 6NF design

special values is one of the difficulties of this design. In case of some types it is not possible to find a special value for each different reason why the information could be missing. For instance, the type BOOLEAN contains only values TRUE, FALSE, and UNKNOWN (that is represented by NULL). A difference between

Fig. 14.3 Design of tables in case of the special values design



the special values design and the SQL NULL design is that in case of the former all the columns are mandatory (have NOT NULL constraints).

One could extend the 6NF design and the special values design in order to represent different kinds of missing information. For instance, one might create a separate table for each possible reason why the name of an artist could be missing in case of the 6NF design. One could use the following special values in case of the columns with the type INTEGER or VARCHAR in case of the special values design.

- -1—Attribute cannot have a value in case of this object.
- -2—Attribute can have a value in case of this object but the value is unknown at the moment.
- -3—Attribute can have a value in case of this object and the value is known but it is not recorded at the moment.

For this experiment, we decided to simplify the designs and did not represent different kinds of missing information. In this way the two designs are comparable with the SQL NULL design that also does not allow us to distinguish between different reasons why the information is missing.

Table 14.1 Criteria that we used to evaluate database designs

| Criterion | Database level | ISO 9126 quality characteristic |
|---|------------------|---------------------------------|
| Complexity of data manipulation operations | External level | Maintainability, usability |
| Schema size | Conceptual level | Maintainability |
| Integrity constraints | Conceptual level | Maintainability, functionality |
| Complexity of schema modification | Conceptual level | Maintainability |
| Performance of data manipulation operations | Internal level | Efficiency |
| Data size | Internal level | Efficiency |

14.2.4 Criteria

Table 14.1 shows criteria that we used to evaluate the designs. For each criterion, we also specify its corresponding database level and ISO 9126 quality characteristics. Table 14.2 shows software measures that correspond to the selected criteria. In case of each selected measure m —the smaller is the measurement result in case of an alternative al , the better is al in the context of the criterion that is associated with m .

We counted the SQL key words in PostgreSQL 8.3 [15] in the data manipulation (DM) language statements as well as in the data definition language statements in order to evaluate the complexity of operations from the point of view of parties who have to perform the operations and write the statements.

Examples of SQL key words are: SELECT, VALUES, INNER, JOIN, and AND.

Piattini et al. [18] define Schema Size measure, which is used to evaluate maintainability of databases, “as the sum of the tables size (TS) in the schema” [18, p. 7] and Table Size measure “as the sum of the total size of the simple columns (TSSC) and the total size of the complex columns” [18, p. 6]. The size of each simple column is one. All the columns are simple columns in case of the designs in this evaluation. Therefore, in this case Table Size of a base table T is equal to the total number of columns in T.

NFK measure, which we used to evaluate the designs in terms of integrity constraints, is the number of foreign keys in the database schema [19].

In case of the performance criterion, we performed the same task five times in case of all the designs and calculated the average result in case of each design.

We used the PostgreSQLTM system-defined function `pg_total_relation_size` [15] to find the total size of base tables (including indexes and toasted data) in case of all the designs.

Table 14.2 Software measures that we used to evaluate database designs

| Criterion | Sub-criterion | Software measure |
|---|------------------------------------|--|
| Complexity of data manipulation operations | Complexity of an INSERT operation | Number of SQL key words in the statements |
| | Complexity of a SELECT operation | Number of SQL key words in the statements |
| Schema size | – | Schema size [18] |
| Integrity constraints | – | NFK [19] |
| Complexity of schema modification | – | Number of SQL key words in the statements |
| Performance of data manipulation operations | Performance of an INSERT operation | Time in seconds |
| | Performance of a SELECT operation | Time in seconds |
| Data size | – | Size of base tables and indexes in megabytes |

14.2.5 Implementation of Databases and Generation of Test Data

We implemented the designs by using the open-source DBMS PostgreSQL™ 8.3 [15]. We implemented the designs in the different schemas of the same database to prevent name conflicts. We created a PHP application to generate the test data. The application generates a set of test data to the main memory and then inserts the data to the tables that have been created based on different designs. One has to use the same data in case of different designs because otherwise the results of measurements would not be comparable.

We generated data about 25,000 songs, 3,000 artists, 5,000 albums, 250 styles, 175,000 playlist entries (instances of *Playing*), and 295,000 instances of relationship type between *Song* and *Album*. We used the following guidelines to generate test data.

- If an attribute a of an entity type E is optional, then the value of a is missing in case of about 10 % of entities with the type E .
- Each name of an artist, an album, and a style consists of between 2 and 100 (end points included) alphanumeric characters.
- Each title of a song consists of between 2 and 255 (end points included) alphanumeric characters.
- There is 1 s difference between the end of one playlist entry and the beginning of another playlist entry. We generated data about playlist entries for one year, starting from January 1st, 2010 00:00.

The computer, where we performed the experiments, had the following characteristics: processor: Intel Core 2 Quad, 3.4 GHz; RAM 4 GB DDRII; Windows XP Professional SP2; PostgreSQL 8.3; Apache HTTP Server 2.2 + *php_pdo_pgsql* extension.

14.3 Results of the Experiment

14.3.1 Relative Importance of Criteria

Figures 14.4 and 14.5 present the relative importance of the criteria (weights) in case of *context 1* and *context 2*, respectively. We used the information that is specified in the description of contexts (see Sect. 14.2.2) to compare the criteria pairwise. For instance, in case of *context 1*, performance of data manipulation (DM) operations is strongly preferred compared to complexity of DM operations, integrity constraints, schema size, and complexity of schema modification. In case of *context 1*, performance of DM operations is very strongly preferred compared to data size. We do not present comparison matrices in the paper due to the space restrictions.

We used web-based AHP tool Web-HIPRE [20] to perform the analysis.

The analysis results (comparisons of the criteria pairwise as well as comparisons of alternatives pairwise in terms of criteria) are presented in the file: http://staff.ttu.ee/~eessaar/files/Missing_data.pdf.

The consistency measure (CM) [20], which indicates the consistency of decision maker, indicated that all the comparison matrices were sufficiently consistent.

14.3.2 Evaluation of Alternatives in Terms of Criteria

Table 14.3 summarizes the results of measurements in case of different alternatives. Next, we describe the tasks that we performed in order to measure the alternatives. One has to bear in mind that it might be possible to solve a task by using different sets of database language statements and hence the results of measurements depend on the selected solutions. The statements, which we used to solve the tasks, are in the file: http://staff.ttu.ee/~eessaar/files/Missing_data.pdf.

In case of the “Complexity of an INSERT operation” sub-criterion we performed the task: “Insert to the database a song, the performer and style of which is unknown, the title is ‘Chiri-Biri-Binn’, the author fee is 6.00 EEK, and the length is 190 s”. The number of SQL key words in a solution of the task is the biggest in case of the 6NF design. In case of the SQL NULL design and the special values design one has to insert a row to the table *Song*. In case of the 6NF design and the current task, one has to insert one row into each of the following tables: *Song*, *Song_artist_unknown*, *Song_title*, *Song_style_unknown*, *Song_authorfee*, and *Song_length*. The INSERT statements must be in one transaction. If entity type *Song* would have more attributes, then we would also need more INSERT statements in case of the 6NF design.

In case of the “Complexity of a SELECT operation” sub-criterion we performed the task: “Find the number of songs, the author fee of which is between

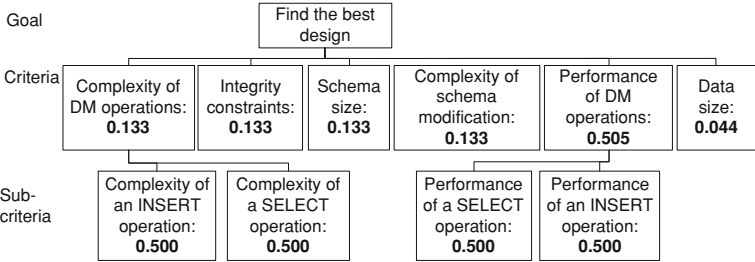


Fig. 14.4 Relative importance of criteria in case of context 1

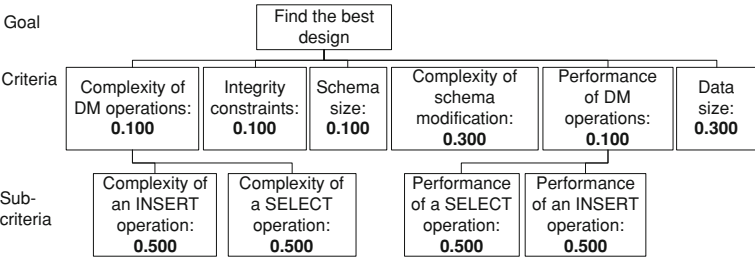


Fig. 14.5 Relative importance of criteria in case of context 2

Table 14.3 Measurement results in case of different designs

| Software measure | SQL NULL design | 6NF design | Special values design |
|--|-----------------|------------|-----------------------|
| Number of SQL key words in statements in case of an INSERT operation | 5 | 21 | 3 |
| Number of SQL key words in statements in case of a SELECT operation | 23 | 35 | 26 |
| Schema size | 20 | 43 | 20 |
| NFK | 5 | 28 | 5 |
| Number of SQL key words in statements in case of a schema modification operation | 14 | 47 | 17 |
| Performance of an INSERT operation in seconds | 210 | 559 | 228 |
| Performance of a SELECT operation in seconds | 0.212 | 0.030 | 0.228 |
| Size of base tables and indexes in megabytes | 33.7 | 74.2 | 33.8 |

4.50–5.50 EEK (end points included), the length of which is between 180–190 s (end points included), and that have been played more than five times during the last three months”. In case of the 6NF design we have to perform the biggest number of join operations to solve the task and hence the number of SQL key words is the biggest in case of this design.

In case of the “Performance of an INSERT operation” sub-criterion we performed the task to insert test data to the tables. [Section 14.2.5](#) contains information about the test data.

In case of the “Performance of a SELECT operation” sub-criterion we used the same task as in case of the “Complexity of a SELECT operation” sub-criterion.

In case of the “Complexity of schema modification” criterion we performed the task: “Modify the schema to make possible registration of the creation year of the song. In case of each song the year can be between 1975 and 2020 (end points included) or it could be unknown”. The number of SQL key words in a solution of the task is the biggest in case of the 6NF design because one has to create separate tables *Song_year* and *Song_year_unknown* as well as enforce foreign key constraints (within one transaction). If we would modify the design so that there is one separate table for each reason why the value of an attribute can be missing, then the number of tables and hence the complexity of statements would be even bigger. The number of SQL key words is bigger in case of the special values design compared to the SQL NULL design because in case of the special values design one has to create NOT NULL constraint to the column *year*. In addition, the CHECK constraint, which should be added to the column *year*, must take into account that one possible value in the column is the special value -1 .

14.3.3 The Results of the Evaluation

In case of each criterion c that has no sub-criteria, we compared all the designs (alternatives) pairwise in terms of c .

In case of each considered criterion c , we divided the results of measurements in case of c (see [Table 14.3](#)) pairwise in order to find out how many times one alternative is better than another in terms of c . For instance, to find out how many times the SQL NULL design is better than the 6NF design in terms of the criterion “Complexity of an INSERT operation”, we performed the calculation $21/5 = 4.2$.

Tables [14.4](#) and [14.5](#) summarize the results of evaluation of the three designs in case of two different contexts (see [Sect. 14.2.2](#)). Row *Relative goodness* presents the final scores of the designs in case of the different contexts. We found them by summarizing the scores of alternatives in case of each design. The bigger is the final score, the better is the design in terms of the goal (find the best design).

Based on the evaluation, the SQL NULL design is the best design and the 6NF design is the worst design in case of both contexts. The 6NF design achieved relatively good results in case of *context 1* because in this case performance of DM operations is very important and the 6NF design has the best measurements results in case of one of its sub-criterion.

We do not claim that the SQL NULL design is always the best and the 6NF design is always the worst. The results depend on contexts, alternatives, criteria, measures, tasks, and solutions of tasks. For instance, if the schema modification statements or data modification statements would be mostly automatically

Table 14.4 The results of evaluation of designs in case of context 1

| Criterion | SQL NULL design | 6NF design | Special values design |
|--|--------------------|---------------|--------------------------|
| Complexity of data manipulation operations | 0.041 | 0.019 | 0.052 |
| Schema size | 0.046 | 0.021 | 0.046 |
| Integrity constraints | 0.052 | 0.009 | 0.052 |
| Complexity of schema modification | 0.053 | 0.016 | 0.044 |
| Performance of data manipulation operations | 0.139 | 0.239 | 0.127 |
| Data size | 0.018 | 0.008 | 0.018 |
| <i>Relative goodness</i> | <i>0.349</i> | <i>0.313</i> | <i>0.339</i> |

Table 14.5 The results of evaluation of designs in case of context 2

| Criterion | SQL NULL design | 6NF design | Special values design |
|--|--------------------|---------------|--------------------------|
| Complexity of data manipulation operation | 0.037 | 0.017 | 0.046 |
| Schema size | 0.041 | 0.019 | 0.041 |
| Integrity constraints | 0.046 | 0.008 | 0.046 |
| Complexity of schema modification | 0.141 | 0.042 | 0.117 |
| Performance of data manipulation operations | 0.027 | 0.047 | 0.025 |
| Data size | 0.122 | 0.056 | 0.122 |
| <i>Relative goodness</i> | <i>0.414</i> | <i>0.189</i> | <i>0.397</i> |

generated by a software program, then the complexity of these statements would not be as important criterion any more. Rönnbäck et al. [6] propose anchor modeling, which can be used for agile information modeling in evolving databases. The modeling method suggests *generation* of database schema where tables are mostly in the sixth normal form (6NF). One of the advantages of using 6NF tables is that all the changes in the schema are done in the form of extensions (by adding new tables) instead of modifying existing tables. It simplifies evolution of databases and applications, which use it. In our evaluation, we only consider the complexity of writing schema modification statements and do not consider other aspects of schema evolution (like transparency or instance migration). Hence, in our evaluation the 6NF design has a low score in case of the “Complexity of schema modification” criterion. It points to the need to continue evaluation of the designs by using larger sets of contexts, criteria, measures, and tasks.

14.4 Conclusions

In this paper, we presented the results of evaluation of three SQL database designs that can be used to deal with missing information. One of the designs uses NULL-marks, another uses tables that are in the sixth normal form, and one uses special data values to represent missing information. The evaluation method is based on the Analytic Hierarchy Process. The method uses the results of measurements in order to compare the designs pairwise. We evaluated the designs in case of two hypothetical contexts and in both cases the design, which uses SQL NULLs, got the best results. However, the design with tables, which are on the sixth normal form, got the best results in case of the “Performance of a SELECT operation” sub-criterion. In addition, some of the problems of using this design can be overcome by using code generators.

Future work includes evaluation of the designs by using larger sets of contexts, criteria, measures, and tasks.

References

1. Melton J (2006) IWD 9075-1:200x(E) information technology—database languages—SQL—Part 1: framework (SQL/framework). Feb 2006
2. Date CJ (2003) An introduction to database systems, 8th edn. Pearson/Addison Wesley, Boston
3. Date CJ, Darwen H (2007) Databases, types, and the relational model. The Third Manifesto, 3rd edn. Addison-Wesley
4. Date CJ (2009) SQL and relational theory. How to write accurate SQL code. O'Reilly, Romulus, MI
5. Date CJ, Darwen H (2010) Database explorations. Essays on the third manifesto and related topics. Trafford Publishing, San Francisco
6. Rönnbäck L, Regardt O, Bergholtz M, Johannesson P, Wohed P (2010) Anchor Modeling. Agile information modeling in evolving data environments. *Data Knowl Eng* 69(12), 1229–1253
7. Karwin B (2010) SQL antipatterns. Avoiding the pitfalls of database programming. The pragmatic bookshelf
8. PredictiveDB, [Online document]. Accessed 5 aug 2011 <http://www.predictivedb.com/>.
9. Codd EF (1990) The relational model for database management: version 2. Addison-Wesley, Menlo Park
10. Liu K-C, Sunderraman R (1990) Indefinite and maybe information in relational databases, *ACM Trans Database Syst* 15(1), 1–39
11. Date CJ, Darwen H (2000) Foundation for future database systems. The third manifesto, 2nd edn. Addison-Wesley, Massachusetts
12. Eessaar E, Soobik M (2012) A decision support method for evaluating database designs. *Comput Sci Inf Syst* 4:345–365
13. Saaty TL (1994) How to make a decision: the analytic hierarchy process. *Interfaces* 24(6):19–43
14. ISO/IEC 9126-1, Software engineering—product quality—Part 1: quality model, first edn.: 2001-06-15

15. PostgreSQL 8.3 Documentation, [Online document]. Accessed 9 Aug 2009 <http://www.postgresql.org/docs/>.
16. Gornik D, UML data modeling profile. In: Rational software white paper TP162, 05/2002
17. Date CJ (2006) The relational database dictionary. A comprehensive glossary of relational terms and concepts, with illustrative examples. O'Reilly, Punta Gorda, FL
18. Piattini M, Calero C, Sahraoui H, Lounis H (2001) Object-relational database metrics. *L'Object* 2:488–494
19. Baroni AL, Calero C, Abreu FB, Piattini M (2006) Object-relational database metrics formalization. In: Sixth international conference on quality software, pp 30–37
20. Mustajoki J, Hamalainen RP (2000) Web-HIPRE: global decision support by value tree and AHP analysis. *INFOR* 38(3):208–220

Chapter 15

Mobile English Learning System: A Conceptual Framework for Malaysian Primary School

Saipunidzam Mahamad, Fatimah Annor Ahmad Rashid, Mohammad
Noor Ibrahim and Rozana Kasbon

Abstract Mobile learning has become apprentice for today's education. With the emerging of mobile technologies, variety of features of current mobile applications could benefit the learners with enhanced learning environment while playing anywhere, anytime. This paper presents a framework of mobile English learning system for Malaysian Primary School for children aged from 9 to 12 years old. It aims to provide both fun and educational features to the learners by offering game-based activities to stimulate their imagination and challenge their curriculum skills such as antonyms, synonyms and sentence structures. The proposed architecture supports cross platform user interface and the framework offers reliability to provide the learner with some erudition input.

15.1 Introduction

Today, mobile applications have grown rapidly. Mobile devices have shown tremendous use of purpose from making and receiving a phone call to send multimedia message and now emerging for playing games [1]. Many new applications

S. Mahamad (✉) · F. A. A. Rashid · M. N. Ibrahim · R. Kasbon
Department of Computer and Information Sciences, Universiti Teknologi PETRONAS,
Bandar Seri Iskandar, 31550, Tronoh, Perak, Malaysia
e-mail: saipunidzam_mahamad@petronas.com.my

F. A. A. Rashid
e-mail: fatimahannor@gmail.com

M. N. Ibrahim
e-mail: mnoor_ibrahim@petronas.com.my

R. Kasbon
e-mail: rozank@petronas.com.my

have been designed to satisfy various types of users need. However, not every mobile application could be used suitably for young learners. Number of applications specifically designed for this target group is still rather small. It is commonly tuned and designed for entertainment and amusement purposes only, which mostly are computer games. No doubt, that most of games available is bidding on enjoyment purpose only leaving other vital elements untouched. In addition, the design of current mobile games fails to stimulate their imagination.

In this digital age, there are signs of motivating potential and possible learning gains of games played on mobile devices. Typical learners may have few educational phone games applications to play, learn and interact with and enhance their critical thinking skills at the same time. Mobile technologies and applications, as learning tools have increasingly become viable with the growing sophistication and affordability of their use, hence growing interest in the field of m-learning. M-learning has been regarded as the potential of learning or as an essential element of any other form of educational development in the future. With this kind of learning, mobile applications with cross-curricular activities are affordable and they have become a key benefit when engaging with classroom activity. The needs of lifelong learning, just-in-time training and retraining led to the development of widely accessible and reusable digital content and learning repositories. Besides, the connection between learners, teachers and caregivers is vital as it facilitates to structure the interchange between student appreciation of technology and training in formal education.

Mobile learning can be divided into three major application types which are SMS, Mobile Application and Mobile Game. The later type is also known as mobile game-based learning [2]. With the advent of mobile communication systems, it enables users to learn more effectively.

This paper proposes an interactive mobile English learning system that is not only providing fun activities during the learning, but also satisfying learners' need to increase their reading comprehension and confidence as reading is the key to success in all subjects [2]. The development focuses on Malaysian primary school, approximately learners in the age range of 9–12 years old. The system features both educational and fun learning features. The system is endeavored to enhance knowledge of students at primary school level with average technical skills on reading comprehension. It also solves user intermittent uncertainties on their English grammar skills. Certainly, such an early conceptual development let students having deeper understanding in the academia, makes them appreciate English as an international and world wide language.

In this paper, an overview of the framework for mobile English learning system is presented. It is not only providing fun activities during the play, but also satisfying children need to increase their learning. In the next section, this paper reviews the related works regarding the studies of mobile learning system, game-based approach, in particular, the effects of current technologies to the environment of learning focuses on Malaysian primary school. The methodology and conceptual framework of the project are discussed along with the result, discussion and conclusion in the last section.

15.2 Background and Related Work

The concept of Edutainment that is a mixture of education and entertainment has long been introduced to the world, since the first educational games for the youngest and since the moral fairy tales, aiming to give ethics lessons to pupils while entertaining them. The power of Internet which been utilized for web based platforms is aimed to providing efficient access to information regarding computer learning applications and to create online communities. However, the target users are limited to only on traditional Internet users, excluding an ongoing population of people that access World Wide Web using mobile devices. Certainly, teaching methods based on educational application expected to be extremely attractive especially to school learners [4, 5]. It shows that computer games raise the efficiency of learning if the developer increase the intrinsic motivation and link the goals winning the game and learning the material.

Learning styles and attitudes toward application of technology innovation in education are vital for understanding of learner factors, such as mobile learning was considered important for designing and implementing educational innovation. Students learn more efficiently when pedagogical procedures are adapted to the students' individual differences [5]. Meanwhile, learning is the process through which persons become the human beings, and it takes place through a variety of media, devices, strategies and processes [6].

Many researches have been conducted on mobile game-based learning systems. One of which has been conducted at University of Trieste named "Mogabal". It is extending the project acronym of MOBILE GAME-BASED Learning, defined it as a sort of game engine rather than a game, as its graphical aspects, rules, educational contents and many other elements can be fully configured and altered thus giving the potentiality for creation of widely different games and game styles [3]. On the other hand, Zaibon and Shiratuddin [2], described mobile learning as a learning technique that happens across locations or e-learning through mobile computational devices. The development use the gamed-based approach for a game play to enhance motivation in learning, engage in knowledge acquisition and improve effectiveness of learning through mobile environment. Its identified main characteristics for the applications are; (1) must be easy to follow and support a playful way of learning; (2) the learning content should be split into small units which require only a reduced span of attention so that game play and learning can take place.

In hypothesizing the theory of mobile learning, several steps have been suggested by a number of researches [7]. First step is to differentiate the unique features of mobile learning evaluation against other types of learning activity. An apparent yet vital distinguish is that learners are persistently on the move. Learning crosswise space, time, topics and moving in and out of engagement with technology is what we are. Secondly, mobile learning should grip the significant learning that happens outside classrooms and lecture hall as people commence and constitute their activities to enable educational progressions and outcomes. Third step is that this kind of learning must be based on modern accounts of preparations

that enable flourishing learning. Lastly, mobile learning must take into consideration of the ubiquitous utilization of personal and shared technology.

The principles of designing educational software are to emphasis on the content and their interaction in order to prevent from failing of costing too much. In developing such application, learning features is given more emphasizing together with attractive graphical user interface. To give extra value to this proposed game, cross platform tool is used in developing it as it cannot only be running via mobile with different operating system but on computer as well.

15.3 System Design and Framework

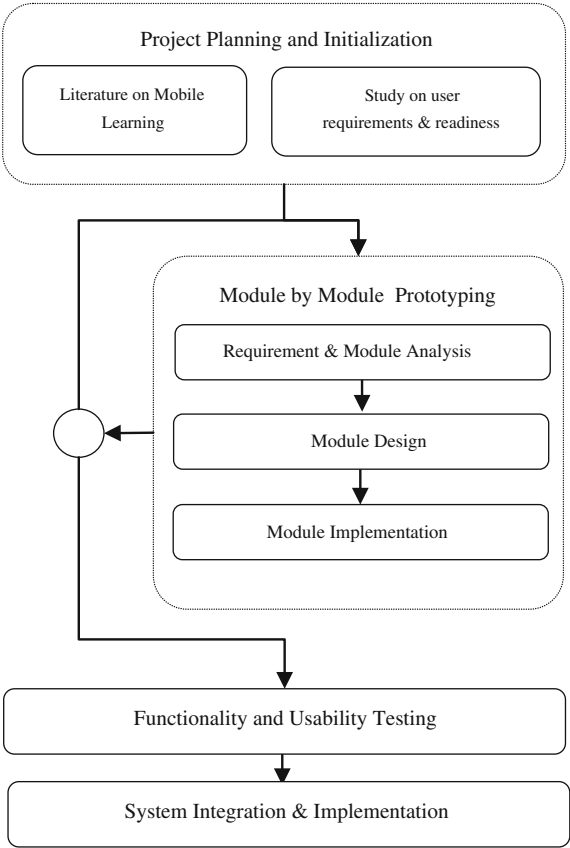
The modified conceptual project framework [8] comprises four main stages is shown in Fig. 15.1. The first stage is the early analysis, which consists of problem initialization and planning. The second stage is the module development followed by testing and system implementation. The functionality and usability testing are important to ensure the developed learning system is robust and free from errors that meets the user requirements.

The overall system architecture adapted from [9] is shown in Fig. 15.2. The development has shown significant contribution and improvement from Malaysian perspective on education environment. It explains all of the conceptualize data gathered before were put together into actual physical model. The model illustrates several level of user authentications i.e. Student, Teacher and Administrator, which have different role of permission in accessing the contents and functionalities. Each of the users has to go through an authentication function before proceeding to the individual function such as lesson, setting and progress tracking. The structural design is developed using J2ME, a cross-platform application and user interface framework. It is a stand-alone package or in combination with the libraries and development tools as a complete software development kit.

The concept of this project is to provide a strong base for grade 3 to grade 6 learners based on Malaysian primary school on reading comprehension skills and language art through Multimedia elements such as text, audio and images. As a challenging and hence interesting platform for Malaysian primary school, level students it is hoped that target users are thrilled by the outcome.

The proposed mobile English learning system is divided into three different modules, namely Player, Information and Play. Player Selection module acts as the entrance for the first time player as well as the existing user. The system involves three main decisions from the player. First, the player has to select his status whether he is a new or existing player. Then, choose either read a plot narration or skip it for existing player. After that, the player can decide which division in the abounded mansion he wants to play first. After winning all the rewards in the reading comprehension activities and collecting the clues, the player has to match the reward which is the victim's personal belonging to the correct victim in order to release the victim.

Fig. 15.1 Conceptual framework



While the Information module acts as the entrance for the user and serves as the mission information doorway. Player can read the plot of the story and get basic information about grammar, the major and common category of grammar. By clicking the action buttons provided, further explanation including an example with image will be provided for each category.

The Play module is where player will undergo their activities. Different division in the abounded mansion will lead to different reading comprehension activities. It serves as the communication channel between the user and the application. Player can test their knowledge and understanding about grammar as well as improve their vocabulary skills. There are two main activities where each activity can directly develop user’s critical thinking. First activity is Knowledge Test where by given a number of words, user may guess which category does the word fall into. The second activity, which is Vocabulary Play will allow the user to extract various words from a given example of phrase and structured a scramble alphabetical and words to form a phrase and sentences according to their understanding.

Figure 15.3 shows the use case diagram for mobile English learning system. It illustrates six use cases, which are select application, choose grammar menu, read

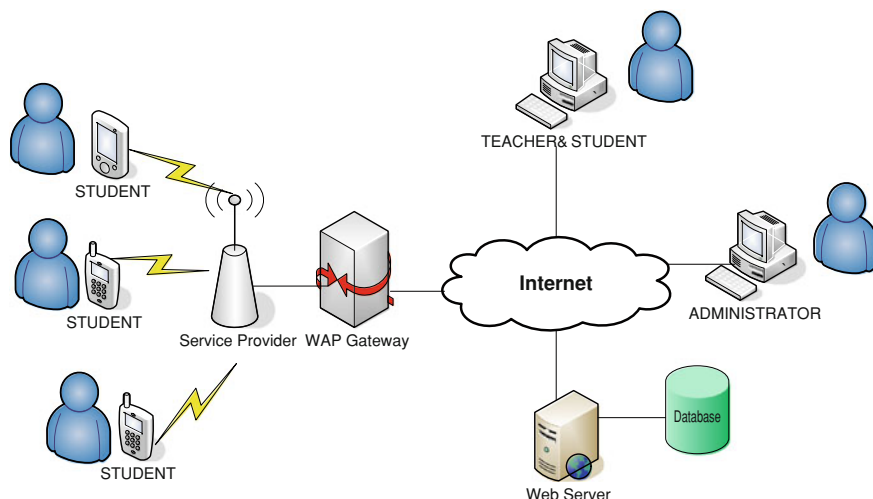


Fig. 15.2 Adapted system architecture [9]

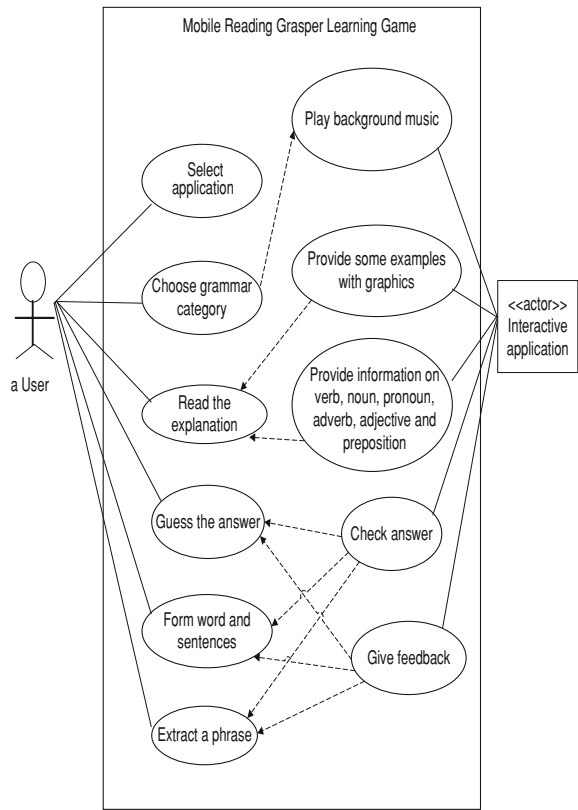
the explanation, guess the answer, form word and sentences and extract a phrase. Two different actors engage in this system, which are the user and the application itself. Three types of relationships involve which are association relationships, include relationship and extends relationship. The diagram later will expend to suite with the adapted system architecture that extend the functionality to teacher, level of difficulties and progress monitoring.

The class diagram shown in Fig. 15.4 depicts classes, which include the attributes and its associations. The user can choose which three categories he wants to explore first where this is a one to six relationships. After selecting the category, the user can choose which sub class he wants to read, understand and identify the example provided by the system and this is a one to one relationship where one category is offering more sub classes. When the user has a better understanding, he can test his knowledge. This is a one to one relationship where user can guess a given example of phrase at time. Besides, user has the opportunity to enhance their vocabulary skill by playing more interactive activities.

15.4 Discussion

Mobile learning system is one of the system that been undergoing a lot of researches. A complete understanding of the m-learning development is necessary inline with development of an interactive graphical application that involved many stages and a lot of animations as well as graphical user interfaces. Designing it would be a tough task, as many elements need to be considered. Other than that,

Fig. 15.3 Use case diagram

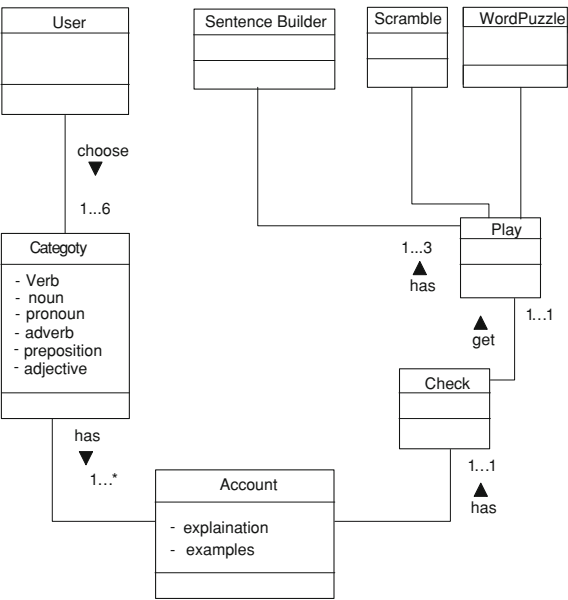


not all development tools can support mobile application. Finding the perfect one and familiarize with it would never be easy. Besides, the criteria for designing an application for mobile also will be different considering the system requirements and the screen resolution. Experience and knowledge with the technology and tools are essential, as this will help during the development of this proposed system.

Recent study demonstrates that mobile games could enhance concentration, thinking and learning time. Several scholars also suggested the concept of massively multiplayer online games for education to be introduced in the classroom. These ideas would support more players simultaneously with the present of internet connection. It is not necessary to be played on personal computer but with new mobile capability, it increased their accessed. It would benefit the learning processes that increase communication on a large scale, to interact meaningfully with people outside the classroom.

In designing application for children, many perspectives have to be taken into considerations. Most children may not fully understand text-based instructions. Children expect to see immediate result of their actions. If nothing happens after

Fig. 15.4 Class diagram



their input, children may repeat their action until something does occur which may cause a chain of unexpected and unwanted events. Children often expect constant auditory and visual feedback although it can be annoying for adult users. The principles of designing software educational software emphasizing on constructing of educational software should be based on some method in order to prevent from failing of costing too much or of being great delayed.

15.5 Conclusion

The developed learning system may provide advantage into overcome the limitations such as the limitations on features and educational or learning functionalities by providing user with fun and education features at the same time. Hence, arming the system with the above-mentioned features, a quality mobile game-based learning, which meets the user expectations, is possible. While learners demanding for games, educators may opt for educational games that can educate their children at the same time as well, instead of just having fun and waste their time for unbeneficial games.

References

1. Mahamad S, Mazlan EM, Kasbon R, Kalid KS, Rusdi NS (2007) Personalised mobile picture puzzle. *World Acad Sci Eng Technol* 25:6–14
2. Zaibon SB, Shiratuddin N (2009) Towards developing mobile game-based learning engineering model. In: *World Congress on Computer Science and Information Engineering*, Los Angeles, California USA
3. Mininel S, Vatta F, Gaion S, Ukovich W, Fanti MP (2009), A customizable game engine for mobile game-based learning. In: *Proceedings of the 2009 IEEE international conference on systems, man, and cybernetics*, San Antonio, TX, USA
4. Schwabe G, Goth C (2005) Mobile learning with a mobile game: design and motivational effects. *J Comput Assist* 21:204–216
5. Chao C (2006) An investigation of learning style differences and attitudes toward digital game-based learning among mobile users. In: *Fourth IEEE international workshop on wireless, mobile and ubiquitous technology in education (ICHIT'06)*
6. Mishra S, Sharma CR (2005) *Interactive multimedia in education and training*. India Idea Group Publishing, India
7. Sharples M, Taylor J, Vavoula G (2005) *Towards a theory of mobile learning*. University of Birmingham, Birmingham
8. Noor Ibrahim M, Mahamad S, Wei ECN (2010) Mobile learning: an application prototype for AVL tree learning object. *Appl Mech Mater* 39:176–181 ISSN 1660-9336 (ISBN-13: 978-0-87849-218-3).
9. Mahamad S, Ibrahim MN, Taib SM (2010) M-learning: a new paradigm of learning mathematics in Malaysia. *Int J Comput Sci Inf Technol (IJCSIT)* 2(4):76–86 ISSN 0975–4660

Chapter 16

Dynamic Cache Miss-Rate Reduction

Mazen AbuZaher, Bayan Alayoubi, Basma Alefeshat
and Abdelwadood Mesleh

Abstract Cache miss rate reduction represents the classical approach to improving cache performance. One method to reduce cache miss rate is higher associativity cache. In this paper we introduce a novel method to enhance cache performance by dynamically choosing the best associativity that can be used to reduce cache miss rate depending on the running program needs. According to our approach cache can be moved from direct mapped to 8-way associative online without the need to perform complex address mapping.

16.1 Introduction

Computer programmers would want unlimited amounts of high performance cache memory. Increasing cache performance can be done by several methods [1] such as reducing cache misses.

Cache misses are divided to three categories [1]: compulsory, capacity, and conflict. Many studies have been done to increase cache performance by reducing cache misses. Some approaches as in [2] try to enhance cache indexing to eliminate conflict misses. Other works focus on cache replacement algorithms as in [3, 4]. The work in [5] uses static scheduling to reduce cache capacity misses. Other methods as in [6] use prediction to increase cache prefetching accuracy.

M. AbuZaher (✉) · B. Alayoubi · B. Alefeshat · A. Mesleh
Computer Engineering Department, Al-Balqa' Applied University, Amman, Jordan
e-mail: mazen.abuzaher@fet.edu.jo

A. Mesleh
e-mail: wadood@fet.edu.jo

The traditional method to reduce conflict misses is using higher associative cache as possible. But the main problem is when increasing associativity; hit time will increase in turn which degraded cache performance.

In implementation, processor must adopt just one type of associativity for the same cache. Since increasing associativity will increase hit time the vast majority of processor caches are direct mapped (one-way), two-way set associative, or four-way set associative [1].

In this work we design cache simulator and use it to introduce a new method to reduce conflict misses depending on higher associativity (up to 8-way). In addition our approach tries to maintain cache hit time to be minimum as possible.

16.2 Our Approach

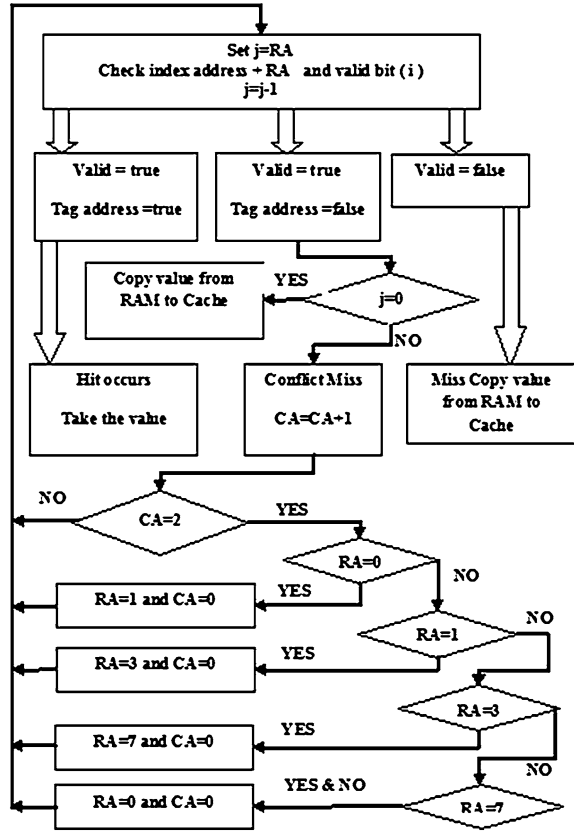
In this work we limit the maximum associativity to be 8-way, that is because of according to rule of thumb that “eight-way set associative is for practical purposes as effective in reducing misses as fully associative” [1].

Our approach steps to enhance cache performance as following:

1. Start with direct mapped cache to gain minimum hit time.
2. Initialize a two bits counter (associativity counter “CA”), and 4-bits register (associativity register “RA”) to zero.
3. CA will be incremented with any conflict miss. And reset if more than 2.
4. If any block on the cache gain more than two conflict misses ($CA = 2$), then, set RA register to one, reset CA, and before going to the next lower level of the memory hierarchy, a second cache entry is checked to see if it matches there. A simple way is to add one to the index field to find the other block. Accordingly, we have a direct mapped cache but it will seem as two-way set associative cache. Now any block has two locations in the cache.
5. Again observe CA if equal two, then, set RA register to three, reset CA, and before going to the next lower level of the memory hierarchy, see if there is match in index field+ RA, index field+ RA-1, and index field+ RA-2 to find the other block. Now any block has four locations in the cache.
6. Again observe CA if equal two, then, set RA register to seven, reset CA, and before going to the next lower level of the memory hierarchy, see if there is match in index field+ RA, index field+ RA-1, index field+ RA-2, index field+ RA-3, index field+ RA-4, index field+ RA-5, and index field+ RA-6 to find the other block. Now any block has eight locations in the cache.
7. To improve hit time, again observe CA if equal two then return to step 1.

According to our approach, the main hardware mapping strategy is direct (one-way), that is to gain minimum hit time. The approach gradually increases associativity to reduce conflict misses and maintains best hit time. If there is no advantage of increases associativity (no conflict misses reduction) the approach returns to direct mapped cache to provide the best hit time. Depending on cache

Fig. 16.1 Enhancement performance approach flowchart



size, RA and CA values can be modified. Figure 16.1 shows the flowchart of our cache enhancement algorithm.

16.3 Results

To test our approach we design our own simulator, which simulates the real environment of our computer. This simulator can give us the hit-miss counters, hit-miss rates and hit-miss times. In addition we also can see the content of cache in any stage.

To use our simulator we create trace files to test our approach and compare it with traditional set associative technique. Table 16.1 shows the miss time in micro second for ten trace files using all mapping strategy (1-way to 8-way) and using our approach.

The trace files are different from each other in size and content. Each program has its own cache size, but all trace files applied using the same processor speed and cache and RAM access time.

Table 16.1 Miss time in microsecond for ten various trace-files

| Trace files | 1-way (direct) μs Miss time | 2-way μs Miss time | 4-way μs Miss time | 8-way μs Miss time | Our approach μs Miss time |
|-------------|--------------------------------|-----------------------|-----------------------|-----------------------|------------------------------|
| Prog.1 | 3 | 2.9 | 2.9 | 2.8 | 2.9 |
| Prog.2 | 3.4 | 3 | 2.8 | 2.9 | 3.2 |
| Prog.3 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| Prog.4 | 1.9 | 2 | 1.9 | 1.9 | 1.9 |
| Prog.5 | 7.3999 | 7.2999 | 6.7999 | 6.9999 | 6.1 |
| Prog.6 | 6.8999 | 6.5999 | 5.5 | 6.1 | 6.5999 |
| Prog.7 | 7.3999 | 7.0999 | 7.0999 | 7.0999 | 6.1 |
| Prog.8 | 8.9 | 8.9 | 8.5 | 8.6 | 7.9 |
| Prog.9 | 9.7 | 9.7 | 9.7 | 10.8 | 8.6 |
| Prog.10 | 7.0999 | 6.1 | 6.4999 | 6.4 | 6.1 |

Table 16.2 Performance comparison between our approach and direct-mapped cache

| Trace files | Performance % compared to 1-way (direct mapped cache) |
|-------------|--|
| Program1 | 3.33 |
| Program2 | 5.88 |
| Program3 | 0 |
| Program4 | 0 |
| Program5 | 17.56 |
| Program6 | 4.34 |
| Program7 | 17.56 |
| Program8 | 11.23 |
| Program9 | 11.34 |
| Program10 | 14.08 |

In Table 16.1 trace programs from 1 to 10 designed to provide different cases of cache operations as following:

- The program 1 is a small program, as we can see in table above the miss time decreases when associativity increases. The table shows that our approach is better than direct mapping and gives the same time for 2-way and 4-way but 8-way gives us less miss time.
- The program 2 is a medium program has repeated value to show the effect of increasing hit rate. As we can see, the values decrease until 8-way which has a small higher difference and our approach is better than just direct mapping.
- The program 3 is a small program all value in it will map to the set zero in all mapping ways, and this program gives same values for miss time for all mapping ways applied including our approach.
- The program 4 is a small program has random values for operands addresses and it shows no change in miss time.
- The program 5 is a medium program has random values for operands addresses and shows decreasing in miss time and a good difference for our approach.

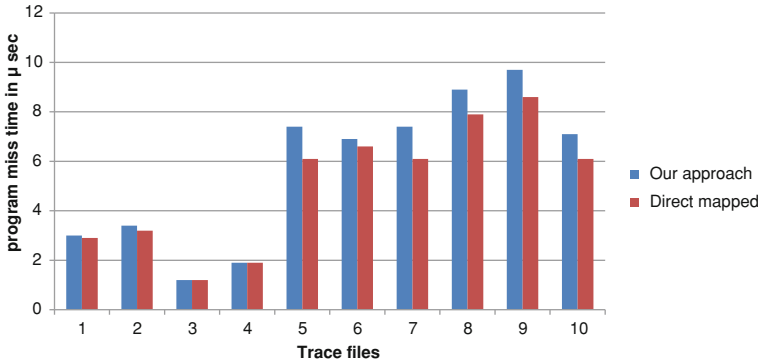


Fig. 16.2 Miss time comparison

- The rest programs are long ones and our approach gives a good optimization for miss time.

Table 16.2 shows the performance optimization percentage for our approach compared to direct mapping. As we can see, the range of our performance enhancement percentage is between 3.3 and 17.56 %, but some programs have no change.

Figure 16.2 shows comparison between our approach and direct cache (1-way) in term of total miss time in each of the ten trace programs. As we can see our approach gives better performance in all programs except two programs which is equal to direct mapped performance.

16.4 Conclusion and Future Work

In this work we have introduced cache simulator program with a new enhancement approach which enhances cache performance by reducing miss rate. The proposed work chooses the best number of cache mapping ways depending on program has been executed. In addition our algorithm is designed to maintain cache hit-time to be minimum. Evaluation results reveal the superiority of the proposed approach compared to traditional set associative technique. As a future work we will try to implement our algorithm using genetic-algorithm to increase performance.

References

1. Hennessy JL, Patterson DA (2003) Computer architecture: a quantitative approach. Morgan Kaufmann Publishers, San Francisco
2. Kharbutli M, Irwin K, Solihin Y, Lee J (2011) Using prime numbers for cache indexing to eliminate conflict misses. This work was supported in part by North Carolina State University, Seoul National University, the Korean Ministry of Education under the BK21 program, and the Korean Ministry of Science and Technology under the National Research Laboratory program

3. Kharbutli M, Solihin Y (2008) Counter-based cache replacement and bypassing algorithms, *IEEE Trans Comput* 57(4):433–447
4. Ghasemzadeh H, Mazrouee S, Moghaddam HG, Shojaei H, Kakoei MR (2006) Hardware implementation of stack-based replacement algorithms. In: *Proceedings of world academy of science, engineering and technology*, vol 16
5. Beyls K, D'Hollander EH (2008) Refactoring intermediately executed code to reduce cache capacity misses. *J Instr Lev Parallelism* 10:1–9 (Submitted 10/06; published 6/08). This article initially appeared in abbreviated form in *ACM computing frontiers* 2006
6. Solihin Y, Lee J, Torrellas J (2003) Correlation prefetching with a user-level memory thread. In: *IEEE Trans Parallel Distributed Syst* 14(6):563–580

Chapter 17

Agent Simulation Group on the Robocup 3D Realization of Basic Motions

Min Zhou, Jia Wu, Hao Zheng, Xiaoming Liu
and Renhao Zhou

Abstract In this paper, we describe the implementation mechanism of the basic movements in Robocup3D Simulation game, study the ZMP and the rotation matrix's application in the agent's action, which compared with simulation tracking data. In addition, we designed and developed a dynamic simulation software used to simulate agent body movements to help us with the in-depth analysis of the agent motion. And we also use evolutionary algorithm to optimize the motion parameters of Agent. Experiments proved that method of rotation matrix and the ZMP in agent movement calculation meet the accuracy requirements such as real-time. The software of dynamic simulation and Evolutionary Algorithm perform well in agent motion analysis and parameter optimization.

M. Zhou (✉) · J. Wu · H. Zheng · X. Liu · R. Zhou
School of Computer Science, China University of Geosciences,
LumoRoad 388, Wuhan 430074, China
e-mail: cugzhoumin@gmail.com

J. Wu
e-mail: wujiawb@126.com

H. Zheng
e-mail: rat604@gmail.com

X. Liu
e-mail: asfion.lxm@gmail.com

R. Zhou
e-mail: renhao.beta@gmail.com

17.1 Introduction

RoboCup (The Robot World Cup) is the robot soccer World Cup tournament, by definition, manufacturing and training robot to football match [1]. The most important purpose of RoboCup robot soccer is to improve the research level in the field of artificial intelligence and robotics, exchange of ideas and the new progress for better promote basic research and applied basic research and Achievements. It's ultimate dream is developed to teach human soccer players can play soccer robots in 2050.

17.2 Two Methods of Improving Agent's Motion

The traditional method of studying agent's action is Geometric, mainly. Analysis and mapping by a simple movement itself has all the characteristics (such as symmetry), the action is divided into several states, each state calculates the angle corresponding to each node and height. Although this method can be worked to some extent the results needed and to meet some of the basic agent's action, for the tedious method and computational complexity this method can only worked with some easy actions, a few state and nodes. In view of this, this paper proposes a method using rotation matrix which is of great applicability with any of agent's actions, most of states and nodes.

17.2.1 The Application of Rotation Matrix in Agent's Motion

Matrix is used very frequently in the action design of Agent, rotation of 3D points, scaling and translation. It's also popular in 2D translation. In three dimensions, rotation matrices are among the simplest algebraic descriptions of rotations, and are used extensively for computations in geometry, physics, and computer graphics. The simplest matrix is defined as a digital form. It has one or more horizontal rows and one or more vertical columns, shown in "Fig. 17.1".

More generally the matrix is used for operating 3D points. It covers a 3D point x, y, z coordinates. We can simply see it as a matrix: $[x \ y \ z]$. Suppose you want to move this point in space, or called translation of the point. Then it can be converted into a matrix. After conversion, including the new coordinates of the point. In the broader application of 3D conversion is the matrix multiplication, commonly used in the scaling and rotation. We can rotate any axis of the three axes and create a matrix for each rotation. Start with the x-axis rotation matrix:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & \sin \theta \\ 0 & -\sin \theta & \cos \theta \end{bmatrix} \quad (17.1)$$

Fig. 17.1 The relationship between coordinate vector and rotation matrix

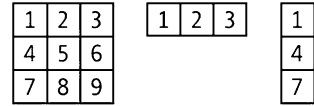
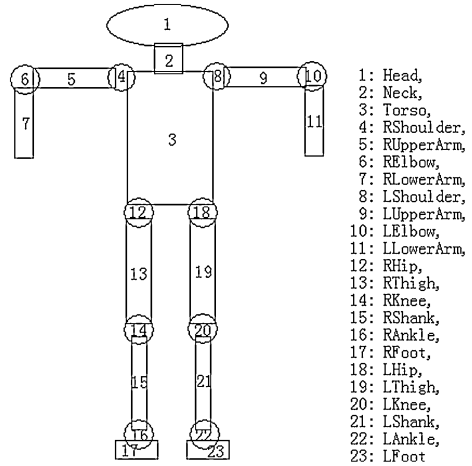


Fig. 17.2 The joints and parts of an agent in Robocup 3D group



Here are some value of sine and cosine. In the discussion of coordinate rotation; we will see that this is actually the x-axis coordinate rotation. Thus, creating a y-axis rotation matrix is very easy:

$$\begin{bmatrix} \cos \theta & 0 & \sin \theta \\ 0 & 1 & 0 \\ -\sin \theta & 0 & \cos \theta \end{bmatrix} \quad (17.2)$$

Finally, z-axis of rotation is:

$$\begin{bmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (17.3)$$

Every robot has 22 joints (every joint only have one degree of freedom) and 13 mass points. To facilitate the calculation, we define the robot as 23 nodes (see “Fig. 17.2”).

“Figure 17.3” is the result of abstracting the joints and parts in the “Fig. 17.2” into node. In the “Fig. 17.4” the Node v_i represents the corresponding of robot’s “node” we have defined in “Fig. 17.3” and the edge e_i represents the value of effect that was made by robot’s joint rotating. So we can use $e_i \cdot e_j$ to express the direct influence that side i imposes on node j and $e_i \cdot e_j$ is the value of effect that side i imposes on side j. In the initial status (all of the degree of robot’s joints angle are 0), the value of effect that node i imposes on side j is $\bar{v}_i v_j$. So, to the “Fig. 17.3”, the value v_i can be calculated by the Eq. 17.4 as follows:

Fig. 17.3 The result of abstracting the joints and parts into nodes

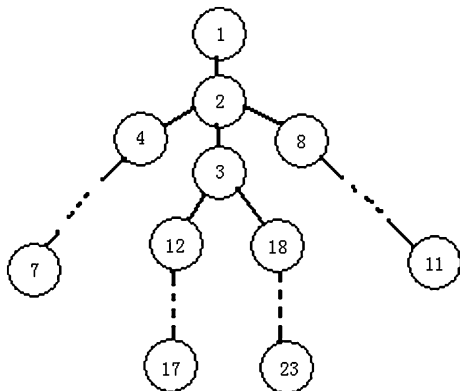
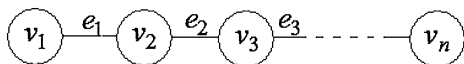


Fig. 17.4 Single line calculation of the relationship between nodes



$$v_i = v_{i-1} + \overrightarrow{v_{i-1}v_i} \cdot \sum_{k=1}^i \prod_{j=1}^k e_j \quad (17.4)$$

In Eq. 17.4, e_j is unknown. If we set the joint axis of rotation $(n_x \ n_y \ n_z)$ in the initial status and define joint angle (θ) as the angle of joint rotation, and, if we rotate robot's joint around its own axis, we can get the rotation matrix of e_j :

$$e = \begin{pmatrix} n_x^2(1 - \cos \theta) + \cos \theta & n_x n_y(1 - \cos \theta) + n_z \sin \theta & n_x n_z(1 - \cos \theta) - n_y \sin \theta \\ n_x n_y(1 - \cos \theta) - n_z \sin \theta & n_y^2(1 - \cos \theta) + \cos \theta & n_y n_z(1 - \cos \theta) + n_x \sin \theta \\ n_x n_z(1 - \cos \theta) + n_y \sin \theta & n_y n_z(1 - \cos \theta) - n_z \sin \theta & n_z^2(1 - \cos \theta) + \cos \theta \end{pmatrix} \quad (17.5)$$

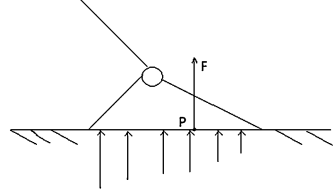
In Eq. 17.5, e_j is known. The relative coordinate in the coordinate system which based on robot's direction can be obtained. Besides, because we have known v_i , by using it in the Eq. 17.6, we can calculate the coordinate of robot's gravity coordinate

$$M_0 = \frac{\sum m_i v_i}{\sum m_i} \quad (17.6)$$

where, M_0 is robot's gravity, m_i is the mass of part i of robot body, v_i is the coordinate of m_i .

Calibrating the result by the direction of robot's body, we can get coordinate of every part of body as well as gravity coordinate. These coordinates are playing very important roles in calculation of robot's action and it is a sign of changing from qualitative computing to precise computing.

Fig. 17.5 Force diagram of agent's foot



17.2.2 Introduction of ZMP

The zero moment point is a very important concept in the motion planning for biped robots and it can be used as a ruler to guarantee the dynamical postural stability of the robot [2]. This concept was introduced in January 1968 by Miomir Vukobratovi at The Third All-Union Congress of Theoretical and Applied Mechanics in Moscow. It can be defined as follows: in “Fig. 17.5”, the upward resultant force from the bottom of the robot's feet is F , P is a sole point that lies in the normal projection of the ankle and ZMP is the point of P .

Vukobratovi describes the stability domain as the projection that the convex region where is made by the non-kicking foot project in the horizontal direction [3]. During one foot supporting robot's body, the projection area is the area of non-kicking foot; during the two feet supporting robot's body, the area is the maximal convex region where is made up by the two feet with ground. To facilitate the description, we define xOy as the plane of ground, besides, z -axis is perpendicular to the ground in an upward direction. So F is the reluctant force in the z -axis, if the value of moment of force of force F is zero, this moment of force point is Zero Moment Point. If ZMP between the area of the supporting feet, the current postural of robot is stable.

During robot walking, there are only normal force, friction force and gravity, all this is a prerequisite for calculating ZMP. From the definition of ZMP, ZMP can be written as follows:

$$p_x = \frac{\int_{x_1}^{x_2} \xi \rho(\xi) d\xi}{\int_{x_1}^{x_2} \rho(\xi) d\xi} \quad (17.7)$$

$$p_y = \frac{\int_{y_1}^{y_2} \xi \mu(\xi) d\xi}{\int_{y_1}^{y_2} \mu(\xi) d\xi} \quad (17.8)$$

The range of force is (x_1, x_2) and (y_1, y_2) and the current force are $\rho(x)$ and $\mu(y)$. In order to facilitate the calculation, we unify all the coordinates to a relative coordinate system which takes robot's head as origin and absolute coordinate system's axis as axis, to realize this process, the only thing we do is to rotate the matrix.

Before the robot falls down, there must be an unstable state. In the process of this state, ZMP's horizontal projection usually beyond the domain of the convex region

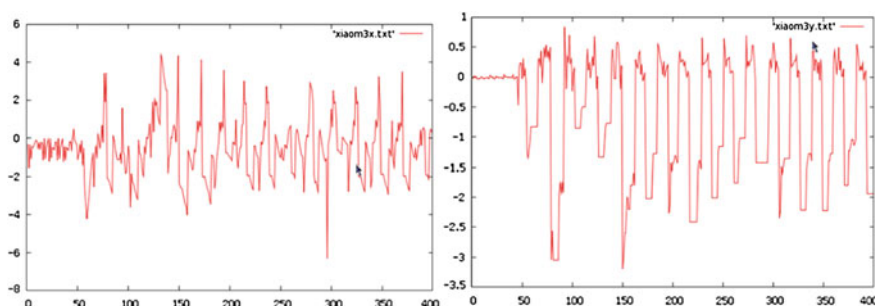


Fig. 17.6 ZMP curve with the step length is 3. The motion is more stable and the ZMP curve is more regular

of feet. From this reason, we can use ZMP to judge whether the robot will be fall down. If we confirm the robot has fell down then we use GYR and other parameter to judge the current postural of the robot.

Generally speaking, the smaller the variance of ZMP the more stable movement of robot (see “Figs. 17.6, 17.7”).

By Adding reasonable rotation of robot’s arms and legs, it can counteract some noise which have a effect on robot’s movement. This is a research interest of our team.

17.3 Two Test Methods of Robot’s Action

17.3.1 *The Simulation Software of Robot’s Body Movement*

17.3.1.1 The Background of Development

To create a good robot motion system is a difficult task. It is extremely abstract and difficult for developer to develop a new movement if there is no model. The developer will be disorientation and don’t know what to do next because they cannot see the movement of robot’s body clearly. Giving inspiration and displaying the image of movement to the developer will greatly reduce the difficult of the development of movement if there is a software that can simulate the robot’s body movement and show the changes in the joint parameters. Therefore, “CUG3D-TEST” test software was born.

17.3.1.2 Software Description

The software’s development is based on the “MFC” library. It has been able to meet the basic test need. When a new idea was born, the developer of movement of

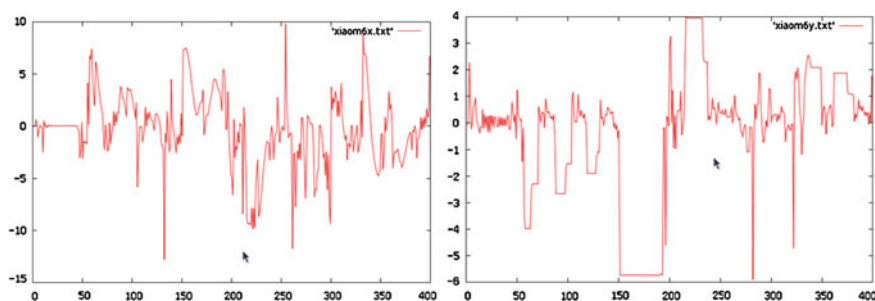


Fig. 17.7 ZMP curve with the step length is 6. The motion is not stable and the ZMP curve is irregular

robots can use the software to simulate his new idea, to see if it is reasonable and feasible, to avoid blind development, and enhancing the image of the virtual and greatly increased the speed of action development. However, the software has some drawbacks, such as the appearance of landscaping, extended to three-dimensional, more human and so on. The software is developed for our group's test of movement function. With the group's growth, we will expand to three-dimensional simulation and provide customer functionality version.

17.3.1.3 Concrete Realization of the Software

At present the software use two-dimensional graphics to display changes in the robot's body and the output parameter to quantitatively determine the robot's physical state. To simulate the leg movement as an example now: Run CUG3D-TEST, in the corresponding dialog box, enter the state of parameters to achieve, and press the "GO" button, the robot will reach its corresponding physical state. Then press "MOVING" to dynamic display details of their physical movement. Also we can choose to output the specific movement parameters and movement status.

Set the hip joint as the fixed point, use the following formula:

$$\begin{cases} x_1 = x_0 + shank \cdot \sin(Lleg[3] - bodyAng) \\ y_1 = y_0 + shank \cdot \cos(Lleg[3] - bodyAng) \end{cases} \quad (17.9)$$

Calculate the precise coordinates of the knee, followed by the same token delivery and the use of each type 02, 03, 04, 05 can accurately calculate the coordinates of the ankle, heel and toe. Use the different parameters when the function is called each time. Using the above formula to calculate each node's coordinates. Then display it with brush. Equal time to time intervals, the simulation of the robot's movement with the software is completed.

$$\begin{cases} x_2 = x_1 - shank \cdot \sin(Lleg[4] + bodyAng - Lleg[3]) \\ y_2 = y_1 + shank \cdot \cos(Lleg[4] + bodyAng - Lleg[3]) \end{cases} \quad (17.10)$$

$$\begin{cases} x_3 = x_2 - footB \cdot \sin(\frac{\pi}{2} + Lleg[3] - bodyAng - Lleg[4] + Lleg[5]) \\ y_3 = y_2 - footB \cdot \cos(\frac{\pi}{2} + Lleg[3] - bodyAng - Lleg[4] + Lleg[5]) \end{cases} \quad (17.11)$$

$$\begin{cases} x_4 = x_3 + footT \cdot \sin(\frac{\pi}{2} + Lleg[3] - bodyAng - Lleg[4] + Lleg[5]) \\ y_4 = y_3 + footT \cdot \cos(\frac{\pi}{2} + Lleg[3] - bodyAng - Lleg[4] + Lleg[5]) \end{cases} \quad (17.12)$$

From the robot dynamic simulation of movement, we can clearly see the movement of the robot's body to determine whether the action is coordinated with the ground with unnecessary friction. If there is unnecessary friction, we can change the movement function according to actual situation.

17.3.2 Evolutionary Algorithm

17.3.2.1 The Outline of Evolutionary Algorithm

Evolutionary algorithms include genetic algorithms, evolution strategies, evolutionary programming and genetic programming four branches [4]. It is based on the natural world and it is not necessary to describe the problem's all characteristics clearly, only under the laws of nature to produce new and better solution. Evolutionary algorithm is an common algorithm of this idea.

Nature is a source of human inspiration. Centuries, use the answer provided by the biosphere to the application to practical problems has proven to be a successful method, and a specialized branch of science of bionics was formed. As is known to all, the answer given by nature is through a long process of the formation of adaptive(The process is called evolution).In addition to the final outcome of evolution, we can also use it to solve some issues which is more complex. Therefore, we need not describe all the characteristics of the problem clearly, only under the laws of nature to produce new and better solution. Evolutionary algorithm is an common algorithm of this idea. Due to its nature of parallel, self-organizing, adaptive and intelligent self-learning features, evolutionary computation has been successfully applied to solving complex problems that is difficult to solve with traditional methods. This paper, we use the genetic algorithm.

17.3.2.2 Concrete Realization of Evolutionary Algorithm

This article focuses on the application of GA in the determination of parameters when robot rotation spot, other sports such as: walk, turn while walking, shoot etc., those should take a further study. Robot rotation involves three parameters:

Degrees of rotation per week: “RobotAngle”; The height of leg lift: Height; Horizontal distance between the legs: YDistance. Main steps are as follows:

Coding:

For any variable $x \in (a, b)$, we can write it as $x = \alpha + \mu \cdot (b - a)$. Among it $\mu \in (0, 1)$. For any i -bit binary, expressed as $(x_1 x_2 \dots x_i \dots x_l)$, then its corresponding decimal number $x_d = \sum_{i=1}^l x_i \cdot 2^{i-1}$. Order $\mu = \frac{x_d}{2^l - 1}$, then we can get the coding conversion between number systems. As the platform's real-time, in order to improve the speed of calculation, we coded with 8.

Initialize population:

Then produce an individual scale, its range is $r_0 \in [0, 20]$, $Height \in [-21, -13]$, $YDistance \in [0, 4]$, this is coded decimal value $x_d \in [0, 255]$.

Crossover:

To reduce the computation, the algorithm uses single-point crossover fashion. Different individuals with different crossover probability: For individuals with higher fitness, they have a high crossover probability on the basis of retention of individual gene. For individuals with lower fitness, they have a lower crossover probability, and there is a certain probability of being eliminated.

$$p_{cl} = \begin{cases} P'_{cl}, f \geq \bar{f} \\ P''_{cl}, f < \bar{f} \end{cases} \quad (17.13)$$

where P'_{cl} , P''_{cl} is adjustable parameters that need to set, $P'_{cl} > P''_{cl}$, f is the individual fitness, \bar{f} is the average individual fitness.

Mutation:

Variation is an important part of biological evolution, it can increase the diversity of groups, avoid premature. “Genetic Search” will become a “random search” if there is too much variation rate, and this will undermine the quality of individual. Some loci may be too early to lose information and cannot be restored if the mutation rate is too small. High mutation rate of the number should be small, because they are more sensitive to the values. For those good race, low figure is a critical range, and their mutation rate should be higher. Therefore, we take “ i ” bit mutation rate “ σ_i ” to make the following calculation:

$$\sigma_i = \sigma_{\min} + \frac{n - i}{n - 1} (\sigma_{\max} - \sigma_{\min}) \quad (17.14)$$

Among it σ_{\min} and σ_{\max} are adjustable parameters, n is the total length of loci. We set σ_{\min} is 0.01, set σ_{\max} is 0.1.

Fitness Function:

As the rotational speed and rotational stability are the main criteria for the robot rotation. We use the fitness function to judge as follows:

$$S_i = \Delta\theta_i - C_1 \cdot t_i - C_2 \cdot \Delta d_i \quad (17.15)$$

where, it s_i represent the i th gene's fitness, $\Delta\theta_i$ represent the i th test's total angle, t_i represent i th test's total number of fall, Δd_i represent i th test's offset distance, C_1 ,

C_2 are adjustable parameters. In the course of our practice, we set $C_1 = 200$, $C_2 = 400$.

Experimental results and analysis:

The following table is based on the above experimental methods and it's parameters, the result of the twentieth generation:

| Score | ro | Height | YDistance | Fitness |
|-------|---------|----------|-----------|---------|
| 1 | 17.8203 | -17.9688 | 3.5625 | 562.948 |
| 2 | 12.6172 | -18.0625 | 2.625 | 467.887 |
| 3 | 8.82031 | -17.2812 | 2.01562 | 368.896 |
| 4 | 15.7109 | -17.1562 | 1.85938 | 350.397 |
| 5 | 6.92188 | -14.6875 | 1.84375 | 318.626 |

17.4 Conclusion

In this paper, we use the rotation matrix method, which greatly making the calculation of agent's motion more simple and agent more humanoid, to achieve the change of manual testing to accurate calculation. The use of ZMP has good performance in judgments of agent's fall and determination of agent's stability. The dynamic simulation software, which achieves the change from digital to visualization and shorten the development cycle, is beneficial for depth analysis of agent's motion. The use of evolutionary algorithms, which embodies the intelligence, makes computing more convenient and rapid and improves the convergence accuracy greatly. In addition, all results described in this article are independent research by our group members, which fully embodies the innovative.

References

1. Craig JJ (2006) Introduction to robotics. Press of Mechanical, Beijing (in Chinese)
2. Xiao N (2008) Humanoid. Press of Science, Beijing (in Chinese)
3. Kajita S (2007) Humanoid. Press of Tsinghua University, Beijing (in Chinese)
4. Tan M, Wang S, Cao Z (2007) Multi-robot systems. Press of Tsinghua University, Beijing (in Chinese)

Chapter 18

Key Generations Model for Mobile Cryptosystems

Rushdi Hamamreh

Abstract Mobile computers and devices may operate in a variety of environments with different security schemes. In this paper we proposed cryptosystem for mobile applications as Short Message Service (SMS) based on block cipher and Hill cipher methods. Many cryptosystems were designed to prevent data from unauthorized access, and some are relatively secure but slow. Others are fast but relatively not secure enough. One of the most efficient cryptosystems is Hill Cipher algorithm which is classified as symmetric encryption. In this paper we provide a solution for the problem of non-invertible matrix by modifying the way of dealing with key matrix, and make all matrices; including non- invertible ones, usable in modified Hill cipher system. Moreover, it will solve the known of pair plaintext and cipher text problem by generating new key matrix for each encrypted block of plaintext, using SHA-512. Since SHA-512 generates 64 integers we can manipulate these integers to become 128 different integers and use them as an input for the matrix; based on the concept that any acceptable data must not be prime.

18.1 Introduction

With the rapid development of mobile networks technology and popularization of mobile device, people can access Internet by mobile device and wireless connection covering the entire mobile communication network (GSM/GPRS/3G/802.11etc) at any moment [1]. Compare with traditional system, the security risk of system based on mobile network is more grave. However, the traditional mobile

R. Hamamreh (✉)

Computer Engineering Department, Al-Quds University, Jerusalem, Palestine
e-mail: rhamamreh@eng.alquds.edu

communication technology does not provide the security services such as authentication, confidentiality, and integrity etc. [1, 2]. To solve this security problem, in this paper, we designed and implemented a mobile cryptosystem using Hill cipher. Cryptography is a mixture of mathematics and computer science. It is the study and the ability of hiding data. Cryptography has increasingly been used to secure information. But secure data of today could be broken in the future. Cryptography is the science of codes and ciphers. It includes many algorithms and techniques that transfer data safely. One of them called Hill cipher, Hill Cipher algorithm is not widely used despite of its linear nature, simplicity and ease of use. Hill Cipher is not widely used since it is easy to know the secret key if pair of plaintext and a cipher text is known [3–5].

In addition, Hill Cipher has a problem of non-invertible matrices; not only the zero determinant Matrices but all non prime determinant matrices relative to modular value. Hence, the unreliability of the system, because of the two previous problems Hill Cipher not widely used [6]. In section four of this paper, we try to make Hill Cipher usable for all determinants in our system. It's important to notice that we have two methods to overcome all Hill Cipher problems. First solve the problem of non invertible matrices which enables us to use the second method.

The rest of this paper is organized as follows. Section first briefly discusses the difference between symmetric and asymmetric cryptosystem. Section two gives brief introduction about Hill Cipher. Section three introduces disadvantages of Hill Cipher through different examples. Section four explains the solution for the Hill Cipher main problem non invertible matrices. Section five explains how to create secret key for every encryption to prevent key discovery. Section sixth describes results and comparison. We conclude our work in section seven.

18.2 Symmetric Versus Asymmetric Encryption

Encryption systems are divided into two main categories, symmetric and asymmetric. Symmetric encryption, also known as secret key or single key, the same key that sender uses to encrypt the data and to decrypt it by the receiver on the other side [7–9]. This system was the only system used earlier to the discovering and developing the public key. In symmetric encryption, a safe way of data transfer must be used to move the secret key between the sender and the receiver [9–11]. Symmetric encryption occurs either by substitution or transposition technique, or by a mixture of both techniques [12].

Symmetric encryption has many advantages over asymmetric in many ways. First, it is faster since it doesn't consume much time in data encryption and decryption. Secondly, it is easier than asymmetric encryption in secret key generation [13]. However, it has some disadvantages, for example, key distribution and sharing of the secret key between the sender and the receiver. Thus, symmetric encryption can achieve a good system performance while asymmetric encryption can provide a high level of security [14, 15].

Asymmetric encryption is the opposite of symmetric encryption in safety, since it doesn't require the sharing of the secret key between the sender and the receiver. The sender has the public key of the receiver. The receiver has his own secret key which is extremely difficult or impossible to know through the public key [14–16]. Asymmetric key can use either the public or secret key to encrypt the data. Also it can use either key in decryption. But asymmetric encryption is slower and very complicated in calculations [15]. Therefore, the nature of the data determines the system of encryption. And every system has own uses. For example, asymmetric encryption may be used in authentication or in sending secret key for symmetric systems [13].

18.3 Block Versus Stream Cipher

Idea of a block cipher: partition the text into relatively large (e.g. 128 bits or 16 integers) blocks and encode each block separately. The encoding of each block generally depends on at most one of the previous blocks.

- the same “key” is used at each block.

Idea of a stream cipher: partition the text into small (e.g. 1 bit) blocks and let the encoding of each block depend on many previous blocks.

- for each block, a different “key” is generated [17].

18.4 Hill Cipher

Hill cipher is an application of modular linear algebra to cryptology [3]. Many researches and papers tried to use Hill cipher algorithm to build a comprehensive cryptosystem, because it has many advantages; it's simple and easy since it uses multiplications of matrices. It's also fast and highly productive also it is very strong substitution technique against a cipher-only attack [18, 19].

However, it has two compound problems in which the second one indirectly depends on the first one. The first problem is that Hill Cipher requires an inverse of each matrix, used in order to decrypt all the matrixes used in the encryption side. And many matrices have no inverse. Therefore, the secret key can't be neither randomly nor mathematically produced as there will be uncertainty of the key validity. In case the key remains constant during the encryption process, it will be easy for the hacker to get it; once he gets a pair of plaintext and cipher text, and this is the second problem [16].

Hill Cipher was invented by Lester S. Hill in 1929 [20, 21]. It's considered as a kind of monoalphabetic polygraphic substitution cipher. It uses the algebraic method. It's also a good example of encrypting data in blocks since it encrypts a group of characters at once. The idea of Hill Cipher is matrices multiplications in

which every character or group of characters in the plaintext is substituted by a character or a group of characters in the cipher text. Each character is assigned to a numerical value [4–7].

To encrypt a block consists of n characters, we need $n \times n$ matrix. During the decryption process, we need the inverse of the matrix used in the encryption. It's important to notice that the inverse of the matrix is calculated depending on “P”, since that the matrices that have inverse are those that have prime determinant relative to modular value “P” [4–7, 17].

The encryption and decryption processes occur through the following mathematical equations.

At encryption side $c = k \times x \bmod p$, where c is the cipher text, x is the plaintext, k is the key matrix, and p is the modular value.

At decryption side $x = k^{-1} \times c \bmod p$

18.5 Hill Cipher Problems

The first problem of Hill cipher is none invertible matrices; since the encrypted text can't be decrypted [10, 14]. Also when the matrix not invertible, two plaintext vector will be mapped into the same cipher text vector. Let us consider the following example.

Through changing the key matrix from the previous example $k = \begin{bmatrix} 1 & 2 \\ 2 & 8 \end{bmatrix}$, then determinant of this key is four, not prime relative to $p = 26$, so no inverse can be found for this key. Also let us have the following two pairs of plaintext:

$$p_1 = \begin{bmatrix} 2 \\ 19 \end{bmatrix}, p_2 = \begin{bmatrix} 2 \\ 6 \end{bmatrix}.$$

$$c_1 = \begin{bmatrix} 1 & 2 \\ 2 & 8 \end{bmatrix} \times \begin{bmatrix} 2 \\ 19 \end{bmatrix} \bmod 26 = \begin{bmatrix} 14 \\ 0 \end{bmatrix}$$

$$c_2 = \begin{bmatrix} 1 & 2 \\ 2 & 8 \end{bmatrix} \times \begin{bmatrix} 2 \\ 6 \end{bmatrix} \bmod 26 = \begin{bmatrix} 14 \\ 0 \end{bmatrix}$$

So $c_1 = c_2$.

If the receiver decrypt this vector (c_1 or c_2), the problem will be to determine from which plaintext vector; p_1 or p_2 , it came from. This is a problem since the matrix determinant is not prime relative to the modular value (26 in this example).

A second problem of Hill Cipher is the known-plaintext attack. Due to Hill Cipher linear nature, the cryptosystem can be broken through the known plaintext attack [22]. An analyzer knows only two pairs of plaintext-cipher text, and then the key matrix can be calculated, from the following equations.

$$\begin{bmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{bmatrix} \times \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} \bmod 26 = \begin{bmatrix} c_{11} & c_{21} \\ c_{12} & c_{22} \end{bmatrix}$$

Example, let $p_1 = \begin{bmatrix} 9 \\ 15 \end{bmatrix}$, $p_2 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$,

$$c_1 = \begin{bmatrix} 15 \\ 1 \end{bmatrix}, c_2 = \begin{bmatrix} 23 \\ 14 \end{bmatrix},$$

Then from the above equation we can calculate key matrix

$$\begin{bmatrix} 9 & 3 \\ 15 & 6 \end{bmatrix} \times \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} \bmod 26 = \begin{bmatrix} 15 & 23 \\ 1 & 14 \end{bmatrix}$$

$$\begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} = \begin{bmatrix} 18 & 17 \\ 7 & 1 \end{bmatrix} \times \begin{bmatrix} 15 & 23 \\ 1 & 14 \end{bmatrix} \bmod 26$$

Then the calculated key is $k = \begin{bmatrix} 1 & 2 \\ 2 & 19 \end{bmatrix}$.

18.6 Invertible Matrices

To make all matrices invertible there are two chooses. First proposed method depend on convert every two characters in encryption side, into three characters in decryption side. At encryption side, the key matrix is used, while at decryption side the normal inverse of key matrix is used. This technique requires some restriction on the maximum value allowed in the key matrix. The second method does not have any restriction on the values of key matrix, but every two characters at the encryption side will convert into four characters in the decryption side.

18.6.1 First Method

- 1 Check if the determinant of the key matrix is zero. If so, add identity matrix, else do nothing. Convert the two vector of plaintext into one numerical value.
- 2 Calculate the three cipher text vectors from the following equations:

$$c_T = k \times (p_1 \times n + p_2)$$

$$c_1 = c_T \bmod n$$

$$c_2 = \text{int}(c_T/n) \bmod n$$

$$c_3 = \text{int}(c_T/n^2) \bmod n$$

- 3 Convert the numerical values into characters.

At the decryption side:

- 1 Check if the determinant of the key matrix is zero. If so, add identity matrix, else do nothing.
- 2 Convert the three vectors of cipher text into numerical values.
- 3 Calculate the two plaintext vectors from the following equations.

$$c_T = c_1 + (n \times (c_2 + (n \times c_3)))$$

$$x = k^{-1} \times c_T$$

$$p_1 = \frac{x}{n} \text{ and } p_2 = x \bmod n$$

- 4 Convert the numerical values into characters.

In this method, numerical values of key have small restriction which is discussed in the following section.

18.6.2 Key Space of Matrices for First Method

$$\text{Let } k = \begin{bmatrix} k_{11} & k_{12} & \cdots & \cdots & k_{1n} \\ k_{21} & k_{22} & \cdots & \cdots & k_{2n} \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ k_{n1} & k_{n2} & \cdots & \cdots & k_{nn} \end{bmatrix}$$

$$p_1 = \begin{bmatrix} p_{11} \\ p_{12} \\ \vdots \\ p_{1n} \end{bmatrix}, p_2 = \begin{bmatrix} p_{21} \\ p_{22} \\ \vdots \\ p_{2n} \end{bmatrix}$$

And the modular value is p . Assume we have the worst case; in this case the values of plaintext vectors are $p - 1$. The problem is to find the values accepted to act as key matrix element. Also assume these elements are also at the worst case scenario; are equal to each other and are equal to y .

$$c_T = k \times (p_1 \times n + p_2)$$

$$c_T = \begin{bmatrix} y & y & \cdots & \cdots & y \\ y & y & \cdots & \cdots & y \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ y & y & \cdots & \cdots & y \end{bmatrix} \times \left(\begin{bmatrix} p-1 \\ p-1 \\ \vdots \\ p-1 \end{bmatrix} \times p + \begin{bmatrix} p-1 \\ p-1 \\ \vdots \\ p-1 \end{bmatrix} \right)$$

$$c_T = n \times y \times (p^2 - 1)$$

To use the previous technique, the following equation must hold:

$$c_T < p^3$$

So we can calculate the maximum value of y .

$$n \times y \times (p^2 - 1) < p^3$$

$$y < \frac{p^3}{n \times (p^2 - 1)}.$$

Where p is the modular value and n is the matrix size.

18.6.3 Second Method

This method is similar to the first one, except that every two characters from encryption side convert into four characters in the decryption side, but no restriction on the key space values.

At the encryption side the equations are:

$$c_T = k \times (p_1 \times n + p_2)$$

$$c_1 = c_T \bmod n$$

$$c_2 = \text{int}(c_T/n) \bmod n$$

$$c_3 = \text{int}(c_T/n^2) \bmod n$$

$$c_4 = \text{int}(c_T/n^3)$$

At the decryption side the equations are:

$$c_T = c_1 + (n \times (c_2 + (n \times (c_3 + n \times c_4))))$$

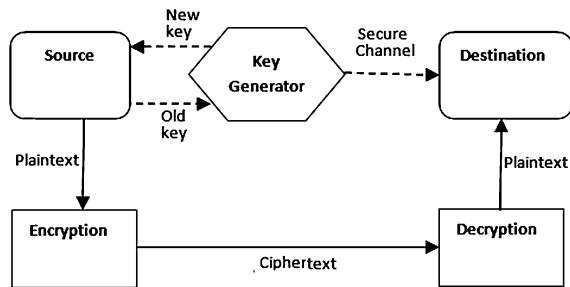
$$x = k^{-1} \times c_T$$

$$p_1 = \frac{x}{n} \text{ and } p_2 = x \bmod n$$

18.7 Multi Key Generation

18.7.1 Secure Hashing Algorithm SHA-512

Since the key matrix accepts any data that makes the determinant not prime relative to the modular value, we can use SHA-512 that generate 64 integers [23].

Fig. 18.1 Key generation

If the result of SHA-512 modulus to modular value, we get 64 integers that is a key matrix element. If modules these integers to the square value of modular value, divide the result with modular value, and cut the remainder, we get new different 64 integers, that can be used as input to generate key matrix element. Using SHA-512 adds extra overhead time to encryption and decryption time. If the key matrix size less than 4×4 , then every time SHA-512 called, four matrix can be formulated; which means for every four blocks of data encrypted, the SHA-512 required to call (save 25 % of encryption and decryption time). But when the key matrix size increase more than 9×9 , the time required by Mousa-Rushdi-Hill Cipher (MRHC) is more than Advanced Encryption Standard (AES). This extra time between AES and MRHC remains the same for any matrix more than 9×9 .

18.7.2 Generating New Key

We has been proposed to generate “new key” next steps:

- 1 Send secure 128 bit using secure channel.
- 2 Use this secret key to generate 128 integers using SHA-512, and save the result in array.
- 3 Check if use all elements of the array matrix, then merge the element of array to generate new 128 bit secrete key and call SHA -512 using this new secret key, else use the reset of element to generate the new key matrix.
- 4 Check if the determinant of the key matrix zero, if so, adds the identity matrix.
- 5 If using the first technique develop the following equation.

$$k = k \bmod (y + 1)$$

- 6 Repeat steps 3 through 5 for each block(s), if required (Fig. 18.1).

18.8 Results and Comparison

We make three comparisons between original Hill cipher, second technique of modified Hill Cipher (MRHC) and AES. Figure 18.2 is when key matrix size is

Fig. 18.2 Encryption and decryption time for matrix size 4×4

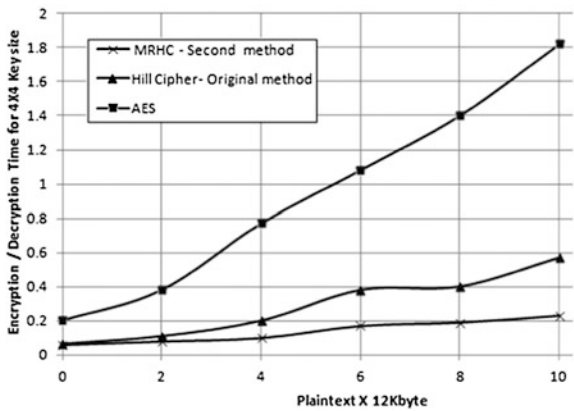
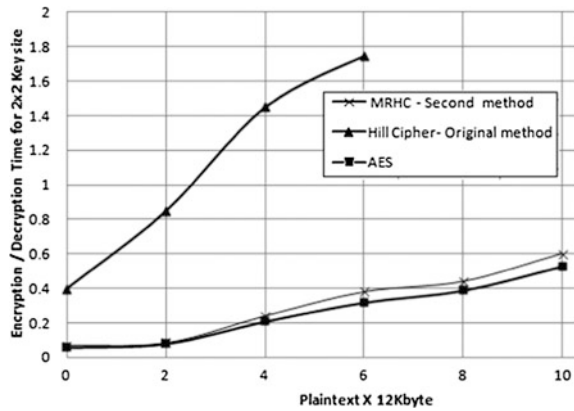


Fig. 18.3 Encryption and decryption time for matrix size 9×9



4×4 and data of plaintext ranged from 12 into 58 KB. Figure 18.3 is when key matrix size is 9×9 . From size 9×9 and up, the MRHC take time more than AES.

18.9 Conclusion

This paper introduced a new method to overcome the non-invertible matrices problem in Hill Cipher. Through the introduced solution, it will be possible to overcome the known plaintext attack.

One of the main advantages for our work is the capability of using any matrices as a key to Hill Cipher algorithm including zero determinant matrices. So, there will be no restriction on key selection. The process of generating a new key for every transmitted block of data makes the algorithm more secure. As generating a

new key doesn't have a mathematical inverse method, using SHA-512, it's extremely difficult for the hacker to calculate the key.

When using SHA-512 for generating new key matrix, if the key matrix is less than or equal to 8×8 , then the modified Hill Cipher MRHC takes less time than AES and for all matrices size less than original Hill Cipher, otherwise the time on MRHC will increase over AES.

References

1. Yu D, Chen N, Tan C (2009) Design and implementation of mobile security access system (MSAS) based on SSL VPN. IEEE first international workshop on education technology and computer science, vol 2. pp 151–155
2. Liao H-C, Lee P-C, Chao Y-H (2007) A location-dependent data encryption approach for enhancing mobile information system security. IEEE the 9th international conference on advanced communication technology, vol 1. pp 625–628
3. Tran BN, Nguyen TD (2008) Modular matrix cipher and its application in authentication protocol. In: Proceedings of the 2008 ninth ACIS international conference on software engineering, artificial intelligence, networking, and parallel/distributed computing. pp 318–323
4. Panigrahy SK, Acharya B (2008) Image encryption using self-invertible key matrix of Hill cipher algorithm. 1st international conference on advances in computing, 21–22 February 2008. pp 1–4
5. Yeh YS, Wu TC, Chang CC, Yang WC (1991) A new cryptosystem using matrix transformation. 25th annual 1991 IEEE international carnhahan conference on security technology, 1991. pp 131–138
6. Acharya B, Jena D (2009) Invertible, involuntary and permutation matrix generation methods for Hill cipher system. In: Proceedings of the 2009 international conference on advanced computer control, 2009. pp 410–414
7. Acharya B, Patra Sk (2008) A novel cryptosystem using matrix transformation. First international conference on emerging trends in engineering and technology, 2008. pp 77–81
8. Gordon AD, Jeffrey A (2004) Types and effects for asymmetric cryptographic protocols. *J Comput Secu* 12:435–483
9. Simmons GJ (1979) Symmetric and asymmetric encryption. *ACM Comput Surv* 11:305–330
10. Obimbo C, Salami B (2007) A parallel algorithm for determining the inverse of a matrix for use in block cipher encryption/decryption. *J Supercomput* 39(2):113–130
11. Dzung D, Crevatin M (2005) Security for industrial communication systems. *Proc IEEE Special Issue Ind Commun Syst* 93(6):1152–1177
12. Schneier Bruce (1996) *Applied cryptography*, 2nd edn. John Wiley, New York
13. Elbirt AJ, Paar C (2005) An instruction-level distributed processor for symmetric-key cryptography. *IEEE Trans Parallel Distributed Syst* 16(5):468–480
14. Acharya B, Rath GS (2008) Novel modified Hill Cipher algorithm. International conference on emerging technologies and applications in engineering, technology and sciences, Rajkot, 13–14 January 2008. pp 126–130
15. Lee SH, Choi L (2008) Accelerating symmetric and asymmetric ciphers with register file extension for multi-word and long-word operation. In: Proceedings of the 2008 international conference on information science and security, 2008. pp 102–107
16. Ismail IA, Amin M, Diab H (2007) How to repair the Hill cipher. *J Zhejiang Univ Sci A* 7(12):2022–2030 (Zhejiang University Press, co-published with Springer)

17. Stallings W (2006) *Cryptography and network security principles and practices*, 4th edn. Prentice Hall, Upper Saddle River
18. Hadi AS, Mahdi AH (2009) Encrypted block code. *Aust J Basic Appl Sci* 10:1315–1318
19. Sastry VUK, Shankar NR (2007) Modified Hill Cipher with interlacing and iteration. *J Comput Sci* 3:854–859
20. Hill LS (1929) Cryptography in an algebraic alphabet. *Am Math Mon* 36(6):306–312
21. Hill LS (1931) Concerning certain linear transformation apparatus of cryptography. *Am Math Mon* 38(3):135–154
22. Romero YR, Garcia RV (2008) Comments on how to repair the Hill cipher. *J Zhejiang Univ Sci A* 9(2):221–214
23. Passing M, Dressler F (2006) Practical evaluation of the performance impact of security mechanisms in sensor networks. The 31st annual IEEE conference on local computer networks, 14–16 November 2006. pp 623–629

Chapter 19

Development of Stakeholder Oriented Corporate Information Security Objectives

Margareth Stoll

Abstract Information, asset and technology are key differentiators for modern organizations. They are faced with a wide, complex and increasingly faster changing range of most different information security requirements of diverse stakeholders, huge potential threats and socio-organizational challenges. Clear, coherent corporate security objectives are required to proper guide and control information security in an organization, to demonstrate information security governance and compliance and to concentrate all security efforts on what matters most. In accordance to action research we develop an innovative stakeholder's oriented process to develop corporate information security objectives: we identify all stakeholders, analyze their needs, deduce security requirements and define the security objectives with priorities and relationships. Based on our research results this process promotes a newly holistic, collaborative, systemic, structured and market driven strategic security approach. In that way information security becomes a new role as success factor or business opportunity to provide enhanced business value and competitive edge.

19.1 Introduction

Due to globalization and ever stronger competition information management and supporting technologies have become key asset and main performance driver for continual innovation and sustainable success. Organizations and their information and technology are faced with security threats from a wide range of sources,

M. Stoll (✉)

University of Innsbruck, Technikerstr. 21a, 6020 Innsbruck, Austria

e-mail: margareth.stoll@uibk.ac.at

including computer-assisted fraud, espionage, sabotage, vandalism, fire or flood. Mobile and cloud computing, off-shoring, social networks, the increasingly interconnected, flexible and virtualized business complexity and dependency are still great challenges for information security. Organizations have to meet also many different legal and regulatory requirements, such as data protection, sound and integer financial practices and internet crime. Lack of information security compliance may result in loss of confidence of customers, partners and shareholders, as well as severe civil and criminal penalties for board members [1]. In this respect information security promotes organizations' success and is integral part of good corporate governance [2–5]. More than 12.934 organizations worldwide have just implemented an information security management system in accordance with ISO/IEC 27001 [6].

Information security for long time was seen fundamentally as an only technical job and integral part of the IT department [2, 3, 7]. The prevailing strategic approach to information security today is the risk management approach [8, 9]. It begins with the identification of assets, threats, and vulnerabilities, followed by risk assessment and the establishment and implementation of counter-measures, such as the use of passwords, antivirus software, firewalls, encryption and others.

But security problems are complex and require also a socio-organizational and human related approach [3, 7, 10–12]. There is a wide, complex and increasingly faster changing range of most different security requirements of diverse stakeholders and huge potential security threats. Organizations have to concentrate all security efforts on what matters most to fulfill unique, enterprise-specific stakeholders', legal and business security requirements and compliance. Also the baseline approach, which implements widely used controls and security standards cannot fulfill effectively and sufficiently the organization specific security requirements. To proper guide security policies and measures, control information security effectiveness, increase accountability and to demonstrate compliance and governance they need firstly to develop their strategic information security objectives and goals [2, 13–15].

Despite these requirements, to the best of our knowledge we found only few approaches for the development of an information security strategy. There is a major concern for information system strategy, which regard in part also information security [16]. Empirical evidence of the practical usefulness of the approaches is missing. As we need more research for information security management topics in general [7, 17], we need also more research focus on organization specific, corporate information security objective development.

The creation of an organization's information security strategy is commonly started by ensuring that the organization's business strategy and mission are met without security problems [2–5]. To focus all information security efforts on enhanced business value and compliance only threat prevention or legal/regulatory compliance is not sufficient.

Information security must provide added value for all stakeholders (such as shareholders, customers, collaborators, suppliers/partners, society). Instead to start from corporate objectives and to align security objectives to them; we start for the

first time from most different stakeholders' expectations and establish information security objectives, which promote stakeholders' satisfaction and competitive edge. In that way we expect to provide information security as enhanced business value for all stakeholders. Information security becomes critical success factor and business enabler. It assists organizations in converting today's threats into tomorrow's opportunities to achieve competitive edge.

How we can develop information security objectives, which promote security as added value for all stakeholders?

Quality Function Deployment (QFD) provides a structured, comprehensive framework aimed at satisfying the customers by analyzing their requirements, the "voice" of the customers [18, 19]. "Quality" means the satisfaction of customer requirements [18, 19]. These demands are translated throughout the development and production process [18, 19] to deliver value [19, 20].

In this paper we provide concrete guidelines to extend and apply for the first time the concept of QFD for developing corporate information security objectives. In accordance with action research [21, 22] we analyze in the next section the current information security research and relevant part of QFD. Then we explain our process (Sect. 19.3). This innovative information security objective development process has been implemented for several years by different small and medium sized organizations of distinct sectors. The obtained experiences, limitations, implications for practice and research are reflected in Sect. 19.4. Section 19.5 provides a conclusion for the paper.

19.2 Theoretical Framework

In this section we summarize requirements for the development of information security objectives and explain current approaches. At the end we present relevant steps and requirements of QFD.

19.2.1 Information Security Research

To achieve effectiveness and sustainability information security must be addressed at the highest levels of the organization. It must be part of or aligned with corporate governance [2–4, 9, 13, 15, 23, 24]. The board and executive management should provide security direction to deliver enhanced business value to all stakeholders [13, 15, 23, 25]. All objectives must be formulated in such a way that compliance and conformance, as well as suitability, adequacy and effectiveness of an information security management system can be measured and improved [2–4, 13–15, 26]. They should support consistent, strategic-aligned decision making, prioritizing investments, optimizing resource allocation while minimizing costs and risks [4, 8, 24].

Security objectives should establish a sense of direction and principles for action [2–4, 10, 13, 23, 26]. Information security requires a holistic, systemic and comprehensive approach by involving collaborators and all constituting dimensions (such as governance, policy, management, organization, processes, best practice, ethical, cultural, legal and technical aspects, insurance, measurement, improvement, certification and other) [1–4, 8, 13–15, 23, 27].

If the collaborators are involved in the establishment, they provide needed business knowledge, contribute to higher performance and a better alignment with business objectives, values, and needs [12].

Following aspects should be taken into account by establishing information security objectives [4, 11, 12, 14, 23]:

- the organization's overall business strategy and objectives,
- the legal, statutory, regulatory, and contractual requirements that an organization, its trading partners, contractors, and service providers have to satisfy and their compliance,
- the particular business requirements, relevant guidelines, as well as feedback and recommendations by interested parties,
- the organizational environment, business circumstances, organization, resource availability, assets, technical environment,
- the results of risk assessments, trends related to threats and vulnerabilities and business continuity requirements,
- the socio-cultural environment with ethical and social values and trust, collaborators participation,
- the results of independent or management reviews, reported security incidents, the status of preventive and corrective actions and the process performance. Recently some researchers propose to use a Balanced Scorecard (BSC) for linking strategic as well as compliance drivers to security objectives [4, 28] or to integrate information security as additional dimension in an IT BSC [29]. But they describe no method to develop stakeholder oriented corporate security objectives, which provide added value for all stakeholders.

19.2.2 Quality Function Deployment

QFD is a set of Total Quality Management tools, techniques and methods developed in the 1960s by Yoji Akao and other leading quality experts [20]. It focuses on delivering products and services that satisfy customers by analyzing customer requirements and translating the demands into design targets and quality assurance points to be used throughout the development and production process [18, 20]. Over the years it has received a lot of extensions and variations [20, 30]. Its implementation results in higher customer satisfaction, promotes internal communication, shorter improvement time, lower costs and larger market share [18–20].

In this paper we will focus at the first steps of QFD to identify and translate customer requirements. Firstly the expressed and latent demands and expectations of the customers in the target marketplace are analyzed by a cross-functional team consisting of members of most different departments, such as marketing, sales, research and development, engineering, design, manufacturing and production, procurement, quality, service, etc. [18, 20]. To determine customer requirements literature, brainstorming, benchmarking, desk panels, discussion groups, interviews, questionnaires and other techniques are used [18]. Further other important characteristics of the target market are studied, including relevant legal and regulatory requirements, contractual and statutory obligations and others.

In the next step the relative importance of all customer requirements are established [20]. The QFD literature proposes therefore many different methods, such as basing it on surveys, or using statistical methods, such as the AHP or other scoring methods (see [22]). After a competition analysis is conducted, the strengths and potential improvements are elaborated and performance objectives are defined. In the next step QFD lists the engineering characteristics and analyze the effect of each engineering characteristic to one or more customer characteristics. Based on the weight of each customer characteristics and the influence of the single engineering characteristics the engineering characteristics are prioritized regarding their criticality for customer satisfaction [20]. The interdisciplinary team seeks consensus on these evaluations, basing them on expert engineering experience, customer responses, and tabulated data from statistical studies or controlled experiments [18]. The majority of QFD applications stop with the completion of these first matrices (or House of Quality) [30]. In the same way the engineering characteristics are translated after throughout the next steps of the development and production process [20].

QFD must be adopted for each organization and project by integrating all relevant methods, diagrams and tools that are most appropriate and overall by focusing on the most relevant aspects [20].

19.3 Stakeholder Oriented Objectives Development

In accordance with action research [21, 22] we developed in consolidation of a holistic information security approach, our practical experiences and QFD (see Sect. 19.2) our stakeholders oriented process to establish organization-specific corporate security objectives. We start right from the top and elaborate or extend the main corporate objectives or corporate policy by integrating information security. Our approach consists of following main steps:

1. identify all relevant stakeholders,
2. define and prioritize all stakeholders' requirements and relevant factors, and
3. translate the requirements in corporate objectives.

Table 19.1 Guidance for stakeholder identification

| |
|--|
| Who are our customers, consumers, users? |
| Who should be addressed? |
| Who is involved? |
| Who is affected indirectly? |
| Who can contribute, cooperate? |
| Who can give us information? |
| Who encourage us? |
| Who has influence? |
| Who decides? |
| Who may have scruples? |

19.3.1 Identify Stakeholders

Based on the economic definition of stakeholders we regard:

- shareholders,
- customers and potential customers of all target markets,
- collaborators,
- suppliers/partners, and
- environment and society.

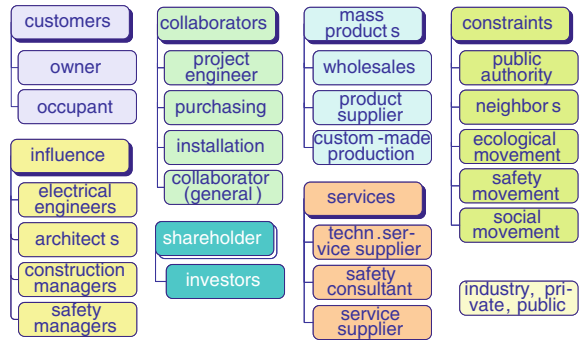
The best way to establish the stakeholders is based on our experiences brainstorming in a cross-functional team consisting of the top management and members of all divisions, departments and service units. If admitted by corporate culture collaborators of all levels should participate. Answers to the questions of Table 19.1 provide a guidance to establish the stakeholders.

We determine the stakeholders as detailed as necessary. Firstly we start with a higher level and go after deeper, if some part of a stakeholders group has different requirements (e.g. customer with particular requirements). The identified stakeholders are grouped by using affinity diagrams or hierarchy diagrams (see Fig. 19.1).

19.3.2 Stakeholders Requirements

To deliver value to the stakeholders we analyze the expressed and latent needs, expectations, problems they face, aspects that are appreciated by their customers and demands of each stakeholders segment [18, 19]. For private customers of an electrician, who invest in high quality and expensive products for example, it is often very important that this will be kept secret. For hotels, for example, the availability of all services including television, data connection, electrician in the rooms and others are essential for the satisfaction of their customers. An assistance for 24 h and 7 days and a short reaction time offer new opportunities to deliver

Fig. 19.1 The stakeholders of an electrician



value to the customer. For the shareholders among others legal compliance and accountability are important. Collaborators expect on one hand the protection of their personal data and on the other hand to be able to access all information as they want.

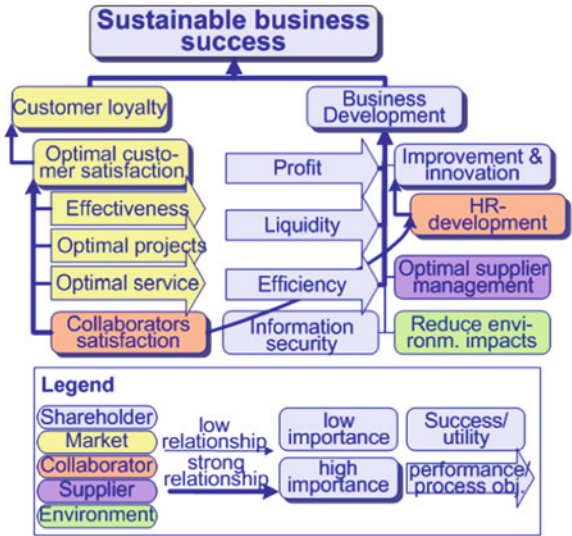
In our case studies we identify also the stakeholder’s requirements in the same cross-functional team (see [Sect.19.3.1](#)) based on discussions and experiences with stakeholders, their feedback and recommendations and by using literature, benchmarking, interviews, questionnaires and others. We used brainstorming with pin boards, post-its and other moderation techniques. Table [19.2](#) offers a general guidance to identify stakeholders’ requirements. The questions must be adopted for each organization.

In the most case studies we have developed a holistic corporate policy, which integrates information security. Otherwise we have to regard also the organization’s overall business strategy and corporate objectives as requirements.

The identified requirements (see concepts of [Fig. 19.2](#)) are grouped by using affinity or hierarchy diagrams. In the next step we analyze the organization advantages and weaknesses in comparison to the competitive environment (SWOT analysis and competitive positioning) and assess potential risks and uncertainty for each requirement group. For the competition analysis we use a scale from 1 to 10, whereby ten represent our most important advantages by regarding also sale value, market barriers, market growth and others. The general risk assessment considers risks of any type, such as economical, financial, market, social, technical, ecological and others. The risks are ranked from 1 to 10, whereby one represents the lowest risk level.

To rank stakeholders’ requirements we made the best experiences with a point scoring scale by institution of each team member. We firstly prioritized the stakeholders and after we prioritized the requirements by regarding the stakeholder’s priority, as well as the competition and risk analysis to promote our opportunities. The ranking is used to find a solution for eventually conflicting requirements and to concentrate us on the key requirements accordingly to the Pareto principle. In some cases it was helpful to visualize the interdependencies of the requirements for prioritizing them. During the development the defined

Fig. 19.2 Corporate objectives and their relationships of an electrician



stakeholder’s requirements and their priority are continually discussed and adjusted. It is essential to balance the different requirements (e.g. between profit, customer satisfaction, collaborators satisfaction) in an optimal way. The process ends when everybody agrees with the results and commits it.

In some cases we used also the analytic hierarchy process (AHP): we prioritize the stakeholders and weight the importance of each requirement from the focus of each stakeholder. These values are multiplied by the stakeholder weights and the results are summed yielding the requirement priority. Most participants considered it as too statistical based and was not able to really commit the results.

19.3.3 Information Security Objectives

Based on the final requirements we elaborate the corporate security objectives by wording in a pro-active, concise, clear, memorable, motivating and attainable form. In some cases we used a requirement—objective matrix to prioritize the objectives and to document how they were developed. Otherwise most of the requirements could be used directly as objectives. Based on the information security objectives we elaborated the corporate (information security) policy.

To visualize and analyze the cause-and-effect relationships of the different objectives we use based on the BSC strategy map approach [31] an influence diagram (see Fig. 19.2), instead of the QFD dependency sheet (the roof of the house).

Table 19.2 Guidance for stakeholders’ requirements

| |
|--|
| Which customers or markets will we target? |
| What does the users/market requires, needs, expect? How do we truly create superior value for our market? |
| How can we promote the corporate objectives of our customers? Which problem they face or will face in the future? What can we offer that they are appreciated by their customers? Who benefits in what way? |
| How can our stakeholders be getting excited? How do we capture value for our stakeholders? |
| How do we distinguish ourselves from competitors? Which reputation will we acquire? |
| Why are we paid? Why do we exist? What is our mission? What should we achieve? What is our vision? |
| What are our success factors? |
| What are the relevant statutory legal and regulatory requirements (for all markets)? What are relevant guidelines, contractual, health, safety and hygienic obligations for all customer segments? |
| What should be changed? |
| What is most important for us? What are our desirable attitudes and behaviours, values, basic assumptions, and beliefs? |
| What factors are to be considered (political, economic, financial, social, technological, physical, environmental and others)? What are the dynamics of the business and markets? |
| What are our internal capabilities,Performance, core competences, strategic tangible and intangible assets, human capital capabilities, resources, knowledge, experiences, technology, organizational enablers, processes, and infrastructure? How can we use technology to do new things? |
| Which are our main risks to face? |

19.4 Evaluation

Action research stands out as an ideal research method to evaluate the IS security objectives development approach. It allows theory refinement and theory testing in practice [21]. Further it is aimed at creating organizational change and solving practical problems through the research [22]. Walsham [32] regards action research as the ideal way to perform involved research.

Our stakeholder oriented corporate information security objective development approach has been implemented successfully by different small and medium-sized organizations (from 27 to 990 employees) of distinct sectors (service, engineering, and public) in Italy, Austria and Germany. The described approach leads to the following results collected by measuring the project process, analyzing audit reports as well as interviewing and observing concerned management and collaborators:

- *Efficiency*: Establishing the corporate (information security) objectives required in small and medium sized organizations a work effort of approximately 1 day. It is important to focus on the main priorities and use efficiently moderation techniques.

- *Market driven, strategic aligned information security*: Prior information security was seen in some organizations as something dangerous and cost-expensive or an imposed legal requirement. With the market focus information security has become success factor or also business opportunity to provide enhanced business value. For an office furniture, e.g. analyzing the data class kept in an office and proposing a data archive protection accordingly to legal requirements becomes a consulting and marketing argument.
- *Holistic, balanced and collaborative*: All developed corporate policies regard organizational, technical and cultural aspects, as well as internal and external factors and balance them. The holistic view and the discussion in cross-functional teams from all relevant organizational units promote security awareness over the whole organization and a common market driven information security view. The developed corporate policy is an excellent basis to communicate this throughout the whole value network and the market.
- *Structured and systematic*: The structured and traceably QFD approach was appreciated overall by the top management. The visualization by influence diagrams are continuously used as basis for strategic and investment decisions, as well as to analyze periodically its actuality and to eventually restart the development process.

19.4.1 Implications for Practice

The strategic approach to information security of an organization must become market and stakeholder oriented to provide enhanced business value.

For many organizations information security can be a unique selling point and contribute to competitive edge (e.g. business or financial services, e-commerce, ICT services, e-government services, health services).

Based on the case study results the impact and relationship of security to corporate objectives should be clearly elaborated. In that way all security related investment, resource and strategic decisions can be based on corporate objectives.

Information security should be fully integrated with the corporate management. It should become part of governance, controlling, corporate culture as well as all business processes and the daily work of everyone.

19.4.2 Limitations and Implications for Research

Limitations of this process are the general difficulties of QFD: the definition of the stakeholder's requirements, unclear correlation among the demands, required management support, the collaboration in cross functional teams, the size of the house of quality and others [30].

Further research is needed to study the challenges in large organizations with numerous internal and external stakeholders and operating in most different countries and cultures.

Our next research focus on the next steps of information security governance with metrics and strategy development, objective deployment, operation, measurement, reporting and continual improvement.

Further research directions are required also on general information security management topics to integrate security based on an interdisciplinary holistic approach with all corporate management activities. In accordance with the increasing importance of information security and growing requirements for customer satisfaction, cost-reduction and efficiency we have to integrate information security in all strategic, tactical and operational processes and activities by regarding all security approaches (governance, socio-organizational, technical). This holistic, interdisciplinary approach will be a great research challenge for the future.

19.5 Conclusion

Organizations need clear, coherent security objectives to properly and effectively guide and control information security and to demonstrate information security governance and compliance. Based on the research results our stakeholders' oriented corporate information security objective development process promotes a holistic, collaborative, systemic, structured and market driven information security approach. In that way information security becomes a new role as success factor or business opportunity to provide enhanced business value and competitive edge. The impact and relationship to corporate objectives of all security investments, resources and strategic decisions should be clearly elaborated. Information security should be fully integrated within all corporate management activities. It should become part of governance, controlling, corporate culture as well as all business processes and the daily work of everyone. Thus a stronger interdisciplinary and holistic information security research approach is required, which integrates information security governance with technical and socio-organizational aspects.

Acknowledgments The research leading to these results was partially funded by the Tyrolean business development agency through the Stiftungs assistenz QE—Lab and the COSEMA project, which is part of the Translational Research program.

References

1. Saint-Germain R (2005) Information security management best practice based on ISO/IEC 17799. *Inf Manag J* 39(4):60–66
2. von Solms SH, Solms RV (2009) Information security governance. Springer, New York

3. Da Veiga A, Eloff JHP (2007) An Information security governance framework. *Inf Syst Manag*, 24. 361–372
4. Sowa S, Tsinas L, Gabriel R (2009) BORIS –Business ORiented management of information security, managing information risk and the economics of security. In: Johnson EM (ed.) *Managing information risk and the economics of security*, Springer, New York pp. 81–97
5. Gordon LA, Loeb MP, Sohail T (2010) Market value of voluntary disclosures concerning information security. *MIS Q* 34(3):567–A2
6. International Standard Organization (2009) ISO Survey of Certifications <http://www.iso.org/iso/survey2009.pdf>
7. Dhillon G, Backhouse J (2001) Current directions in IS security research: towards socio-organizational perspectives. *Inf Syst J* 11(2):127–153
8. Beebe NL, Rao VS (2009) Improving organizational information security strategy via meso-level application of situational crime prevention to the risk management process. *Commun AIS* 2009(26):329–358
9. Straub DW, Welke RJ (1998) Coping with systems risk. *MIS Q* 22(4):441–469
10. Siponen M, Vance A (2010) Neutralization: new insights into the problem of employee information systems security policy violations. *MIS Q* 34(3):487–A12
11. Bulgurcu B, Cavusoglu H, Benbasat I (2010) Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness. *MIS Q* 34(3):523–A7
12. Spears JL, Barki H (2010) User participation in information systems security risk management. *MIS Q* 34(3):503–A5
13. ISO/IEC27001, ISO/IEC 27001:2005 (2005) Information Technology, Security techniques, Information security management systems requirements. International Standard Organization, Geneva
14. National Institute of Standards and Technology (NIST) (2008) Federal information technology security assessment framework. <http://csrc.nist.gov/libproxy.unibz.it/drivers/documents/Federal-IT-Security-Assessment-Framework.pdf>
15. IT Governance Institute (ITGI) (2006) Information security governance: guidance for boards of directors and executive management. IT Governance Institute, rolling meadows, IL
16. Chen DQ, Mocker M, Preston DS, Teubner A (2010) Information systems strategy: reconceptualization, measurement, and implications. *MIS Q* 34(2):233–A8
17. Siponen M, Willison R (2009) Information security management standards: problems and solutions. *Inf Manag* 46(5):267–270
18. Akao Y (1990) Quality function deployment integrating customer requirements into product design. Productivity Press, Cambridge
19. Mazur GH (1994) QFD for small business, a shortcut through the Maze of Matrices In: [The sixth symposium on quality function deployment, Novi, Michigan, June 12–14]. http://www.mazur.net/works/-sme_qfd.pdf
20. Li Y, Huang M, Chin K, Luo X, Han Y (2011) Integrating preference analysis and balanced scorecard to product planning house of quality. *Comput Ind Eng* 60(2):256–268
21. Barkerville R, Wood-Harper A (1998) Diversity in information systems action research methods. *Eur J Inf Syst* 7(2):90
22. Baskerville R, Myers MD (2004) Special issue on action research in information systems: making is research relevant to practice—foreword. *MIS Q* 28(3):329–335
23. ISO/IEC27002, ISO/IEC 27002:2005 (2005), Information Technology, security techniques, code of practice for information security management. International Standard Organization Geneva
24. Great Britain Office of Government Commerce (OGC) (2007) Service strategy (SS) : ITIL. TSO The Stationery Office, London
25. OECD (2004) Principles of corporate governance <http://www.oecd.org>
26. IT Governance Institute (ITGI) (2007) COBIT 4.1: framework, control objectives, management guidelines, maturity models. IT Governance Institute Rolling Meadows

27. Stoll M, Breu R (2012) Integrating information security governance and standard based management systems. In: Gupta M (ed) Strategic and practical approaches for information security governance, technologies and applied solutions, IGI Global
28. Jaquith A (2007) Security metrics: replacing fear, uncertainty, and doubt. Addison-Wesley, Upper Saddle River
29. Baschin A (2001) Die Balanced Scorecard für Ihren IT-Bereich: ein Leitfaden für Aufbau und Einführung. : Campus-Ver, Frankfurt/Main u.a.l
30. Carnevalli JA, Miguel PC (2008) Review, analysis and classification of the literature on QFD—types of research, difficulties and benefits. *Int J Prod Econ* 114(2):737–754
31. Kaplan RS, Norton DP (2000) Having trouble with your strategy? Then map it. *Harv Bus Rev* 78 (5):167–176
32. Walsham G (2006) Doing interpretive research. *Eur J Inf Syst* 15(3):320–330

Chapter 20

Stakeholder Oriented Information Security Reporting

Margareth Stoll

Abstract Organizations have to meet most different enterprise-specific stakeholders', business, standard, legal and regulatory information security requirements. They are faced with a wide range of potential security threats and socio-organizational challenges. To invest all security efforts effectively the collaborators and partners of the whole value chain must be aware how they contribute to achieve common objectives and compliance. This is scarcely supported by fragmented approaches. To bridge the gaps we analyze accordingly to a design-science approach the different requirements and present a coherent and systematic stakeholder oriented information security reporting model. The comprehensive, systemic and structured reporting approach demonstrates the value of information security and sustains informed decision making to invest security efforts pro-actively, effectively and efficiently. The stakeholder oriented focus on security reporting offer new impacts for practice and a wide range of most different research questions.

20.1 Introduction

In today's complex, interconnected world information security must become a critical success factor, which assists organizations to create enhanced business value. Organizations have to concentrate all security efforts on what matters most to fulfill enterprise-specific stakeholders', legal and business security requirements and compliance. In this paper we use an economic definition of stakeholders,

M. Stoll (✉)

University of Innsbruck, Technikerstr. 21a, 6020, Innsbruck, Tyrol, Austria
e-mail: margareth.stoll@uibk.ac.at

which regards shareholders, customers and potential customers of all target markets, collaborators, suppliers/partners and the environment and society.

An increasingly faster changing environment (market, customer, technology, law or regulations) requires a continual adaption of information security processes, controls and procedures. The effectiveness and performance of the security management and the actual risk and compliance situation must be continually evaluated and improved [1–8]. Effectively implemented security measurements demonstrate the value of information security to top management, face informed decision making, provide information on compliance, improve security confidence and enable stakeholders to improve ongoing information security [2, 3, 6, 9]. This is a critical success factor for sustainable information security [10].

Different measurement requirements were defined by best practices, such as Control Objectives for Information and related Technology (COBIT) [3], the Information Technology Infrastructure Library (ITIL) [4] and by measurement models, such as ISO/IEC 27004 [1] and NIST 800-55 [2].

Huge technical, operational metrics and some for economic or risk issues were proposed (e.g., [11, 12]). Measurement taxonomies have been developed (e.g., [6, 13]). Information security problems are complex and require a collaborative approach [7, 14]. But measurement is still considered in a fragmented way [6, 15]. One of the most common mistakes of security performance management is to saturate the organization with metrics that have little meaning to the target audience [12]. Each stakeholder needs appropriate, comprehensible and useful reports accordingly to the security requirements for his area of responsibility [2, 9, 14].

Thus the stakeholders of all organizational and technical levels over the whole value chain need firstly coherent security objectives [2, 8, 16, 17]. The effectiveness of achieving these objectives including performance, risk and compliance can be controlled in a stakeholder oriented way.

Despite these requirements there is a lack of a holistic, systematic method to cascade/report information security objectives/measurement results in a coherent and systemic way [6]. This leads us to the following key research questions:

How can we deploy security objectives to the stakeholders of all levels over the whole value chain in a coherent way?

How can we aggregate security metrics to control the achievement of the established objectives in a comprehensive, stakeholder oriented and systematic way?

The remainder of this paper is structured as follows: in the next section we analyze reporting requirements and current research approaches, standard requirements and different best practices. Based on these and expert interviews we establish in the third section the requirements and explain in the fourth section our stakeholder oriented information security reporting model. In the fifth section we evaluate the fulfillment of the established requirements, reflect limitations and deduce implications for practice and research. The last section provides a conclusion for the paper.

20.2 Problem Analysis

Information security must be part of or aligned with corporate governance [4, 5, 7, 8, 10, 16, 17]. The board and executive management must provide and review the strategic direction by establishing objectives and monitoring the implementation and performance [10, 16, 17]. Operational, tactical and strategic security objectives should be linked transparently to corporate objectives [2, 5, 7, 10]. Measurement data should be extracted on the operational level, compiled and integrated to perform measurement and monitoring on the tactical level. The board and executive management need strategic level reports for related management decisions [2, 3, 5–8, 16].

The established objectives and the performance of security management should be reviewed and improved at planned intervals or if significant changes occur to ensure ongoing suitability and effectiveness [2, 3, 5, 7, 8, 14, 16]. Security metrics should support the detection of security events and the identification of attempted security breaches, incidents and previously undetected or unknown security issues [1, 16].

Information security requires a holistic, systemic and comprehensive approach by involving all collaborators and partners, as well as all constituting dimensions (such as governance, policy, management, organization, processes, best practices, ethic, culture, legal aspects, technology, measurement, improvement, certification and others) [2, 5, 7, 8, 10, 14, 16–18].

Information security objectives should be communicated in a form that is relevant, accessible and understandable to establish a sense of direction and principles for action [3, 7, 10, 16, 19, 20].

Technical oriented information security approaches are based overall on enterprise architecture, which is defined by ISO/IEC 42010:2007 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”. Architecture oriented frameworks start from the business processes, describe the information architecture, align the logical IT services, go through the application architecture and the whole technology architecture until the infrastructure and physical environment. The interdependency of the different components of all architecture levels are documented [4, 21, 22].

Recently some researchers propose to use a Balanced Scorecard (BSC) for linking strategic as well as compliance drivers to security objectives [5, 23, 24] or to integrate information security as additional dimension in an IT BSC [25]. The BSC was founded by Kaplan and Norton [26, 27] and provides a framework to develop a strategy, translate the strategy into objectives and measures, plan operations, and finally to monitor and learn to continually test and adapt the strategy [28]. Kaplan and Norton developed the strategy map to visualize the strategy as a chain of cause-and-effect relationships among strategic objectives [29]. BORIS uses transferring tables to propagate compliance and strategic

business requirements to security objectives [5]. Jaquith [24] cascades the business security scorecard in departmental scorecards. Herath et al. [23] declare this as further research issue.

20.3 Requirements

Based on interviews with top management, chief information security officers, security experts and collaborators of different levels and departments organizations face great challenges to:

- demonstrate and underline the importance of information security in the “voice” of the top management,
- define coherent security objectives for the collaborators and partners of all levels over the whole value chain,
- control security efforts and report results in a stakeholder oriented, coherent and systematic way.

Current security objective deployment approaches stop either at the departments or business process level or start from the process level. Additionally to presented approaches we aim to close this gap. We develop a model to establish stakeholder oriented objectives and reports for the collaborators and partners of all organizational and technical levels over the whole value chain in a comprehensive, coherent and systematic way. In consideration of the increasingly interconnected, cross-organizational business and based on the research framework and the challenges in practice we elaborated in accordance to a design science approach [30, 31] the main requirements for a stakeholder oriented reporting model (see Table 20.1).

20.4 Model

In accordance to a design science methodology [30, 31] we developed in consolidation of current information security research discussed above and the established requirements a stakeholder oriented information security reporting model.

20.4.1 Objective Deployment

As a first step we elaborate corporate information security objectives. To focus all security efforts on enhanced business value and compliance we proposed an innovative stakeholders’ oriented information security objective development

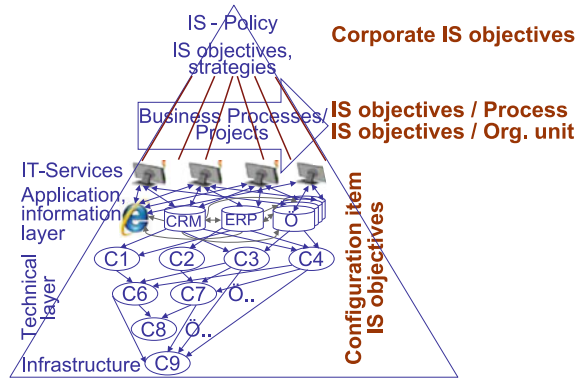
Table 20.1 Main requirements

| Requirement | Description |
|-------------------------------|--|
| Stakeholder oriented | Assign understandable and relevant security objectives to all collaborators/partners and communicate the measurement results in a meaningful and useful way. |
| Strategic aligned | Align security objectives transparently to corporate objectives and provide appropriate reports to control their achievement and the performance of security efforts. |
| Comprehensive | Deduce security objectives from corporate security objectives down over all organizational and technical levels of the whole value chain. Report security issues to collaborators/partners of all levels over the whole value chain (additional cross-organizational requirement). |
| Coherent and systematic | Deduce all security objectives from corporate security objectives in a coherent and systematic way. Extract detailed measurement data on the operational level, consolidate, interpret and aggregate them to monitor the achievement of the objectives of upper levels. |
| Holistic | Follow a holistic information security approach, which integrates the different security dimensions (such as governance, culture, awareness, organization, processes, operation, technology, infrastructure, legal, regulatory and business requirements, best practices, certification and others). |
| Promote continual improvement | Control and report continually the effectiveness and performance of security efforts and detect changed internal or external conditions or security issues to promote ongoing information security and continual improvement. |

approach [32]: we identify all relevant stakeholders, analyze their needs, deduce and prioritize security requirements and all relevant factors (inclusive legal, regulatory and standard requirements) and translate these in corporate security objectives with priorities and relationships. To visualize and analyze the cause-and-effect relationships between the different corporate objectives we use an influence diagram. In that way the relationship of information security to other relevant corporate objectives are established. Thus the entire organization struggles to accomplish stakeholders’ requirements inclusive information security, as well as legal, regulatory and standard compliance.

Information security objectives for all core business processes are deduced from corporate security objectives by using a matrix. The rows assign relevant objectives to each business process (columns). Additionally to the deduced objectives we analyze with the collaborators concerned unique security requirements for each business process and define eventually further security objectives (e.g., performance objectives). We take into account special stakeholders’, business, contractual, legal, regulatory requirements, relevant standards and best practices, used assets, risks, threat trends and cultural, technological, environmental and organizational aspects. In the same way the security objectives are broken down for all lower business processes and integrated with process objectives.

Fig. 20.1 Objective deployment



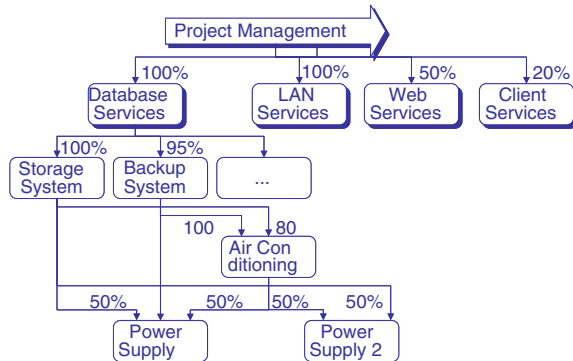
In the next step we analyze with the collaborators concerned all business processes over the whole value chain to identify additional unique information security requirements at single process steps with assigned roles and responsible organizational units or partners. We describe also the entire non-technical assets (such as information, data, documents, archives and others) and their security requirements. Based on the defined security objectives for each business process and the involved organizational units/partners, as well as additional security requirements of single process steps with assigned roles and responsible units/partners we deduce security objectives for all non-technical organizational units or partners by using a matrix (see Fig. 20.1: upper part).

In accordance to ITIL [4], COBIT [3] and the IT architecture oriented approaches we define with the collaborators concerned the *information security configuration model* (see Fig. 20.1: lower part starting from business processes). We describe for all business processes the aligned IT services with dependencies and relationships. For all these IT services we document their supporting configuration items over the whole value chain going always deeper through all relevant technical layers until the physical infrastructure (see Fig. 20.1: lower part starting from business processes). In line with the definitions of ITIL we understand by a configuration item any component that needs to be managed in order to deliver an IT service, such as software, hardware, infrastructure, documentation [4].

Starting from the assigned information security objectives to the business processes we deduce the security objectives to all aligned IT services by regarding the dependency of business processes from IT services. The project management process, for example (see Fig. 20.2), is defined as 100 % depending on database services (because the project management process is down without database services), 100 % depending on local area network services and due to limited support without web services for 50 % on internet services, as well as 20 % on single client services (due to the redundancy of clients the collaborators can continue to work with another workplace).

The highest requirement value for each security objective of all upper items is inherited by regarding the dependencies to lower items. If we assume, for example,

Fig. 20.2 Part of a configuration model with dependencies



a required availability of 4 for the project management process (see Fig. 20.2) a value of two (50 % dependency) is inherited to the web services. The security objectives are deduced down over all levels to all configuration items and eventually further objectives for special security requirements are assigned.

In the next step we assign to each item appropriate roles and responsibilities. Thus all technical collaborators and partners receive coherent and stakeholder specific information security objectives, which are deduced from corporate security objectives in a systematic, structured and transparent way.

For each security objective of all levels and over the whole value chain we determine with the collaborators concerned goals. On upper levels we define long-term, medium-term and annual goals. As deeper as we come down the planning period becomes always shorter. For each goal we define appropriate metrics or indicators, targets, measurement methods and assign roles and responsibilities (see [33]) to monitor their achievement (measurement implementation plan). Additional measurement methods are defined to detect errors, incidents, breaches or previously unknown information security issues. Apart from the measurement results we use also data acquired through audits, self assessments, tests, scanning, reviews and external information (e.g., environment observation, interviews, and technical surveys/reports, feedback from market, supplier and partner).

20.4.2 Reporting

In accordance to the defined measurement implementation plan detailed measurement data are extracted on the operational level, consolidated and interpreted. The responsible of each lowest item receives all relevant operational reports to control the achievement of planned targets. Considering the defined dependencies and relationships of the configuration model we aggregate the measurement results of lower levels and report them on upper levels by using business intelligence tools. It is the reverse way to the objective deployment. Thus the responsible for each objective receives relevant reports to control the achievement of planned

targets. The impact of the unavailability of the web services (see Fig. 20.2), for example, is calculated by the product of the importance of the project management process (e.g. percentage of profit produced by projects) and the 50 % dependency. In that way the achievement of established information security objectives is evaluated and reported in a systemic and overall stakeholder oriented way. The board and executive management receives necessary information to take strategic-aligned and cost benefit balanced decisions.

The responsible of each item, process or organizational unit monitors continually the measurement results, as well as events, incidents, breaches or previously unknown security issues to take appropriate corrective actions and to propose preventive actions in accordance to established processes, if necessary. In that way the effectiveness is improved; problems are prevented; impacts and volume of breaches and incidents are reduced.

On demand (e.g. for problem or trend analysis) or if measurement results exceed defined thresholds, patterns or targets (e.g., time-limits) the results are communicated to the responsible of all next upper items. Corporate culture, the information preferences of the responsible of each upper item and the defined criticality of upper items influence the reporting flow. In that way the responsible of upper items receive escalated and attention required operating information only or on demand. They are not overloaded. If power supply 1 (see Fig. 20.2), for example, has a problem the backup system is down due to its 100 % dependency. The responsible for the backup system receives an alert concurrently with the responsible for the power supply 1. If the problem is not resolved, for example one hour before the next backup starts, the responsible for the database services receives an alert, too. Depending on the criticality of the project management process, the corporate culture and the information preferences of the project management process owner he receives also alerts. In that way necessary actions are planned on time to prevent further problems.

Based on the configuration model the cause analysis of security events and the security impacts of projects, improvements and investments are checked in a systemic way and all security efforts are prioritized in accordance to coherent objectives.

The actuality and accuracy of the whole configuration model must be maintained ongoing. The suitability of the established security objectives and the usefulness and effectiveness of security reporting must be reviewed at planned intervals or if significant changes occur. If objectives changes, they must be deployed another time to all lower levels. To monitor their achievement goals with appropriate metrics or indicators, targets, measurement methods must be defined and roles with responsibilities assigned. The security reporting must be adapted accordingly. Thus the stakeholder oriented security reporting promotes ongoing information security in today's complex, interconnected and increasingly faster changing world. All information security efforts are focused in a sustainable, coherent and systematically way to provide enhanced business value.

20.5 Evaluation

In this section we present in accordance to a design science approach [30, 31] the results of our stakeholder oriented security reporting model in terms of established requirements (see Table 20.1). After we describe limitations and deduce implications for practice and research.

20.5.1 *Fulfillment of the Requirements*

The stakeholder oriented information security reporting model fulfills all elaborated requirements (see Table 20.1):

Stakeholder oriented: The responsible of each item, all organizational unit manager, business process owner and the top management receive clear, unique and coherent security objectives and relevant reports to control the achievement of assigned objectives and the performance of their security efforts. Information security becomes part of the daily work of all collaborators/partners. While the responsible of each item controls measurement results, as well as events, incidents, breaches or previously unknown security issues continually and takes necessary actions, the responsible of upper items receive only attention requiring information or on demand. In that way only meaningful and useful information are reported.

Strategic aligned: Corporate security objectives, explicitly linked to corporate objectives, are deployed systematically and coherently to all business processes, organizational units and configuration items. The alignment of each security measure, project, improvement and investment to corporate objectives demonstrates the value of information security and sustains informed decision making to invest security efforts effectively and efficiently.

Coherent and systematic: The influence diagram, matrices and the configuration model link the security objectives with the related measurement results of all levels and organizational units to corporate objectives in a coherent, systematic and transparent way. The systematic event reporting facilitates cause analysis and brings together most different information in a systemic and cooperative manner.

Comprehensive: The corporate information security objectives are deployed coherently and systematically down over all organizational and technical levels of the whole value chain. The measurement results and eventually occurred events from the whole value chain are reported vice versa over all levels. The security efforts of the entire value chain are focused on coherent objectives.

Holistic: The comprehensive approach regards the strategic, tactical and operational level over the whole value chain. By establishing corporate objectives and by analyzing specific additional objectives of processes, units or items we take into account organizational, technical and cultural aspects, as well as internal (e.g. stakeholders' requirements, awareness, assets, results of the risk assessment, threat trends and others) and external factors, such as stakeholders', business,

contractual, legal, regulatory and relevant standards and best practice requirements, environmental conditions and information (e.g. technical surveys/reports, feedback from market, supplier and partner). Thus the security reporting model integrates all most different information security dimensions.

Promote continual improvement: Additionally to the deduced objectives we integrate also performance objectives and special operational objectives to improve effectiveness, prevent problems and reduce impacts and volume of breaches and incidents. The systemic, strategic-aligned and structured reporting model provides an optimal basis to operate in a coherent, pro-active, creative and flexible way on all changing requirements, conditions and opportunities. The suitability of the established information security objectives and the usefulness and effectiveness of security reporting is reviewed at planned intervals or if significant changes occur. Thus an ongoing effectiveness and performance of the information security management is promoted.

Based on case studies results our objective deployment approach promotes stakeholder oriented, coherent and strategic-aligned security objectives in an effective, efficient and sustainable way. The systemic approach over all organizational and technical layers and units of the whole value chain develops a big picture of all objectives and items with dependencies and possible impacts. Thus the awareness of business drivers, supporting technology and supported business processes and the understanding for the work and requirements of other functional and technical organizational units was increased. Coherent security objectives create a common security understanding to enhance business benefits and compliance.

20.5.2 Limitations and Implications

Opposite to the advantages is the effort for the establishment of the configuration model. An appropriate level of detail is essential to implement stakeholder oriented information security reporting in an effective and efficient way.

In the next step we will empirically evaluate the practical usefulness and challenges of the entire model in different medium-sized and large organizations.

Clear understandable, coherent objectives and useful stakeholder oriented security reports for all involved collaborators/partners of the whole value chain should be adapted in practice. In that way all stakeholders receive a clear direction, can act pro-actively, flexibly and control their effectiveness. The impacts of stakeholder oriented reporting on security commitment and awareness should be empirically investigated.

The role of security reporting for continual security improvement should be further investigated. This offers a wide research field for most different disciplines, such as communication theory, motivation theory, knowledge management, organizational learning and others.

The ever stronger interconnection and virtualization will promote the request for coherent and systematic cross-organizational security reporting.

Information security reporting should become an integral part of corporate management in practice. The business value of information security should be explained by aligning all security measures, projects, improvements and investments to corporate objectives. In that way the top management can take cost benefit balanced strategic decisions and all security efforts are focused on enhanced business value. This holistic, integrated, interdisciplinary information security approach will be a great research challenge for the future.

20.6 Conclusion

In today's complex, interconnected world information security must become a critical success factor, which assists organizations to create in common enhanced business value and compliance. Thus organizations have to establish for each collaborator and partner of all organizational and technical levels over the whole value chain clear understandable, unique security objectives, which are systematically and coherently deduced from corporate information security objectives by integrating additional requirements. Clear objectives provide direction. Each collaborator and partner can take ownership to act pro-actively, effectively and flexibly. The achievement of these coherent security objectives, the performance of security efforts and the contribution of security measures, projects, improvements and investments to corporate objectives are controlled by relevant stakeholder oriented security reports. In that way information security becomes an integral part of corporate management and all collaborate to accomplish common security objectives including legal, regulatory and standard compliance.

We call for a stronger objective oriented and integrated information security approach for all technical levels and non-technical units/partners over the whole value chain. This offers a wide range of most different interdisciplinary future research directions.

Acknowledgments The research leading to these results was partially funded by the Tyrolean business development agency through the Stiftungsassistentz QE—Lab and the COSEMA project, which is part of the Translational Research program.

References

1. ISO/IEC27004 (2009) ISO/IEC 27004 information technology, security techniques, information security management measurement. International Standard Organization, Geneva, Switzerland
2. National Institute of Standards and Technology (NIST) Performance measurement guide for information security, special publication 800-55 revision 1. Available at <http://csrc.nist.gov/publications/-nistpubs/800-55-Rev1/SP800-55-rev1.pdf>

3. IT Governance Institute (ITGI) (2007) COBIT 4.1: Framework, control objectives, management guidelines, maturity models. IT Governance Institute, Rolling Meadows, IL
4. OGC (Great Britain Office of Government Commerce) (2007) Service design (SD): ITIL, The Stationery Office (TSO), London
5. Sowa S, Tsinas L, Gabriel R (2009) BORIS—Business oriented management of information security, managing information risk and the economics of security. In: Johnson EM (ed) *Managing information risk and the economics of security*. Springer US, New York, pp 81–97
6. Savola R (2007) Towards a security metrics taxonomy for the information and communication technology industry. In: International conference on software engineering advances, 2007. ICSEA 2007, pp 60–66
7. Da Veiga A, Eloff JHP (2007) An information security governance framework. *Inf Syst Manag* 24(4):361–372
8. von Solms SH, Solms RV (2009) *Information security governance*. Springer, New York
9. Humphreys E (2007) Implementing the ISO/IEC 27001 information security management system standard. Artech House, Boston
10. ISO/IEC27002 (2005) ISO/IEC 27002:2005 information technology, security techniques, code of practice for information security management. International Standard Organization, Geneva, Switzerland
11. Böhme R (2010) Security metrics and security investment models. In: Echizen I, Kunihiro N, Sasaki R (eds) *Advances in information and computer security*, Springer Berlin/Heidelberg, pp 10–24
12. Herrmann DS (2007) Complete guide to security and privacy metrics: measuring regulatory compliance, operational resilience, and ROI. Auerbach, Boca Raton
13. Vaughn RB Jr, Henning R, Siraj A (2003) Information assurance measures and metrics—state of practice and proposed taxonomy. In: *System sciences, 2003. Proceedings of the 36th annual Hawaii international conference*, pp 10–19
14. ISACA (2009) An introduction to the business model for information security. Available at <http://www.isaca.org/Knowledge-Center/Research/Documents/Intro-Bus-Model-InfoSec-22Jan09-Research.pdf>
15. Verendel V (2009) Quantified security is a weak hypothesis. In: NSPW'09: new security paradigms workshop, Oxford, UK, 8–11 September 2009
16. ISO/IEC27001 (2005) ISO/IEC 27001:2005 information technology, security techniques, information security management systems requirements. International Standard Organization, Geneva, Switzerland
17. IT Governance Institute (ITGI) (2006) *Information security governance: guidance for boards of directors and executive management*, IT Governance Institute, Rolling Meadows, IL
18. Saint-Germain R (2005) Information security management best practice based on ISO/IEC 17799. *Inf Manag J* 39(4):60–66
19. Bulgurcu B, Cavusoglu H, Benbasat I (2010) Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness. *MIS Q* 34(3):523–A7
20. Puhakainen P, Siponen M (2010) Improving employees' compliance through information systems security training: an action research study. *MIS Q* 34(4):767–A4
21. McGee AR, Vasireddy SR, Chen Xie SR, Picklesimer DD, Chandrashekhara U, Richman SH (2004) A framework for ensuring network security. *Bell Labs Tech J* 8(4):7–27
22. Trèek D (2003) An integral framework for information systems security management. *Comput Secur* 22(4):337–360
23. Herath T, Herath H, Bremser WG (2010) Balanced scorecard implementation of security strategies: a framework for IT security performance management. *Inf Syst Manag* 27(1):72–81
24. Jaquith A (2007) *Security metrics: replacing fear, uncertainty, and doubt*. Addison-Wesley, Upper Saddle River
25. Baschin A (2001) *Die Balanced Scorecard für Ihren IT-Bereich: ein Leitfaden für Aufbau und Einführung*. Campus-Verl., Frankfurt/Main u.a

26. Kaplan RS, Norton DP (1992) The balanced scorecard—measures that drive performance. *Harv Bus Rev* 70(1):71–79
27. Kaplan RS, Norton DP (1996) The balanced scorecard: translating strategy into action. Harvard Business School Press, Boston
28. Kaplan RS, Norton DP (2008) Mastering the management system. *Harv Bus Rev* 86(1):62–77
29. Kaplan RS, Norton DP (2000) Having trouble with your strategy? Then map it. *Harv Bus Rev* 78(5):167–176
30. Hevner A, Chatterjee S (2010) Design research in information systems. Springer Science + Business Media, LLC., Boston
31. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S (2007) A design science research methodology for information systems research. *J Manag Inf Syst* 24(3):45–77
32. Stoll M, Breu R Development of stakeholder oriented corporate information security objectives (in press)
33. Stoll M, Breu R (2010) Information security measurement roles and responsibilities. In: IEEE international conference telecommunication and networking, TeNe2010, Bridgeport

Chapter 21

Experimenting with Watchdog Implementation on a Real-Life Ad hoc Network: Monitoring Selfish Behavior

Tirthankar Ghosh and Tian Hou

Abstract A watchdog monitors neighboring nodes in a multihop wireless ad-hoc network to determine whether they are forwarding packets or not. In this paper we have presented detailed study of our experiments with implementing watchdog functionality on a real-life ad-hoc network testbed. Experiments were conducted on our university campus to test the functionality under different network topologies.

21.1 Introduction

The concept of watchdog was first proposed by Marti et al. [1]. They implemented the function on Dynamic Source Routing (DSR) [2] protocol, where the nodes use source routing to send packets. Monitoring neighboring nodes in a source routing environment is easier as the nodes have the entire route in the header structure, and know all hops in the path. Over the course of several years many studies [3–5] have used the watchdog concept in monitoring neighboring nodes, but they have been limited mostly to simulation-based analysis.

There are two phases of neighbor monitoring in a multihop ad-hoc network scenario: monitoring during the control phase, and monitoring during the data transfer phase. During the first phase, watchdog monitors whether the neighboring

T. Ghosh (✉) · T. Hou

Department of Computer Science and Information Technology,
St. Cloud State University, St. Cloud MN, USA
e-mail: tghosh@stcloudstate.edu

T. Hou

e-mail: hoti0001@stcloudstate.edu

nodes are forwarding routing control messages, while in the second phase monitoring is done to ensure that the nodes in the path are forwarding data as agreed. In our study, we have implemented watchdog to monitor neighbors during the control phase on top of the Ad-Hoc On-Demand Distance Vector (AODV) [6] routing protocol.

The rest of the paper is organized as follows. [Section 21.2](#) discusses design and implementation of watchdog, followed by detailed discussion on experimental setup and tests in [Sect. 21.3](#). Finally, [Sect. 21.4](#) concludes the paper and highlights some of the future work.

21.2 Design and Implementation

21.2.1 Watchdog Design

Selfish behavior of a node in a multihop ad-hoc network results from its unwillingness to forward packets to save its battery power. As normal network operation depends on mutual trust among nodes in forwarding packets, selfish behavior results in disruption of the network operation. There are two broad approaches to ensure that nodes are forwarding packets as agreed: the first approach deals with providing incentives to nodes [3, 7, 8, 9], and the second approach is based on monitoring neighboring nodes [1, 3, 5]. Both approaches have their relative pros and cons. In the first approach a game theoretic strategy needs to be designed and implemented where the incentives outweigh nodes' selfish intention. On the other hand, the second approach is difficult to implement and consumes significant computing resources especially on power-intensive applications. Typically, watchdog is based on the second approach where each node monitors its one-hop neighbors. Due to the resource-intensive nature of watchdog, our approach to the problem is to have a dedicated watchdog in the network which will be implemented as a mobile agent. The watchdog will move around and monitor nodes whether they are forwarding packets. This study is conducted only with implementing watchdog to monitor control packets, we are working towards the implementation for monitoring data packets which will be presented in future.

The essential idea behind the design of watchdog is that by recording the IP address of previous hop from which RREQs were received by a neighboring node, watchdog determines if the neighboring node has forwarded the packets. Since AODV only stores the originator IP address and the destination IP address, the structure of AODV needs to be modified to include the IP address that is recorded when transmitting control packets. To fully achieve the goal of knowing if a particular neighboring node has forwarded RREQs, the watchdog utilizes the following functionalities of AODV:

- Buffering a record for every RREQ broadcasted by any node for a time interval that is defined for path discovery

- Identifying each broadcasted RREQ packet with the originator IP address and the RREQ ID
- Converting IP addresses from a network byte order to a string in dotted decimal format by which information about a node associated with receiving and forwarding RREQ packets can be learnt.

Overall, the watchdog traces which neighboring node has assisted in disseminating RREQs during route discovery. So the information obtained through this mechanism can be used to determine which node has not participated in the transmission of route request packets in the networks.

21.2.2 Implementation

We have used AODV-UU [10], Uppsala University implementation of AODV routing protocol, to implement our watchdog functionality. Figure 21.1 below show the flowchart to illustrate the process when a RREQ is received by watchdog.

The existing RREQ structure is modified to include a new field called *hold_addr*, which is used to hold the address of previous hop of RREQ during path discovery. The value of *hold_addr* is updated before a RREQ is forwarded to the next hop. This information stored by *hold_addr* is key to watchdog operation as it tells watchdog the ID of the previous node from which RREQ is received.

In AODV, any node acts on only the first RREQ and ignores any subsequent RREQs with the same RREQ ID. In the process of Fig. 21.1, instead of ignoring the duplicated RREQs, watchdog does the address checking. When a RREQ is received by watchdog, it checks if it has a record for the RREQ in its buffer. If it does not have one, then it determines if the previous hop of the RREQ is the originator by comparing the IP address of the previous hop of the RREQ with the IP address of originator in the RREQ. If the RREQ is not directly transmitted from the originator meaning that the previous hop of the RREQ is an intermediate node, which has retransmitted the packet, then watchdog prints out information telling it has seen that the forwarding node receiving the RREQ from its predecessor transmitted the RREQ. The original RREQ module uses function *ip_to_str()* to convert IP addresses from a network byte order to a string in dots-and-number format when logging AODV information. Watchdog utilizes this feature, so it can tell from which IP a forwarding node receives a control packet and re-forwards it to its neighbors. Notice that all AODV message types including RREQ are received via User Datagram Protocol (UDP) and that normal IP header processing applies. This enables the IP address of previous hop of RREQ as a parameter to be used directly in the function *rreq_process()* when processing control packets.

On the other hand, if a record is found for the RREQ, watchdog checks the RREQ record timer since each RREQ is buffered for a *PATH_DISCOVERY_TIME* interval. If the timer is not expired, and the RREQ is transmitted by

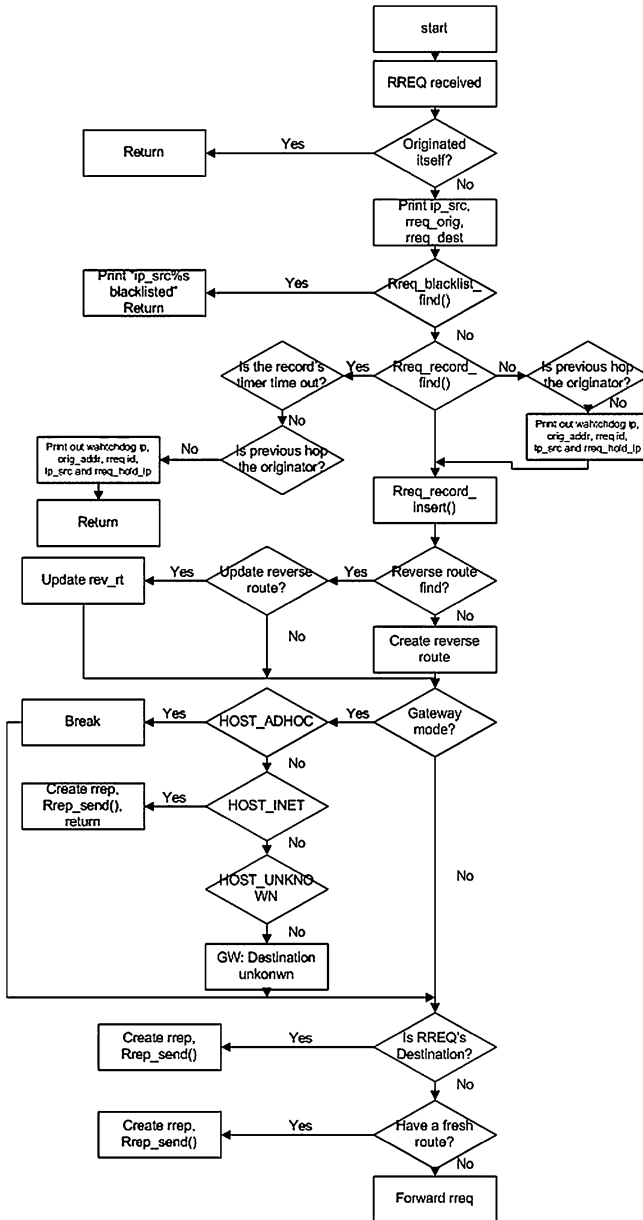
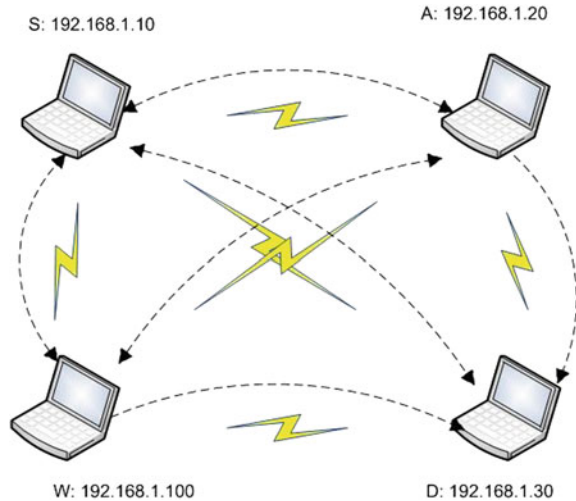


Fig. 21.1 Flowchart illustrating watchdog function in a node

an intermediate node, watchdog then prints out information saying that it has observed that the forwarding node receiving the RREQ from its previous hop transmitted the packet.

Fig. 21.2 Scenario 1: all nodes in one another's transmission range



21.3 Experimental Setup and Test

21.3.1 Experimental Setup

Our testbed was built with four HP Evo N600c laptops running Fedora 7 with kernel version 2.6.21-1, equipped with Orinoco Gold b/g wireless cards. The wireless card drivers were set up in ad-hoc mode using 802.11b standard. The transmission power of each card was reduced to 2 mW during each run of the experiments. This was done to scale down the area of coverage. We selected AODV-UU [10], the AODV implementation developed by Uppsala University, Sweden, as our routing protocol. MadWiFi driver was used for the wireless cards.

Three types of scenarios were developed for the test. To illustrate the test for each scenario, a corresponding network topology created for the scenario is explained with figure. The sample output generated for each scenario during the execution of each test is presented with interpretation for verification of watchdog operation. For simplicity, the following capital letters are used with the network setup to represent the name of each node deployed in networks:

S: source node

A, B: any intermediate node

D: destination node

W: node running watchdog

21.3.2 Tests

Scenario 1. The first scenario developed for the test is that all ad hoc nodes deployed in the network are within each other's transmission range. Figure 21.2 illustrates the network topology.

As in Fig. 21.2, the source S broadcasted RREQs containing destination IP address 192.168.1.30 to its neighbor W, A and D for a path discovery. Since all the nodes can see one another, W and A retransmitted the packets as they checked they were not the destination, however, S, A, and D did not re-process the duplicated RREQs, because AODV nodes act on only the first RREQ with the same RREQ ID. Thus, A is the only node that W has seen forwarding the RREQ. There are two log files (*aodvd.log* and *aodvd.rtlog*) that were generated for watchdog when running the test for scenario 1. The following shows a portion of AODV logs, which provides information for verification for the watchdog operation.

Let us look at the process that takes place when a RREQ is received by the watchdog written to */var/log/aodvd.log*.

```

12:54:34.341                rreq_process:
ip_src=192.168.1.10 rreq_orig=192.168.1.10
rreq_dest=192.168.1.30

12:54:34.341                rreq_record_insert:
Buffering RREQ 192.168.1.10 rreq_id=2
time=5600

12:54:34.342    log_pkt_fields:    rreq-
>flags: rreq->hopcount=0 rreq->rreq_id=2

12:54:34.342    log_pkt_fields:    rreq-
>dest_addr:192.168.1.30
rreq->dest_seqno=0

12:54:34.342    log_pkt_fields:    rreq-
>orig_addr:192.168.1.10
rreq->orig_seqno=4

12:54:34.342    rreq_forward: forwarding
RREQ src=192.168.1.10, rreq_id=2
12:54:34.342    aodv_socket_send: AODV msg
to 255.255.255.255 ttl=5 size=28

12:54:34.343                rreq_process:
ip_src=192.168.1.20 rreq_orig=192.168.1.10
rreq_dest=192.168.1.30

12:54:34.343                rreq_process:
Watchdog:192.168.1.100 has seen that node
192.168.1.20 receiving
RREQ(src=192.168.1.10, rreq_id=2) from
192.168.1.10 forwarded it to its neighbors

```

Table 21.1 Routing table for scenario 1

| Destination | Next hop | HC | St. | Seqno | Expire | Iface |
|--------------|--------------|----|-----|-------|--------|-------|
| 192.168.1.20 | 192.168.1.20 | 1 | VAL | 1 | 2431 | ath0 |
| 192.168.1.10 | 192.168.1.10 | 1 | VAL | 6 | 4950 | ath0 |
| 192.168.1.10 | 192.168.1.30 | 1 | VAL | 1 | 6135 | ath0 |

In the debug output of the process above that contains the RREQ information one can see that it contains the hop count, the RREQ ID, the originator and destination IP addresses, and the originator and destination sequence numbers (highlighted in red). The first entry in the output tells the RREQ is the first time received by watchdog from IP 192.168.1.10 (the previous hop of the RREQ). The following entry indicates that watchdog inserts a record for the RREQ. The second entry from the bottom in the output illustrates that watchdog receives a duplicated RREQ and then drops the packet, while the last entry (highlighted in blue) describes that A (192.168.1.20) forwarded the RREQ (src=192.168.1.10, rreq_id=2) for S (192.168.1.10).

Second is an example of the AODV internal routing table generated for watchdog, which is corresponding to the RREQ process illustrated in debug output previously.

As shown in the routing table in Table 21.1 above, watchdog establishes one hop with every other node accordingly since all nodes are within each other’s transmission range.

Scenario 2. In the second scenario, the network topology was developed with three hops. As illustrated in Fig. 21.3, S and A were placed out of each other’s transmission range with W in the middle. Similarly, W and D were deployed out of each other’s transmission range, and A was placed in between.

In Fig. 21.3, W monitors its one-hop neighbor node A as S sends out RREQs for seeking a route to D. In addition, W also serves as a forwarding hop in the process of the route discovery. As expected, W has observed that A retransmitted the RREQs to its neighbors. A sample log including debug outputs and routing table printouts drawn from *aodvd.log* and *aodvd.rtlog* files for scenario 2 is presented in below.

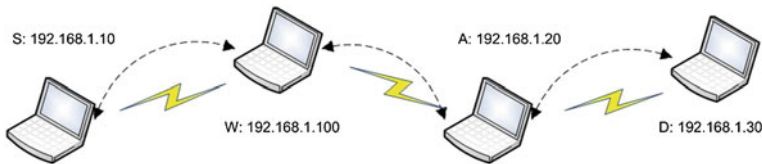


Fig. 21.3 Scenario 2: watchdog in 2-hop from destination

```

15:55:00.100                                rreq_process:
ip_src=192.168.1.10 rreq_orig=192.168.1.10
rreq_dest=192.168.1.30

15:55:00.101                                rreq_record_insert:
Buffering RREQ 192.168.1.10 rreq_id=81
time=5600

15:55:00.101    log_pkt_fields:    rreq-
>flags: rreq->hopcount=0 rreq->rreq_id=81
15:55:00.101    log_pkt_fields:    rreq-
>dest_addr:192.168.1.30           rreq-
>dest_seqno=24

15:55:00.101    log_pkt_fields:    rreq-
>orig_addr:192.168.1.10           rreq-
>orig_seqno=89

15:55:00.101    rreq_forward: forwarding
RREQ src=192.168.1.10, rreq_id=81

15:55:00.101    aadv_socket_send: AODV msg
to 255.255.255.255 ttl=4 size=28

15:55:00.104                                rreq_process:
ip_src=192.168.1.20 rreq_orig=192.168.1.10
rreq_dest=192.168.1.30

15:55:00.104                                rreq_process:
Watchdog:192.168.1.100 has seen that node
192.168.1.20 receiving
RREQ(src=192.168.1.10, rreq_id=81) from
192.168.1.100 forwarded it to its
neighbors

```

As shown in the debug output above, the first two entries illustrate that watchdog receives a new RREQ with ID “81” from IP 192.168.1.10 and buffers a record for the RREQ. The next three entries contain information of the RREQ (highlighted in red). Entry 6 and 7 indicate that the RREQ is forwarded and sent through a socket. The eighth entry indicates that a duplicated RREQ is received from A (192.168.1.20) and the last entry (highlighted in blue) tells that A forwarded the RREQ (src=192.168.1.10, rreq_id=81) for W (192.168.1.100).

The output above is a look at the process of RREP corresponding to the process illustrated for the RREQ previously. The first entry indicates a RREP is received.

Table 21.2 Routing table for scenario 2

| Destination | Next hop | HC | St. | Seqno | Expire | Iface | Precursor |
|--------------|--------------|----|-----|-------|--------|-------|--------------|
| 192.168.1.30 | 192.168.1.20 | 2 | VAL | 24 | 5367 | ath0 | 192.168.1.10 |
| 192.168.1.20 | 192.168.1.20 | 1 | VAL | 1 | 2366 | ath0 | |
| 192.168.1.10 | 192.168.1.10 | 1 | VAL | 89 | 4879 | ath0 | |

```
15:55:00.108
aodv_socket_process_packet: Received RREP
15:55:00.108      rrep_process:      from
192.168.1.20      about      192.168.1.10-
>192.168.1.30

15:55:00.108      log_pkt_fields:      rrep-
>flags: rrep->hcnt=1

15:55:00.108      log_pkt_fields:      rrep-
>dest_addr:192.168.1.30      rrep-
>dest_seqno=24

15:55:00.108      log_pkt_fields:      rrep-
>orig_addr:192.168.1.10      rrep-
>lifetime=6000

15:55:00.108      nl_send_add_route_msg:
ADD/UPDATE:      192.168.1.30:192.168.1.20
ifindex=8
15:55:00.109      rrep_forward: Forwarding
RREP to 192.168.1.10

15:55:00.109 aodv_socket_send: AODV msg
to 192.168.1.10 ttl=253 size=20

15:55:00.109      precursor_add: Adding
precursor 192.168.1.10 to rte 192.168.1.30
```

The second entry shows the RREP is transmitted by A (192.168.1.20). Entry 3, 4, and 5 reveal the information of the RREP. Entry 6 indicates a forward route is added. Entry 7 and 8 tell that RREP is forwarded and sent in the AODV message to S (192.168.1.10). Last entry adds precursor S to the route entry (192.168.1.30).

Table 21.2 shows the routing table of watchdog under scenario 2. As shown in the table, W (192.168.1.100) establishes two hops to D (192.168.1.30); whereas it creates one hop to both A (192.168.1.20) and S (192.168.1.10) respectively.

Scenario 3. In this last scenario, the network topology was developed for two hops. As illustrated in Fig. 21.4, S and W were placed out of each other's transmission range with A and B in the middle. Due to the limited number of mobile laptops, W was not deployed as the destination node in the network. Instead, it served as a forwarding node again. The reason to design such network topology is that W can monitor multiple neighbors in the network (while lacking of available

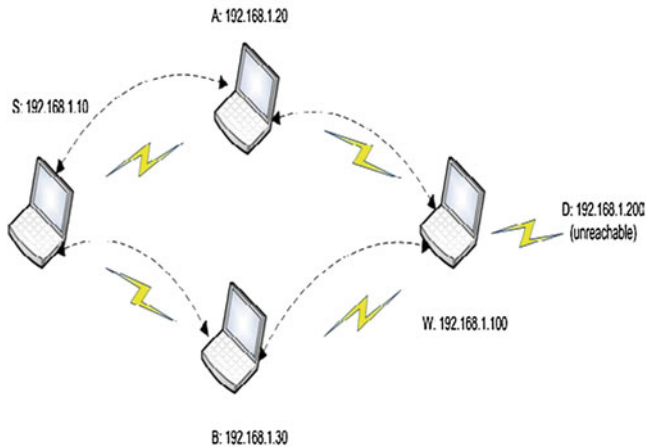


Fig. 21.4 Scenario 3: watchdog as destination

ad hoc devices, which is only in this case). To test the watchdog operation in such scenario, we simply pinged an unreachable IP address i.e., 192.168.1.200.

The following debug output and routing table printout provide information that verifies the witness of watchdog on A and B in the transmission of RREQs.

```

11:46:24.771          rreq_process:
ip_src=192.168.1.30 rreq_orig=192.168.1.10
rreq_dest=192.168.1.200

11:46:24.771          rreq_process:
Watchdog:192.168.1.100 has seen that node
192.168.1.30 receiving
RREQ(src=192.168.1.10, rreq_id=46) from
192.168.1.10 forwarded it to its neighbors

11:46:24.771          rreq_record_insert:
Buffering RREQ 192.168.1.10 rreq_id=46
time=500

11:46:24.771  log_pkt_fields:  rreq-
>flags: rreq->hopcount=1 rreq->rreq_id=46

11:46:24.771  log_pkt_fields:  rreq-
>dest_addr:192.168.1.200      rreq-
>dest_seqno=0

11:46:24.771  log_pkt_fields:  rreq-
>orig_addr:192.168.1.10      rreq-
>orig_seqno=48

```

```

11:46:24.771      rt_table_update:      rt-
>next_hop=192.168.1.20,
new_net_hop=192.168.1.30

11:46:24.771      nl_send_add_route_msg:
ADD/UPDATE:      192.168.1.10:192.168.1.30
ifindex=8

11:46:24.771      rreq_forward: forwarding
RREQ src=192.168.1.10, rreq_id=46

11:46:24.772      aodv_socket_send: AODV msg
to 255.255.255.255 ttl=33 size=28

11:46:24.772      rreq_process:
ip_src=192.168.1.20 rreq_orig=192.168.1.10
rreq_dest=192.168.1.200

11:46:24.772      rreq_process:
Watchdog:192.168.1.100 has seen that node
192.168.1.20 receiving
RREQ(src=192.168.1.10, rreq_id=46) from
192.168.1.10 forwarded it to its neighbors

```

Table 21.3 Routing table for scenario 3

| Destination | Next hop | HC | St. | Seqno | Expire | Iface |
|--------------|--------------|----|-----|-------|--------|-------|
| 192.168.1.30 | 192.168.1.30 | 1 | VAL | 1 | 2759 | ath0 |
| 192.168.1.10 | 192.168.1.30 | 2 | VAL | 2 | 48 | ath0 |
| 192.168.1.20 | 192.168.1.20 | 1 | VAL | 1 | 2760 | ath0 |

As shown in the context of the output above, the first and last two entries (highlighted in blue) need to be discussed. The first entry indicates that a new RREQ is received by watchdog from B (192.168.1.30) and the second entry tells that B forwarded the RREQ for S (192.168.1.10); whereas, the second last entry illustrates that the watchdog receives a duplicated RREQ with the same ID from A (192.168.1.20) and the last entry depicts that A forwarded the RREQ for S (192.168.1.10).

For other entries, the messages are interpreted in order as follows: The third entry buffers a record for the RREQ. Entry 4, 5, and 6 display the RREQ information (highlighted in red). Entry 7 updates the routing table. Entry 8 adds the forward route. Entry 9 and 10 indicate the RREQ is forwarded and sent in AODV message.

The routing table in Table 21.3 above illustrates the establishment of hops of watchdog with other nodes. As shown in the routing table, watchdog establishes one hop with both A and B and two hops with S.

21.4 Conclusion and Future Work

In this experimental study, we have implemented watchdog on a real-life wireless multihop ad-hoc network testbed, and conducted experiments with different network topologies to test its functionality. The implementation was done on Ad-Hoc On-Demand Distance Vector (AODV) routing protocol.

The study discussed here has been conducted only with implementing watchdog to monitor control packets, we are working towards the implementation for monitoring data packets which will be presented in future. Also, watchdog needs to be implemented on all nodes, and a detailed study needs to be conducted to monitor the energy utilization of the nodes under such condition.

References

1. Marti S, Giuli TJ, Lai K, Baker M (2000) Mitigating routing misbehavior in mobile ad hoc networks. In: Proceedings of the 6th annual international conference on Mobile computing and networking (MobiCom), August 06–11, 2000, Boston, MA
2. Johnson DB, Maltz DA (1999) The dynamic source routing protocol for mobile ad hoc networks. Internet draft, MANET working group, IETF, Oct
3. Buchegger S, Boudec JL (2002) Performance analysis of the CONFIDANT protocol (cooperation of nodes: fairness in dynamic ad-hoc networks). In: MOBIHOC'02, Switzerland, June 9–11
4. Pirzada AA, McDonald C (2004) Establishing trust in pure ad-hoc networks. In: 27th Australian computer science conference, the University of Otago, Dunedin, New Zealand
5. Ghosh T, Pissinou N, Makki K (2005) Towards designing a trusted routing solution in mobile ad hoc networks. ACM J Mobile Netw Appl (MONET) Special Iss Non-Cooperative Wirel Netw Comput 10(6):985–995
6. Perkins C, Royer E (1999) Ad hoc on-demand distance vector routing. In: Proceedings of IEEE workshop on mobile computing systems and applications
7. He Q, Wu D, Khosla P (2004) SORI: a secure and objective reputation-based incentive scheme for ad-hoc networks. In: WCNC 2004
8. Michiardi P, Molva R (2002) CORE: a collaborative reputation mechanism to enforce node cooperation in mobile ad hoc networks. In: Proceedings of the 6th IFIP communications and multimedia security conference, Portoroz, Slovenia
9. Buttyán L, Hubaux JP (2002) Stimulating cooperation in self-organizing mobile ad hoc networks. MONET J Mobile Netw 10(4):477–486
10. AODV Implementation, (AODV-UU), Department of Information Technology, Uppsala University (Sweden). <http://core.it.uu.se/core/index.php/AODV-UU>

Chapter 22

Power Consumption Evaluation for Cooperative Localization Services

Patrick Seeling

Abstract Current mobile applications oftentimes require power-consuming localization services. In this paper, we outline the co-localization approach, where nodes share their location with peers, enabling a reduction in the costs of localization when a precise location fix is desired. While several works in this domain compare the accuracy of localization techniques in cooperative scenarios, we focus our evaluation on the power consumption and accuracy that can be achieved. We present a first model and evaluation using statistics and traces derived from two human mobility models. We find that for 15 min intervals of location requests, a cooperative localization approach can reduce the costs associated with localization if half of the nodal peer encounters are with location-sharing nodes and GPS is usable about half the time.

22.1 Introduction

Location-based services are vital components of many application scenarios in the mobile design space. Two of the most commonly utilized localization methods are (i) The satellite-based Global Positioning System (GPS), which delivers a high accuracy, and (ii) Cellular or other wireless network based (triangulation in some cases), which typically delivers a rough approximation.

In [1], the authors investigate several different methods of maintaining a constant location fix on a mobile terminal by different means, amongst them location sharing of neighboring devices via Bluetooth for different time intervals.

P. Seeling (✉)

Department of Computer Science, Central Michigan University,
Mount Pleasant, MI 48859, USA
e-mail: pseeling@ieee.org

While for some application scenarios a constant location fix is desirable, other scenarios, such as the one we consider here, can benefit from regarding a desired location determination in regular intervals.

Other research efforts in the past were directed at optimizing the cooperation in localization problems, to a large degree in sensor network environments, see, e.g., [2, 3], while other recent research efforts were directed towards cellular networks, see, e.g., [4–6], and indoor scenarios, see, e.g., [7].

Our approach is different from these previous efforts in that we provide a general evaluation framework combining the accuracy and the power consumption in cooperative (co-) localization and individual localization scenarios. We additionally broaden the evaluation to include ZigBee-enabled devices for obtaining a close location fix, as in future networking scenarios, proliferated sensor networks can be utilized to provide additional information in a localized context even for cellular users. We utilize previous findings for the power consumption of wireless network interfaces outlined in, e.g., [8, 9], to determine numeric values for our performance evaluation.

The remainder of this paper is structured as follows. In Sect. 22.2, we present the overall system approach which we follow throughout the paper and its corresponding model in Sect. 22.3. We subsequently provide a numeric performance analysis of our approach in Sect. 22.4 before we conclude in Sect. 22.5.

22.2 Co-Localization Overview

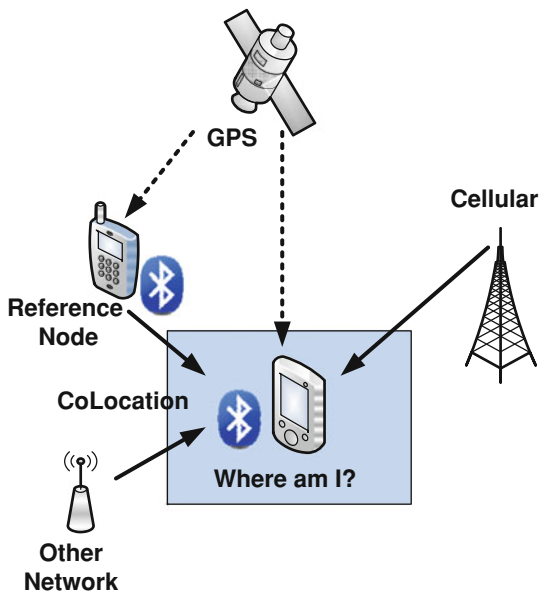
The co-localization model allows nodes that are unable to obtain a location fix through GPS or want to conserve battery power to obtain their location information through neighboring nodes. The neighboring nodes can be fixed or mobile and are presumed to update their respective locations over time. We illustrate the overall co-localization model in Fig. 22.1.

In the co-localization model, the mobile terminal under consideration is presumed to be able to obtain a location fix at any time through the cellular network as a fallback. Obtaining the location based on cellular network (including triangulation), however, infers heavy accuracy penalties. In contrast, if being able to use the GPS, the mobile terminal can determine its location with high accuracy. This, however, comes typically at the expense of higher power consumption during the period of obtaining a satellite fix.

If some nodes would share their self-determined location through low-power local networks, e.g., Bluetooth, other nodes could infer their location based on signal strength and shared location to an accuracy that is higher than that obtained through the cellular network (including potential triangulation), but at a cost lower than incurred by utilizing GPS. This comes especially into play when regarding indoor scenarios, where a GPS fix might not be possible, however inferring a location based on nodes close to the outside or sensor networks that announce their location indoors, e.g., indoor positioning systems, is feasible.

Other Network GPS CoLocation Where am i ? Reference Node Cellular

Fig. 22.1 Overview of the co-localization model, where a node can obtain its location through the cellular network, GPS, other nodes that are sharing their current or last location, or other networks, such as local sensor networks



22.3 Co-Localization Model

The performance metric we use throughout this paper is the localization cost C , determined as energy cost of location inaccuracy, i.e., how much energy was spent on determining the distance from the device from its actual position. Throughout this paper, we assume that nodes strive to determine their location as accurately as possible and with the least amount of energy to maximize their utility.

We denote t_{GPS} as the time required to obtain a GPS fix on the current location and let P_{GPS} denote the power consumed by the mobile device to obtain the GPS fix. We denote the inaccuracy of an obtained GPS location as I_{GPS} as simple distance in m from a device's actual position, e.g., $I_{GPS} = 3\text{m}$. We define

$$L_{GPS} = t_{GPS} \cdot P_{GPS} \cdot I_{GPS} \quad (22.1)$$

as performance metric (price) for using the GPS components of the mobile device in obtaining the current location.

Following the outlined model, we denote the inaccuracy for a location obtained through Bluetooth as I_{BT} , whereby we assume that a location obtained through Bluetooth connections is more inaccurate, e.g., $I_{BT} = 13\text{m}$. In this scenario, the accuracy for Bluetooth is determined by the accuracy of a GPS location fix and the typical range of a Bluetooth connection, here assumed with 10 meters. We similarly obtain a ZigBee location accuracy of $I_{ZB} = 28\text{ m}$, assuming a general range of 25 m for ZigBee network nodes. We provide an overview of the assumed values for the various location technologies evaluated in this paper in Table 22.1.

Table 22.1 Typical values for the location determination using GPS, cellular network, Bluetooth and ZigBee radio interfaces

| Type | Time | Power | Inaccuracy | L |
|----------------|---------|---------|------------|-------|
| Type | t [s] | P [W] | I [m] | [Wsm] |
| GPS | 15 | 0.4 | 3 | 18 |
| Network (NET) | 1 | 0.25 | 1,000 | 250 |
| Bluetooth (BT) | 5 | 0.15 | 13 | 9.75 |
| ZigBee (ZB) | 3 | 0.075 | 28 | 6.3 |

In addition to the location determination outlined in Eq. 22.1, the probability for being able to determine a location through neighbor inquiry has to be taken into account. Let $P(C)_{BT,ZB}$ denote the probability that a neighboring node can be queried for its location and that the location has been recently obtained (i.e., is relevant). We denote that GPS can be used (which is not always the case, e.g., indoors) using the indicator function $P(\cdot)_{GPS}$, i.e., if indoors $P(\cdot)_{GPS} = 0$ and if outdoors $P(\cdot)_{GPS} = 1$.

In the following, we assume that the cellular network location is always available as a fall-back alternative to an initial attempt of co-localization, i.e., the mobile device is presumed to be always connected to a cellular network. performance metric

$$C_{GPS} = P(\cdot)_{GPS} \cdot L_{GPS} + (1 - P(\cdot)_{GPS}) \cdot (L_{GPS} + L_{NET}),$$

where $L_{NET} = t_{NET} \cdot P_{NET} \cdot I_{NET}$ and L_{GPS} as in Eq. 22.1.

In turn, we derive the Bluetooth co-localization performance metric as

$$C_{BT} = L_{BT} + (1 - P(C)_{BT}) \cdot (L_{BT} + C_{GPS}) \quad (22.3)$$

and the ZigBee co-localization performance metric synonymously. Using the values outlined in Table 22.1, we calculate $C_{NET} = 250[\text{Wsm}]$ and using Eq. 22.2, we calculate $C_{GPS} = 18 [\text{Wsm}]$ for outdoors and $C_{GPS} = 268 [\text{Wsm}]$ for indoors.

22.4 Performance Evaluation

For an initial evaluation of the advantage of the co-localization approach, we use a Bluetooth scenario. Initially, we need to quantify $P(C_{BT})$ as function of (i) the probability that other nodes are encountered and in range $P(e)$, (ii) the probability that the encounter lasts for more than the time required to scan for devices and services $P(t_{enc} \geq t_{BT})$ and (iii) The probability that the node encountered for a specific time shares its location $P(S)$. In the following, we evaluate the co-localization approach based on two exemplary Human Mobility Models (HMM), referring to, e.g., [10], for a comparative overview of mobility models.

22.4.1 Working Day Model

We use the contact durations presented in [11] as $P(t_{enc} \geq 14s) \approx 0.85$ and the probability for an encounter given that a location determination is required can be estimated similarly from [11] as $P(e) \approx 0.4$ assuming that approximately every 15 min a localization is required. (We note that this is in contrast to [1], where continuous localization requirements are assumed.) Using these example values, we can calculate $P(C_{BT}) \approx 0.17$, assuming that about half the encountered nodes share their location, i.e., assuming $P(S) = 0.5$. This yields an approximate $C_{BT} \approx 240$ [Wsm] indoors and $C_{BT} \approx 33$ [Wsm] outdoors. Similarly, we obtain $C_{ZB} \approx 240$ [Wsm] indoors and $C_{ZB} \approx 33$ [Wsm] outdoors for ZigBee localization, assuming $P(C_{ZB}) = 0.2$.

We note that for this scenario, it is always beneficial for a node to use the co-localization approach when inside and requiring location updates approximately every 15 min. Overall, we note that this model follows the typical preference of power consumption of the different network technologies in play.

22.4.2 Small World in Motion Traces

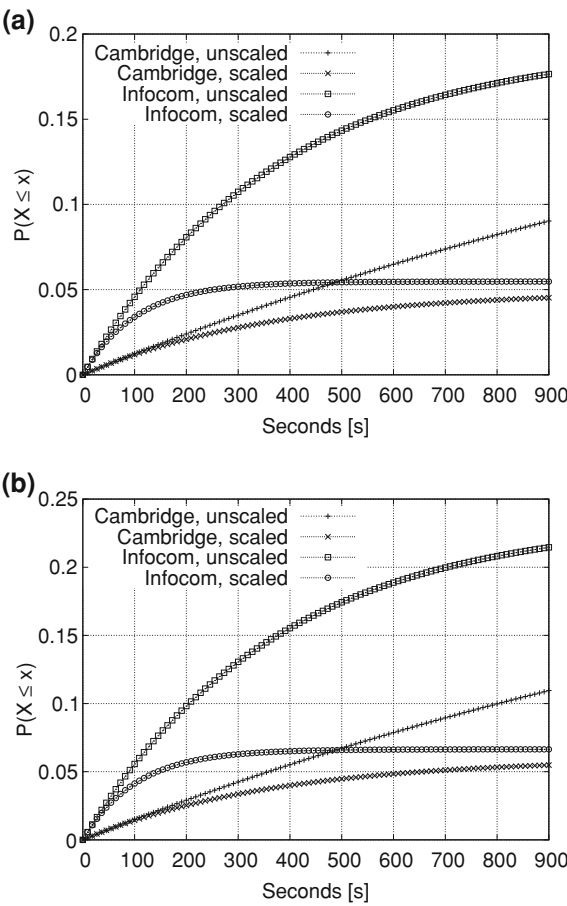
In the following, we evaluate the performance of the different co-localization approaches based on the traces generated by the Small World in Motion (SWIM) model's implementation [12]. Specifically, we evaluate the *Cambridge* and *Infocom* original as well as phoenix-scaled traces from 2006, which are publicly available at [13]. These traces contain the meeting and leaving timestamps for 2,000 nodes generated by the SWIM model and based on original captured data. We perform additional processing on these traces by initially dropping entries for which no meeting match can be found (i.e., where we cannot calculate a connection duration). Note that we assume that an encounter in the trace files corresponds to the nodes being in range, independently of the network technology used for location sharing. Next, we calculate the connection durations at the full second time scale and calculate the probabilities for an encounter lasting at least t_{BT}, t_{ZB} seconds from the thus generated data as $P(T \geq t_{BT,ZB})$ for each trace as in Table 22.2.

In addition, we calculate the probability for an encounter occurring with a node that shares its location and assume that half of the nodes observed will be sharing their location with others. For this outlined scenario, we determine the probability for an encounter with a sharing node as follows. We assume that all even numbered nodes in the traces have sharing enabled, while all uneven numbered nodes in the trace have location sharing disabled. From the trace, we can directly determine the probability of encounters of even nodes, assuming Independence of both underlying events. From the various traces, we furthermore determine the average inter-arrival times for meeting a location-sharing node as in Table 22.2, assuming an exponential distribution. We illustrate the resulting probabilities of

Table 22.2 Probabilities for nodal encounter lasting at least t_{BT}, t_{ZB} seconds and inter-arrival times for the encounters captured in the original and scaled Cambridge 2006 and Infocom 2006 SWIM traces [12]

| | Cambridge | | Infocom | |
|--------------------|-----------|----------|----------|---------|
| | Original | Scaled | Original | Scaled |
| $P(T \geq t_{BT})$ | 0.23659 | 0.049458 | 0.19354 | 0.05473 |
| $P(T \geq t_{ZB})$ | 0.28699 | 0.059955 | 0.23533 | 0.06642 |
| λ^{-1} | 1872.6 | 364.76 | 370.56 | 102.94 |

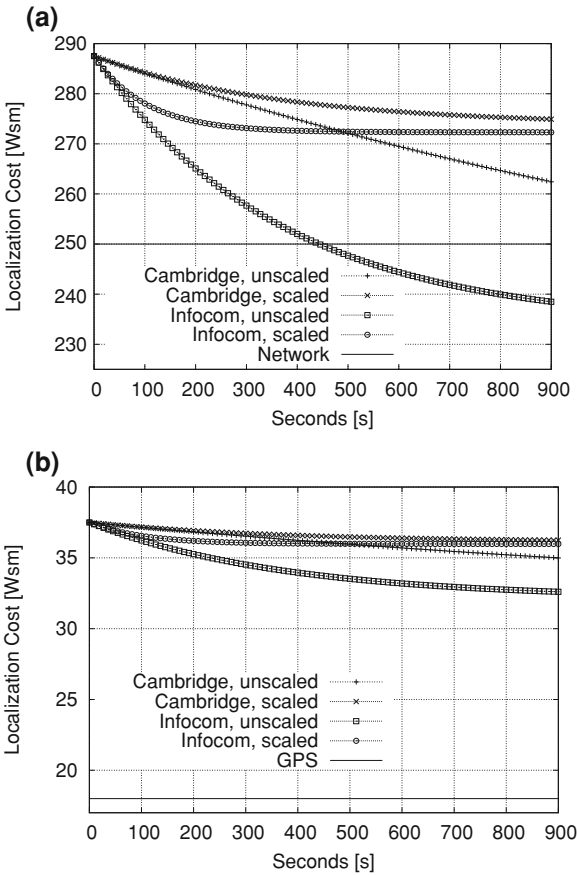
Fig. 22.2 Probabilities for encountering a location sharing node within the next x seconds for at least t_{BT}, t_{ZB} seconds calculated from the original and scaled Cambridge and Infocom 2006 SWIM traces [12] assuming half the nodes share their location



successful determination of a location by using a nearby node in Fig. 22.2 for Bluetooth and in Fig. 22.2 for ZigBee.

We observe that for Bluetooth and ZigBee cases alike, the scaled traces yield higher probabilities of encountering a sharing node compared to their unscaled

Fig. 22.3 Localization costs for Bluetooth co-localization C_{BT} if 50 of encountered peer nodes share their location for indoor and outdoor scenarios calculated from the original and scaled Cambridge and Infocom 2006 SWIM traces

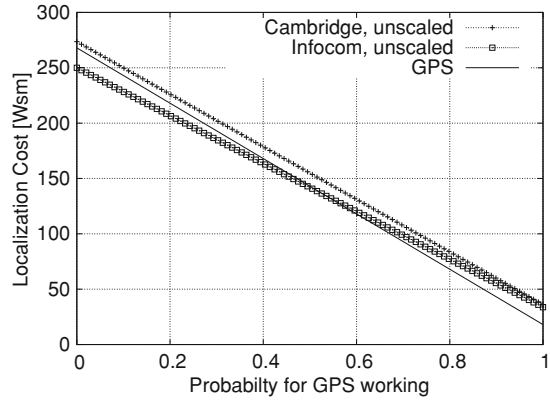


counterparts. We additionally observe that the overall chance of encountering a sharing node using the Cambridge traces is fairly small (around five percent) when compared to the unscaled Infocom trace, which approaches 18 and 21 for Bluetooth and ZigBee cases, respectively.

Next, we investigate the effect of these probabilities on the co-localization performance metric for the Bluetooth scenario, noting that the ZigBee approach yields similar results. We illustrate the co-localization costs as function of the probability of encountering a sharing node as C_{BT} in Fig. 22.3 for indoors and outdoors.

We observe that for the indoors case, the co-localization costs are significantly higher than for the network-based localization costs for both Cambridge and the scaled Infocom traces for times of encountering sharing nodes. The unscaled Infocom trace, however, exhibits a break-even point, after which a cost-minimizing node would benefit from switching to the co-localization approach and first attempt to discover its location based on neighboring peers. In the outdoors scenario illustrated in Fig. 22.3, we observe that the costs associated with the co-localization

Fig. 22.4 Localization costs for expected encounters after 444 seconds when nodes have different probabilities of being indoor or outdoor based on the original (unscaled) Cambridge and Infocom 2006 SWIM traces



approach cannot reach the assumed low localization costs for GPS-based location determination. This can readily be explained with the low probability of encountering a sharing node as well as the smaller differences in location inaccuracy and power, which results in closer values of C_{BT} and C_{GPS} when $P(\cdot) = 1$.

The break-even case occurs at 444 integer seconds of time to encounter a sharing node in the unscaled Infocom scenario, i.e., here would be indifference between choosing the co-localization or the network approach when regarding the indoors scenario. We now investigate the influence of the probability of being inside (or the assumed fraction of time a node spends inside) on the costs of the co-localization approach in comparison to the GPS costs in Fig. 22.4.

We note that throughout the range of possibilities of the GPS acquiring a fix by being outside, using the co-localization approach with Bluetooth based on the unscaled Cambridge trace never yields a benefit. For movement patterns that are based on the unscaled Infocom trace, however, we note that the costs for using the co-localization approach are lower if approximately half the time a GPS fix is not possible.

22.5 Conclusion and Outlook

In this paper, we introduced the co-localization approach for localization of nodes through neighboring nodes. Through evaluation of a cost-based function that incorporates location inaccuracy and power used with human mobility models, we found that this is a feasible approach which minimizes localization costs if overall nodes spend approximately half their time in places without GPS access and half of the encountered peer nodes share their location with others.

In future works, we intend to (i) Perform simulations to validate our approach in dynamic environments and (ii) Develop demonstration implementations across multiple mobile operating systems to evaluate the actual device-specific saving that can be realized following our approach.

References

1. Paek J, Kim J, Govindan R (2010) Energy-efficient rate-adaptive gps-based positioning for smartphones. In: Proceedings of the 8th international conference on mobile systems, applications, and services, ser. MobiSys '10. ACM, New York, pp 299–314.<http://doi.acm.org/10.1145/1814433.1814463>
2. Reghelin R, Fröhlich A (2006) A decentralized location system for sensor networks using cooperative calibration and heuristics. In: Proceedings of the 9th ACM international symposium on modeling analysis and simulation of wireless and mobile systems. ACM, pp 139–146
3. Patwari N, Ash J, Kyperountas S, Hero A III, Moses R, Correal N (2005) Locating the nodes: cooperative localization in wireless sensor networks. *Signal Proc Mag IEEE* 22(4):54–69
4. He Z, Ma Y, Tafazolli R (2011) Cooperative localization in a distributed base station scenario. In: Vehicular technology conference, VTC Spring, 2011 IEEE, 73rd IEEE, pp 1–5
5. Chen Y, Chou C, Chen C (2011) Cooperative localization for wireless and mobile social networking service (SNS). In: Wireless communications and mobile computing conference (IWCMC), 7th International. IEEE, pp 1952–1957
6. Masajedian S, Khoshbin H (2004) Cooperative location management method in next generation cellular networks. In: Computers and communications, Proceedings ISCC 2004. Ninth Int Symp IEEE, vol 1, pp 525–530
7. Thompson B, Buehrer R (2011) Cooperative indoor position location using reflected estimations. In: Wireless conference 2011-sustainable wireless technologies (European wireless), 11th European. VDE, pp 1–6
8. Pering T, Agarwal Y, Gupta R, Want R (2006) Coolspots: reducing the power consumption of wireless mobile devices with multiple radio interfaces. In: Proceedings of the 4th international conference on mobile systems, applications and services, ser. MobiSys '06, ACM, New York, pp 220–232. <http://doi.acm.org/10.1145/1134680.1134704>
9. Evaluation of context distribution methods via bluetooth and wlan: Insights gained while examining battery power consumption. 5th international ICST conference on mobile and ubiquitous systems: computing, networking and services, May 2008
10. Karamshuk D, Boldrini C, Conti M, Passarella A (2011) Human mobility models for opportunistic networks. *IEEE Communications Magazine*, 2011, accepted for publication
11. Ekman F, Keränen A, Karvo J, Ott J (2008) Working day movement model. In: Proceeding of the 1st ACM SIGMOBILE workshop on mobility models, ser. MobilityModels '08, ACM, New York, pp 33–40. Online]. Available: <http://doi.acm.org/10.1145/1374688.1374695>
12. Mei A, Stefa J (2009) Swim: a simple model to generate small mobile worlds. In: INFOCOM 2009, IEEE, pp 2106–2113
13. Kosta SMA, Stefa J Swim: small world in motion—modeling human mobility, website. <http://swim.di.uniroma1.it>

Chapter 23

A Modified Banker's Algorithm

Youming Li

Abstract In this paper, we propose a new algorithm for the classical deadlock avoidance problem. The new algorithm has the simplicity as Banker's algorithm in terms of data structures used in the algorithm; in particular, there is no graph model of any kind is required. The original Banker's algorithm has time complexity of $O(n^3d)$, where n is the number of processes and d is the number of resources (as per certain literatures, it is misquoted as $\Theta(n^2d)$ which refers as the safety check part), in comparison, our algorithm has time complexity of $O(n^2d^2 + ndM)$, where M is the total number of resource units. When the total resource units are fixed as a constant and number of resource types is fixed, which is a reasonable assumption, the complexity of our algorithm is $O(n^2d^2)$, a significant improvement over the original Banker's algorithm. The space-complexity of our algorithm is $O(nd^2)$ which is worse than the original Banker's algorithm.

23.1 Introduction

In operating systems, there exist at least three strategies dealing with deadlocks for concurrent processes, namely deadlock prevention, avoidance and detection, in the decreasing order of handling extent. There are numerous literatures on deadlock detection and avoidance. To name a few, [1–5] and [6, 7] are among classical papers. See also textbooks [8, 9] on operating systems. Our late works [10, 11] proposed two algorithms for the deadlock detection.

Y. Li (✉)

Department of Computer Sciences, Georgia Southern University,
Statesboro, GA 30460, USA
e-mail: yming@georgiasouthern.edu

Deadlock avoidance tries to contain the system in a safe state so the deadlock will never occur; The Banker's algorithm is a classical algorithm on deadlock avoidance. We shall study this algorithm by proposing a new algorithm for its core part, namely the safe state detection.

We first introduce the related context and terminologies that will be used throughout in this paper. In a concurrent computer system, let $S = \{P_i : i = 1, \dots, n\}$ be a set of n processes, and $\{T_j : j = 1, \dots, d\}$ be a set of d reusable resources types competed by the processes. Each resource type has the M_j available units in total. Each process P_i is already allocated with the resources, grouped as a vector $A_i = (a_{i1}, \dots, a_{id})^t$. At the same time, it is requesting additional resources, grouped as a vector $R_i = (r_{i1}, \dots, r_{id})^t$. Further each process has a maximum demand vector $X_i = (x_{i1}, \dots, x_{id})^t$ for the resources.

The core of the Banker's algorithm is the safe state detection. The system is said to be in a safe state if there exists a sequence of execution so that the deadlock can be avoided. Namely for there exists a $\sigma \in S_n$ the permutation group on the set $\{1, \dots, n\}$, so that

$$R_{\sigma(i)} \leq F + \sum_{1 \leq k \leq i-1} A_{\sigma(k)}$$

Here $F = (f_1, \dots, f_d)^t$ is the free resource vector, and the addition and comparison for vectors is done component-wise.

The safety detection part in the original Banker's algorithm is basically a sequence of linear searches with each step being finding a process whose request vector is less than or equal to the current available resource vector. Therefore the worst time-complexity for this part is of the order $\sum_{1 \leq k \leq n} (dk) = \Theta(n^2d)$. Finally, the whole Banker's algorithm tests safety property by pretending the resource request for each process is satisfied. Thus the total complexity for the Banker's algorithm is thus of order $O(n^3d)$.

Although the polynomial dependence of the complexity on n makes it tractable, the cubic dependence is undesirable when the number of processes is large. In this paper, we propose a strategy to handle the safe state detection part with the time-complexity that is roughly linearly dependent on the number of processes. More specifically, the safety detection part algorithm is of the time-complexity $O[(nd + \sum_{1 \leq j \leq d} M_j)d]$.

Thus the complexity for the Banker's algorithm is

$$O[(nd + \sum_{1 \leq k \leq d} M_k)nd],$$

which is of quadratic dependence on the number of processes.

A realistically reasonable assumption is that the numbers of resource type M_k are fixed constants (and relatively small). In this case, the complexity of the newer Banker's algorithm is of the order

$$O(d^2n^2 + Cnd^2) = O(n^2d^2).$$

This can be further reduced to $O(n^2)$ when the number of processes is much larger than the number of resource types and total units of each resource type.

For the space-complexity, our algorithm is slightly worse than that of the original Banker's algorithm depending on d , the size of the number of resource types.

23.2 Description of the Algorithm

Now we describe the algorithm that is based on a kind of greedy method. We use a modified "Algorithm B" from our early work [11]. That algorithm was created for deadlock detection. We stress that the safe state detection algorithm is the logical negation in terms of the output, of the deadlock detection algorithm. Therefore they have the same time complexity. Here, for the sake of completeness, we shall briefly describe the safety detection algorithm in its own context for the purpose of deadlock avoidance, and simplified in a number of ways.

23.2.1 The Description of Safety Detection Algorithm

Input with Data Structures: System of n processes with the following information: the total available resource vector $M = (M_1, \dots, M_d)^t$, the processes' allocated resource matrix $A = (A_1, \dots, A_n)$, the processes' resource requesting matrix $R = (R_1, \dots, R_n)$, and the processes' maximum demand matrix $X = (X_1, \dots, X_n)$.

Output: TRUE is the system can be made safe, namely, deadlock free; FALSE otherwise.

Extra Data Structures: d permutations on the set $\{1, \dots, n\}$ and $d \times d$ matrix of permutations. The first one can be implemented as $n \times d$ matrix or a two-dimensional array, and second one as a three-dimensional $d \times d \times n$ array.

Steps:

1. Compute resource available vector $F = M - A$;
2. Trivial Rejection Check: If any column of R is not less than either F or the corresponding column in the maximum demand matrix X , return FALSE;
3. Use counting sort algorithm to sort each resource types request vector by the processes; record the permutation vector $S = (s_1, \dots, s_d)$ Compute the $d \times d$ permutation matrix $T = [t_{ij}]$ where $t_{ij} = s_i s_j^{-1}$ for all i, j from 1 to d .
4. Greedily checks which process can have its request satisfied. This is done by fixing all and any resource type i check largest position in the sorted resource vector in R_{si} whose value is less than F_s . To verify that this last position's corresponding process, say process m , can have its request satisfied for all other resource types, the algorithm simply uses the already computed permutations p_{**m} from the permutation matrix P , to do a trivial comparison to other

available resource types. Further if the process is verified satisfied for a resource type, it will not be verified in the next iteration.

There are 2 possibilities in Step 4

- (i) The algorithm cannot find a process with its resource request being satisfied. In this case, simply return FALSE.
- (ii) The algorithm found a process with its all resource type satisfied. In this case, the algorithm de-allocates the process's allocation vector (making it available) and repeats Step 4 for other processes:

$$F = F + A_k$$

5. Return TRUE.

23.2.2 Time and Space Complexities

The core difference between our safety detection algorithm and the one used in original Banker's algorithm is of course the Step 4. Here we used top-down approach guided by greedy philosophy. Namely the algorithm always selects a process with largest request for a resource type, in the hope that if it can be scheduled, it then releases largest free resource allocation for that type. The time efficiency of our algorithm lies in the pre-computed permutation matrix which keeps track of all sorted order requests, and the fact that this checking process is done in a disjoint way, i.e., if the process is already checked against a resource type, then next iteration will skip the very same procedure. The correctness of the algorithm can be proved similarly as in [10].

We omit here the full analysis of the time complexity as it is similar to the Algorithm B in [11]. But we briefly list it for each step.

Step 1 takes $O(d)$. Step 2 takes $O(n)$. Step 3 takes $O(dn + (M_1 + \dots + M_d))$ for the courting sort part. The construction of the permutation matrix takes $O(d^2n)$ as each entry in the matrix takes $O(n)$ to complete (See the Lemma in [11]). Step 4 worst case time (with all possible iterations) takes $O(d(dn + M_1 + \dots + M_d))$.

Therefore the entire safety detection algorithm has time-complexity of $O(d^2n + d \sum_{1 \leq j \leq d} M_j)$.

We now find the space-complexity. We note it is more than that of the original safety state detection algorithm mainly due to the storage requirement for the permutation matrix. More specifically the space-complexity for the Extra Data Structure part in the algorithm is $O(dn + d^2n) = O(d^2n)$. Here the unit of the complexity is the storage for a single integer type. This is d -time worse than that of the original algorithm.

We now can apply the safety detection algorithm to Banker's algorithm. The most obvious way is to linearly check the safety for each process assuming it is

pre-allocated to the system. With this approach, the time complexity of the Banker's algorithm is $O(d^2n^2 + dn \sum_{1 \leq j \leq d} M_j)$.

Assuming the M_j are small and fixed, then the time-complexity for the Banker's algorithm is $O(d^2n^2)$ with the linear series of applications of our safety detection algorithm.

23.3 Conclusion and Further Consideration

We have obtained an improved algorithm over the original Banker's algorithm for deadlock avoidance for concurrent systems with reusable resource types. This new algorithm has the main advantage being n -times faster than the original Banker's algorithm, when the number of resource types is fixed and the maximum number of units for each resource is fixed; this is a reasonable assumption in practice.

There is further possibility on the improvements of the Banker's algorithm by a different application of our safety detection algorithm other than the linear verification for each process assuming it is pre-allocated. We will investigate this potentially new strategy. Currently, we expect that it can be incorporated in the safety detection algorithm, and it can achieve the overall time-complexity of the Banker's algorithm to be $O(d^2n \log(n))$, again assuming M_j are fixed constants.

References

1. Dijkstra EW (1965) Cooperating Sequential Processes. Technical report, Technological University, Eindhoven, The Netherlands, pp 43–112
2. Gold EM (1978) Deadlock prediction: easy and difficult cases. *SIAM J Comput* 7:320–336
3. Habermann AN (1969) Prevention of system deadlocks, *Commun ACM* 12.7:373–377, 385
4. Holt RC (1971) Comments on prevention of system deadlocks. *Commun ACM* 14(1): 179–196
5. Holt RC (1972) Some deadlock properties of computer systems. *Comput Surv* 4(3):179–196
6. Minura T (1980) Testing deadlock-freedom of computer systems. *J ACM* 27(2):270–280
7. Suguyama Y, Araki T, Okui J, Kasami T (1977) Complexity of the deadlock avoidance problem. *Trans Inst Electron Comm Eng Jpn* J60-D 4:251–258
8. Silberschatz A, Galvin PB, Gagne G (2002) *Operating system concepts*. Wiley, New York
9. Stallings W (1997) *Operating systems, internals and design principles*. Prentice-Hall, Upper Saddle River
10. Li Y, Cook R (2007) A new algorithm and asymptotical properties for deadlock detection problem for computer systems with reusable resource types. *Advances and innovations in systems, computing sciences and software engineering*. Springer, Heidelberg, pp 509–512
11. Li Y et al (2010) On Dijkstra's algorithm for deadlock detection. *Advanced techniques in computing sciences and software engineering*. Springer, Heidelberg, pp 385–387

Chapter 24

Courses Enrollment Pattern Analysis

Nur Fatihah Abdul Rahim, Shakirah Mohd Taib
and Saipunidzam Mahamad

Abstract Continuous research and works on the implementation of Apriori algorithm into educational environment proved that Association rule has evolved. This paper analyze enrollment pattern of final semester students from a university and discover interesting knowledge that may help the timetable committee in reducing clash conflict and for future decision making along the gradual process. The factor of clash conflict has been identified based on increasing number group of students whose did not follow the course structure guideline that has been designed for them. By producing prototype that adopts Apriori algorithm, it reveals the strong rules based on the extracted data. It is potential to apply into existing timetable preparation process to added new value in order to increase system performance while reduce number of class clashes. The rules or patterns produced by the prototype will then be compared with the existing course structure to evaluate the similarity and connection between courses.

N. F. A. Rahim (✉) · S. M. Taib · S. Mahamad
Faculty of Science and Information Technology, Department of Computer and Information
Sciences, Universiti Teknologi Petronas, Bandar Seri Iskandar, 31550
ronoh, Perak, Malaysia
e-mail: fathafazwan@gmail.com

S. M. Taib
e-mail: shakita@petronas.com.my

S. Mahamad
e-mail: saipunidzam_mahamad@petronas.com.my

24.1 Introduction

Scheduling is an essential business process where every organizations and learning institutions nowadays start to implement a systematic procedures or steps to overcome any clash or business interruption. This kind of interruption leads to low performance, increase number of complaints, data redundancy and unreliable output. In order to gain comprehensive satisfaction from both internal and external party in the organization, data mining framework has been practice into the process. The initial idea of Association Rule Discovery is to achieve cost optimization being implemented by market analyst by studying the behavior of their customer on the purchasing trend. This study adopts the same concept in a different domain, which is education and focuses on analyzing enrollment patterns in higher education institutions.

The scheduling process in any learning institution involves many stages or activities before the course timetable produced [1]. Usually before the arrangement starts, timetable's committee set a meeting to discuss the historical data from past enrollment. Despite analyzing the historical data by using spreadsheets and display the statistics by visual aids, with adoption of Association Rule technique; Apriori, can benefit the academic committee by predicting possible clashes for the coming semester.

The potential core problem of conflict has been identified by the respective institution which is the unusual enrollment made by some students instead of referring to the recommended modules set by the institution itself. For this study, enrollment data of final semester students is referred as a case study since there is high possibility of timetable problem for this group. Priority was given to this group pertaining to the completion of study may affect other group that follow the standard course structure designed by registrar department.

This paper presents the analysis on course enrollment at the university using Apriori algorithm by exploring the output of new knowledge or interesting rules. The analysis gathered the course enrollment pattern with the concept of Association Rule Discovery into timetable preparation process. The project is designed for the timetable committee during timetable preparation phase as an added value to discover previous enrollment pattern by students. Prior to complexity issue produce by large database, data training or testing data is created to be use during the testing and analysis stages

24.2 Concept and Theory

24.2.1 Timetabling Process

Timetabling or scheduling is process of setting an order and time for planned events. It is all about arranging a series of learning subject for the given time. According to [1] research journal, "course scheduling in colleges and universities is

an NP-complete operation” which means that the computational time required to find the solution increases exponentially with problem size. Clashing occur if the timetabling team place less priority to students as supported by [2] “... timetables are planned according to the characteristic of individual departments and institutes, with little attention to students’ interest and needs in career development”.

According to [3], the problem of timetabling is to assign various resources to the timetable that consistent with the following two fundamental constraints; neither students nor staff can be in more than one place at once and there must be enough space for the students in all venues. Thus, this paper discusses the significance of identified patterns in students’ enrollment data that offers helpful and constructive recommendations to the academic planners to enhance their timetabling process.

24.2.2 Apriori Algorithm

Due to excessive amount of data produced by any transactional system urged data mining researchers to discover pattern and find useful information or knowledge to improve business process in the related area. Apparently mining association rules are widely implemented in marketing, sales forecasting, inventory management and other related fields.

Apriori principle relies on the concept of support, s and confidence, c .

$$A \rightarrow B \quad (24.1)$$

For instance, when equation in (24.1) applied, s define as a fraction of transaction that contain both A and B , while confidence, c measures how often items in B appear in transactions that contain A . Apriori algorithm was pioneered in 1994 and became as a reference to solve association rules [4]. Based on its principle, if the itemset is frequent, then all of its subsets must also be frequent [5]. Frequent itemset is when support satisfies the minimum support set by user. First, the set of frequent 1-itemsets is found. This set is representing $L1$ and then $L1$ is used to generate candidate of $L2$, the set of frequent for 2-itemsets, which is used to create the next generation of $L3$, and so on, until it converged where no more frequent k -itemsets can be created. During the process, those infrequent itemsets are eliminated by applying pruning if not satisfy the *minsup*. Hence, k passes on the database where k is size of the largest frequent itemset. This will lead to the drawback of algorithm if size of k is extreme large.

24.2.3 Evolution of Association Rules Discovery

The ideal data mining starts with retail stores in analyzing customer behavior and focus on cross-selling strategy. Data mining is widely applied in business applications including market segmentation, customer profiling, fraud detection, evaluation of retail promotions, credit risk analysis for insurance policy, and in some military operations [6]. Now, data mining or as known as Knowledge Discovery

could be applied to others industries including education. Educational data mining is an interesting research area, which extracts useful, previously unknown patterns from educational database for better understanding, improved educational performance and assessment of the student learning process [7]. It is concerned with developing methods for exploring the unique types of data that come from educational environment which include students' results repository.

A study has been made by [8] regarding the implementation of association rule to identify students' failure patterns. The identified patterns are analyzed to offer a helpful and constructive recommendations to the academic planners in higher institutions of learning to enhance their decision making process. Intensive research on association rules improved the effectiveness of educational purposes. The domain knowledge in data mining is essential in association rule mining in order to find fast and effective association rule mining algorithm [9].

24.3 System Design

The project development is segmented into three phases:

24.3.1 Phase 1: Planning and Critical Review of Related Works

The problem has been identified and study of project background, which is on timetable preparation process applied to schools, colleges and universities done on this stage. The study of project background is conducted through interview session at and survey is conducted to final year students. Likewise, the critical review of existing data mining applications in the market is performed in literature review part.

24.3.2 Phase 2: Data Mining Processes and Analytical Technique

Enrollment data from previous semesters were collected. However, for the purpose of data integrity only several data that correlated to the project were selected and were cleaned through preprocessing before mining process. Algorithm was tested and the process of determine suitable parameters: *minsup* and *minconf* for the prototype is necessary to produce adequate rules.

24.3.3 Phase 3: Design and Pattern Analysis

The prototype was designed specifically to help the timetable committee to prepare a forecasting report based on student enrollment data in term of unusual

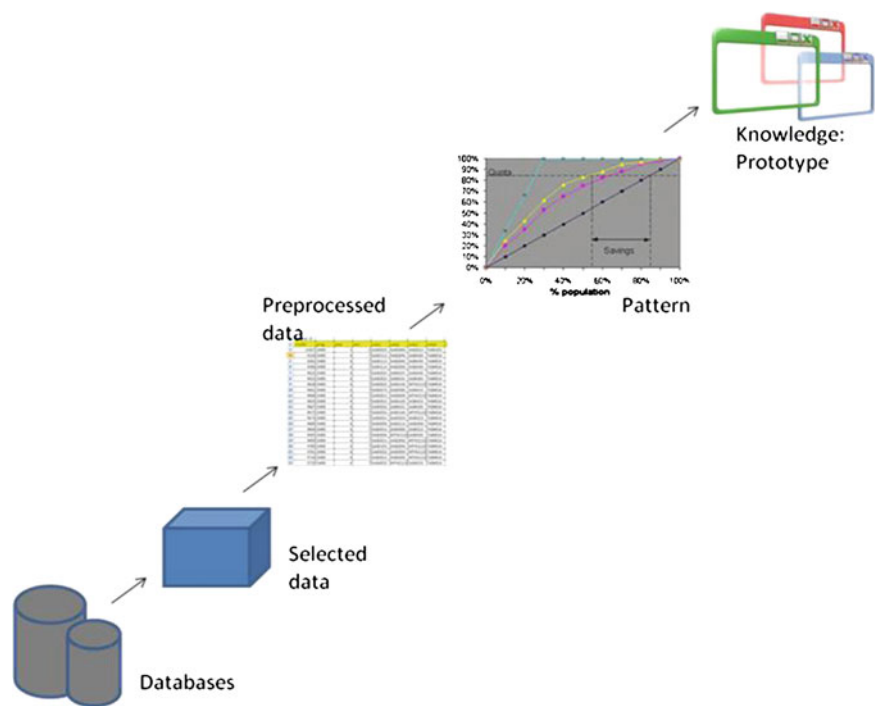


Fig. 24.1 Mining process

enrollment that may lead to conflict problem, pattern of specific group of students, and which subjects that the students will preferably enroll together based on the study of their past enrollment patterns. Figure 24.1 shows the overall project activities.

24.4 Result and Discussion

Data gathering process aims to collect as much information about the entire project. Therefore two methods are involved in collecting data, which are survey and interview at Universiti Tekonologi Petronas (UTP) as a case study. A total of 94 respondents from different programmes of final year students at the university participated in these survey and interview.

Figure 24.2 shows the methods used by the students on selecting the course for the course enrollment. There are 56.4 % of the students were choose the ideal method by following the course structure in the university guideline. About 29.8 % of them made the decision based on the classmates’ preference. Mean-while, another 13.8 % of the students are following the course structure that

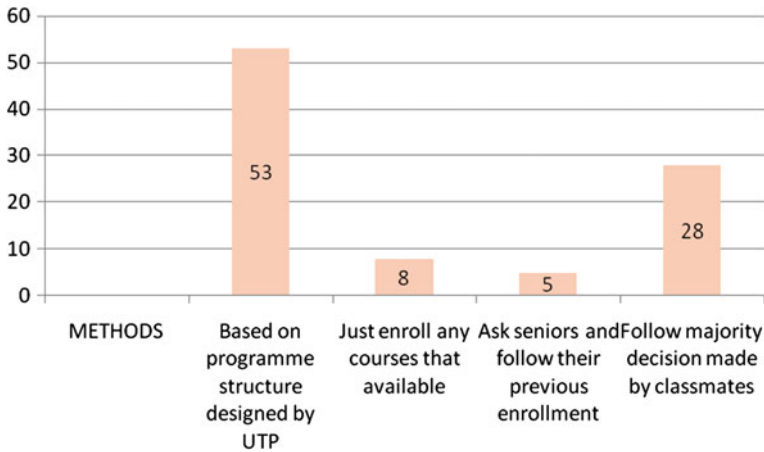


Fig. 24.2 Methods of choosing subject in next enrollment

recommends them to register based course availability or following a suggestion from seniors. Thus, such situation will derived a problem likes timetable clashing.

The prototype of the system comprises of three main functions as shown in Fig. 24.3.

- Read file and extract data
Students' enrollment data were in.txt format
- Run Apriori analysis
The program starts when user clicks on button run. It will display k-itemset and any subjects that meet the parameters set by program as rules.
- Compare rules generated with generic course structure
Rules generated by the system then being compared with the generic course structure designed by the university.

In the development process, the parameters for this system; minimum confidence, minsup and minimum confidence, minconf were set as 6 and 0.9 respectively. An association rule is interesting if it satisfies user-specified minsup and minconf. There are no specific guidelines in setting both parameters. According to the interestingness of s and c stated by [10] if minsup is low, we may extract too many spurious patterns involving items with substantially different support level. It means that too many items may leads to overfitting data. Likewise, if minsup is high, we may miss many interesting patterns occurring at low levels of support because the frequent itemset only exist when support more or equal to minsup. Usually minsup is set by the user lower than the minconf itself.

Figure 24.4 shows the general pseudocode to implement Apriori algorithm into the system.

While Fig. 24.5 shows the simple steps that were developed based on Apriori algorithm.

Fig. 24.3 System activity diagram

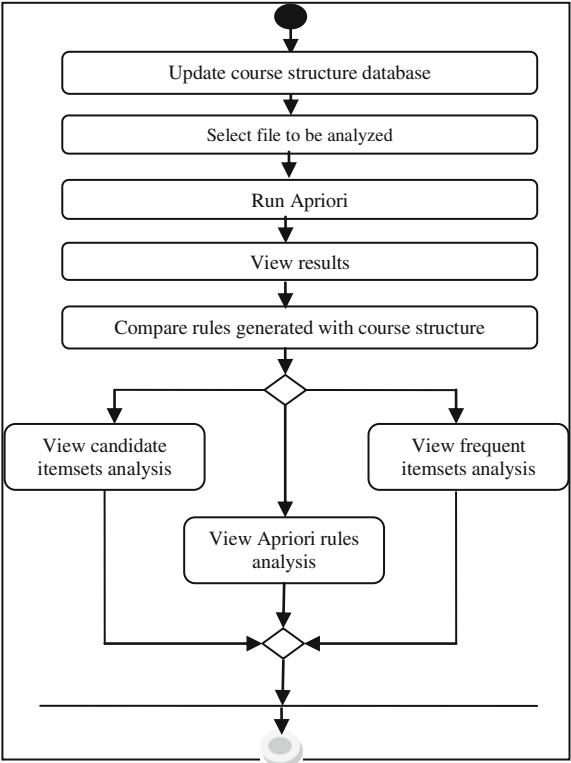


Fig. 24.4 Apriori algorithm

```
L1 = {large 1-itemsets}
For (k=2; Lk-1 )≠ ∅; k++) do begin
    Ck =apriori-gen(Lk-1); // new candidates
    Forall transactions t ∈ D do begin
        C't = subset (Ck,t) //candidates
        contained in t
        Forall candidates c ∈ Ct do
            c.count++
        end
        Lk = {c ∈ Ct | c.count ≥ minsup}
    end
end
return UkLk
```

The interestingness of association rule can be described as shown in Table 24.1 where we manipulate minsup and minconf to find the best parameters to be set in the system prototype.

Table 24.2 shows data of students’ enrollments from final semester students in Business Information Systems (BIS) for data testing purposes.

Figure 24.6 shows the output of association rules with 4 rules generated from the system. All rules are 100 % strong in confidence, which mean the relationship

Fig. 24.5 Steps of implementation

- Step 1: Both parameters minsup and minconf were set and accept the student enrollment course as the input data set.
- Step 2: Determine the support count for all the items as s (courses in enrollment data as items).
- Step 3: Select the frequent items; item with $s \geq \text{minsup}$
- Step 4: The set candidate k- item is generated by 1- extension of the large (k-1) itemsets generated in step3
- Step 5: Support for the candidate k-itemsets are generated by a pass over the database.
- Step 6; Itemset that do not have minsup are discarded and the remaining itemsets are called large k-itemsets.
- Step 7 : The process is repeated until no more large item.
- Step 8: The interesting rules are determined based on the minimum confidence.

Table 24.1 Relationship between minconf and minsup and number of rules generated

| Sample of data (program) | minsup minconf | 6 0.8 | 6 0.9 | 7 0.8 | 7 0.9 |
|--------------------------|-------------------|----------|----------|----------|----------|
| ME | #rules | 8 | 5 | 8 | 5 |
| CHE | #rules | 2 | 2 | 2 | 2 |
| BIS | #rules | 4 | 4 | 2 | 2 |
| ICT | #rules | 31 | 22 | 9 | 3 |
| CV | #rules | 12 | 11 | 1 | 13 |
| EE | #rules | 43 | 21 | 43 | 21 |
| PE | #rules | 68 | 68 | 68 | 68 |

between items is equal, or above *minsup* and *minconf*. For instance, the rule of SAB4343 \rightarrow TAB 4014 indicates that subject of SAB4343 (Data Mining and Knowledge Discovery) will be enrolled together with TAB 4014 (Final Year Project II) by most of BIS students in their final semester. Apparently, course SAB4343 is not in their module structure at that semester. This unusual enrollment will be an input delivered to timetable planner for further analysis.

Based on the interview conducted, the reason why the major subjects are not included in final semester of BIS study is to allow students to focus on their final year project. But somehow, rules generated by Apriori may support the relevancy of course structure designed by university as shown in Table 24 3. Both courses are related in the existing course structure by the rule of MPW2133 \rightarrow TAB 4014.

Table 24.2 Sample dataset of his final semester students

| STUDENT | SUBJECT1 | SUBJECT2 | SUBJECT3 | SUBJECT4 |
|---------|----------|----------|----------|----------|
| S1 | GAB3113 | HAB2043 | SAB4343 | TAB 4014 |
| S2 | GAB3113 | HAB2043 | SAB4343, | TAB 4014 |
| S3 | GAB3113 | HAB2043 | SAB4343 | TAB 4014 |
| S4 | GAB3023 | SAB4333 | SAB4343 | TAB 4014 |
| S5 | GAB3023 | SAB4333 | SAB4343 | TAB 4014 |
| S6 | GAB3023 | GAB3143 | MPW2133 | TAB 4014 |
| S7 | GAB3023 | GAB3143 | MPW2133 | TAB 4014 |
| S8 | GAB3013 | GAB3023 | SAB4223 | TAB 4014 |
| Sn | .. | .. | .. | .. |

Fig. 24.6 Generated association rules

```

Strong rules discovered[Association Rules]
HAB2043=>TAB4014          100%
SAB4343=>TAB4014          100%
GAB3023=>TAB4014          100%
MPW2133=>TAB4014          100%
Result:4rules generated

```

Table 24.3 Generic courses structure for his final semester

| Subject | Course code |
|------------------------------------|---|
| Final year project II | TAB 4014 |
| Small business Entrepreneurship | GAB3093 |
| Corporate ethics | GAB3013 |
| Malaysian studies | MPW2133 |
| ^a Any minor electives | GAB3023, GAB3073, GAB3083, GAB 3123, GAB 3133, GAB3143, GAB3103, GAB3063, GAB3053, GAB2043, GAB3113, GAB3153 |

^a It can be more than one at one enrollment

24.5 Conclusion

This project highlights the importance of data mining technique to discover hidden pattern on students' enrollment. The new knowledge that is an enrollment pattern was discovered by the prototype is recommended to be included during the timetable preparation process. The result shows the association rules discovery on the enrollment data would benefit the planners as it reveals the types of students' group and their behavior towards course enrollment. They can predict which courses that are commonly enrolled together and compare with the existing course structure set by the university. For future system development, this prototype should be extended with some additional functions such as graph aids and automated pattern interpretation.

References

1. Yu J, Hu Z (2008) Using formal methods to design a class scheduling system. IEEE Xplore 56–59. doi:[10.1109/CSSE.2008.804](https://doi.org/10.1109/CSSE.2008.804)
2. Yu Y et al (2008) On the application of data mining technique and genetic algorithm to an automatic course scheduling system. IEEE Xplore 400–404
3. Burke EK, Elliman DG, Weare RF (1995) The automation of the timetabling process in higher education. J Educ Technol Syst 23(4):257–266
4. Ahmed Ayad M (2000) A new algorithm for incremental mining of constrained association rules. Alexandria University
5. Tan K et al (2006) Introduction to data mining, Pearson International Edition, p 33
6. Witten IH, Frank E (2000) Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufman, California
7. Dogan B, Camurcu AY (2008) Association rule mining from an intelligent tutor. J Educ Technol Syst 36(4):433–447
8. Oladipupo OO, Oyelade OJ (2010) Knowledge discovery from students. Repository: association rule mining approach. Int J Comput Sci Secur 4(2):199–204
9. Motwani et al (2009) Use of domain knowledge for fast mining association rules. In: Proceedings of the international multiconference of engineers and computer scientists (IMECS) 2009, vol I, 18–20 March, Hong Kong
10. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. VLDB-94

Chapter 25

Integration of Safety and Smartness Using Cloud Services: An Insight to Future

Neha Tekriwal, Madhumita and P. Venkata Krishna

Abstract The problem of terrorist attacks has become a critical issue in today's world. To provide maximum security to the people, fastest detection and evacuation means are required. We aim at integrating intelligence and smartness using cloud based services to enable public safety. A smart city has a ubiquity of technology and inter-connected with government and private sub-systems to provide early response and security to the people. In this paper, we discuss an approach to detect remote explosives using wireless sensors and generate different levels of alert to notify the public. Using GPS tracking technology, the shortest routes are shown to users. Thus by integrating existing with upcoming technology, we provide a solution to make smarter and timelier decisions.

25.1 Introduction

In a country like India where there is a high rate of terrorist and criminal activities, security and safety of the general public has become one of the most important concerns of the government. According to data from the South Asia Terrorism Portal database [1] there were 1902 deaths in India in 2010 due to terrorist and insurgent activities. There has been a significant rise in the use of Improvised

N. Tekriwal · Madhumita · P. V. Krishna (✉)
School of Computing Science and Engineering, VIT University, Vellore,
Tamil Nadu, India
e-mail: parimalavk@gmail.com

N. Tekriwal
e-mail: ntekriwal@gmail.com

Madhumita
e-mail: madhumita154@gmail.com

Explosive Devices (IED). There has been 1062 effective and successful IED attacks in 2010 in Afghanistan. Around 273 monthly IED attacks take place outside Iraq and Afghanistan [2]. The main issue that lies here is how to safeguard the lives of innocent people who are victim of such atrocities. Though the smart city concept is new, there is a lot that can be done in this field.

In a smart city, life of the people is made easy with the ubiquity of technology. There will be devices and systems such as wireless sensor networks, radio frequency identification (RFID), wireless and mobile communications and internet-enabled devices. A wireless sensor network is a wireless network consisting of devices using sensors deployed at different locations to monitor the environment. These sensors contain a network topology where each node is connected to the main node called the base station which act as a gateway between the end users and the sensor nodes. These sensor nodes are also called motes which communicate with the other connected nodes to process and collect the information.

These technologies are being used more and more for accessing and processing information as well as communication. U-city is an urban development in the field of smart city [3]. It is designed to focus on the users, and provide personalized service and user interface for all the citizens. The U-city aims at proposing an electronic lifestyle, such that the interactions take place through ubiquitous computing. This city is under development and currently operational. Smart city aims at networking the entire city such that each person and device is interconnected to each other. High speed unlimited internet connection is available to all. In the system individuals are connected to a cloud network. Internet based cloud computing for data collection and monitoring is done.

Cloud Computing refers to the integration of hardware and software services over the internet to provide multiple access in terms of applications and resources. These cloud networks are further connected to other clouds to form a network of cloud. This cloud network becomes the platform for data monitoring and surveillance. Figure 25.1 defines the cloud architecture that will be used in the proposed architecture. The data from the sensors is streamed to the analytic software running on the cloud [4]. The system is beneficial in the terms of real-time processing and accessing of data and information. Cloud computing provides a scalable processing power as there are multiple networks being used and reduces the overall cost factor in the deployment of network. Thus, this system enhances the applicability and pertinence of the data being monitored and improves the decision making.

The future of the smart city in India revamps the existing security and safety assessments. The system aims at securing public places by monitoring any threats like explosives or arms (Fig. 25.2).

The remainder of this paper is structured as follows. Section 25.2 tells us about the related work, Sect. 25.3 presents the solving approach to provide public safety in the smart city. We have subdivided this section into two parts. Section 25.3.1 gives an overview of the architecture proposed and Sect. 25.3.2 provides the strategy proposed. Section 25.4 deals with the challenging issues and Sect. 25.5 provides some countermeasures in case of GPS failure. Section 25.6 presents the conclusions and results.

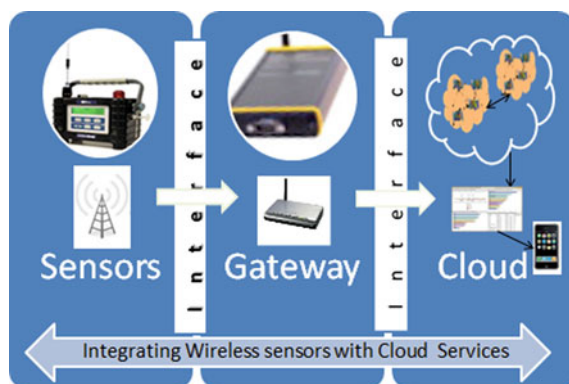


Fig. 25.1 Integration of sensors with cloud services

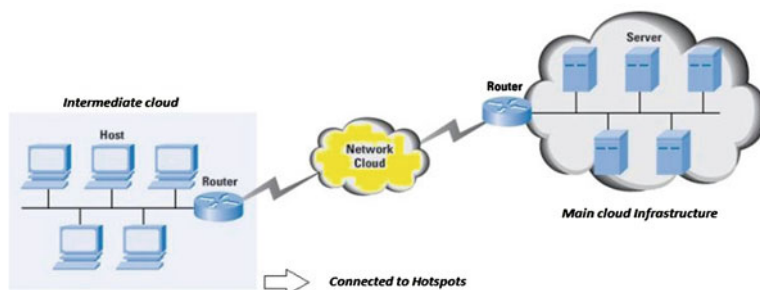


Fig. 25.2 Cloud network

25.2 Related Works

We present here some work that has been done in this field in the literature. One of the ongoing projects to detect explosive threats is Bomb factory detection by Network of Advanced Sensors (BONAS) [5]. In this project, the explosive sensors are deployed in sensitive locations and concealed easily, to detect the precursors used in IED productions.

Researchers at the Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE in Wachtberg have built a prototype security system Hazardous Material Localization and Person Tracking (HAMLeT) to alert security personnel to suspicious individuals by tracking down hidden explosives [6].

There is another ongoing research by a group from Kingston University's Digital Imaging Research Center (DIRC), to develop a system to automatically analyze multi-camera networks and footage before and after any trigger incident [7].

A ‘Tag and Track’ surveillance system was developed by security company Ipsotek [8]. This system can allow security to tag suspicious individuals, and track them across multiple cameras.

Another method that has been proposed is wireless sensing of bombs using neural networks, and positioning through Kalman Filter and GPS tracking [9].

iWEDS (Intelligent explosive detection and terrorist tracking system using wireless sensor network) proposes a method to track explosives through gas and chemical sensors [10]. The two outputs are combined and the explosive name is retrieved from the database. The trigger control unit informs the neighbors to track the target.

The past works enable us to know the presence of explosives, and in some cases, their location or type. There is no method proposed for evacuation of the people in the area without creating panic. There is no mechanism to alert the people. Also, there is no method proposed lest the GPS tracker should fail to work.

In this paper, we propose an explosive detection and evacuation plan to alert and secure the people of any threat. We also propose alternative methods to employ in case of GPS failure.

25.3 Solving Approach

The explosive sensors are deployed in the entire city. The nanosensors provide a full-fledged platform for the trace detection of explosives [11]. It allows mass deployment of these tiny sensors in the public hotspots at inexpensive rate. In a multi-integrated environment the application of chemo-sensors and nano sensors for security operation and diligence can be accomplished.

Wireless sensor connectivity is provided to everyone in the smart city at all the public hotspots and the Global Positioning System (GPS) [12] tracker is enabled to pull the instantaneous locations of the users connected to it. High security evacuation points are constructed. The positions of evacuation points in the area are overlaid on maps using Google Map API.

We have designed a rough solution for providing safety in a smart city architecture and proposed the evacuation plan in case of any attack in the next sub-section.

25.3.1 Architecture

In our architecture (see Fig. 25.3), we have used the state-of-the-art sensors, CCTVs, facial recognition software, Global Positioning Satellites, and interactive screens which are described below.

State-of-the-art explosive sensors. These sensors monitor the environment to detect traces of chemicals in air. These sensors are capable of detecting hidden explosives and can be very effective in providing a warning in advance [13].

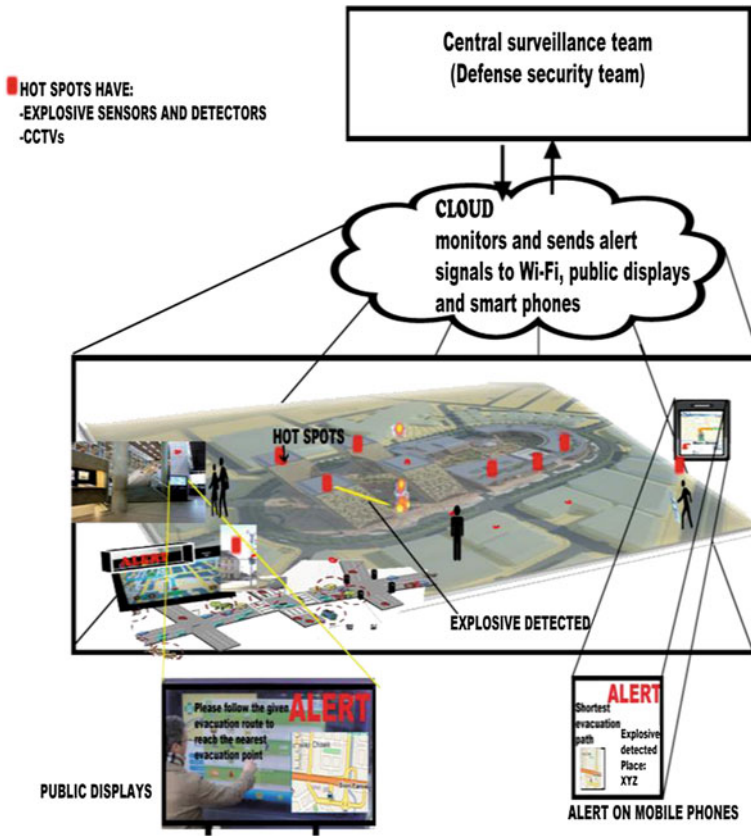


Fig. 25.3 Design of a Smart City to implement safety

CCTVs. The CCTV cameras are installed at every hotspot. They continuously process the images of the entire area, in order to keep a track of events occurring there.

Facial recognition software. This software processes the scanned images obtained by CCTVs to identify the individuals uniquely by comparing it to the existing database [14].

Global Positioning Satellites. This is satellite based navigation system that gives the location-based information. This location positioning technique is used to locate the users in an area, and to generate the shortest route to the nearest evacuation point.

Interactive screens. Interactive public displays are large screen digital displays that can interact with users and return the data collected over the network. These screens are used to display the alert status and the evacuation routes to the user [15].

In this design, a central surveillance system acts as a backbone for monitoring the entire city through cloud services. The sensors send the data gathered to the 'central nervous system' dynamically. The design shows the alerts generated on detecting any explosive which is then displayed on public displays and internet-enabled devices along with the evacuation routes. This design is multi purposed and can be used in other services like traffic monitoring, pollution control, user interfaces etc. depending upon the sensors used.

25.3.2 Remote Explosive Detection and Evacuation Plan

The hotspots are continuously monitored and tracked. The data collected at different hotspots are shared using cloud network to provide central assessment and monitoring of the city. The tracking process follows the following Remote Explosive Detection and Evacuation Plan (REDEP) (see Fig. 25.4):

1. All the nodes are initialized and synchronized.
2. The CCTVs return the images of everyone in the area.
3. Facial recognition software uniquely identifies the individuals.
4. Individuals with a history of criminal records cause a variable 'criminal' to be set as true.
5. The state-of-art sensors detect the chemical traces of explosive vapors in air, and sets a variable 'evacuation' to be true.
6. The output from 4 and 5 are used to generate alert levels:
 - (a) If 'evacuation' is true, high level alert is generated and evacuation plan is followed.
 - (b) If 'criminal' is true, medium level alert is generated, and surveillance team gets into action to verify the alert.
 - (c) If complaints about local crimes have been lodged, then a low level alert is generated.

For the evacuation plan, an alert message is broadcast to the immediate cloud. The message is then forwarded to the main cloud which sends alert messages and evacuation routes to its users in that area with the help of GPS tracking. Alert messages are also displayed on public displays and interactive screens installed in area. Appropriate evacuation routes are displayed on them. The users are notified through e-mail, SMS, HTTP POST and public addressing system.

GPS locator pulls out the instantaneous location of the people. Shortest route to the nearest evacuation point in the area is generated from that location (see Fig. 25.5). The evacuation routes are generated using tools such as EvacSpace [16] used in emergency situations like this. Users are requested to follow the given route to avoid panic. After people reach the evacuation point, the person's profile is saved and listed in the safe zone. Along with that, a message is broadcast on the user profile informing that he is safe. Relatives and acquaintances can easily find

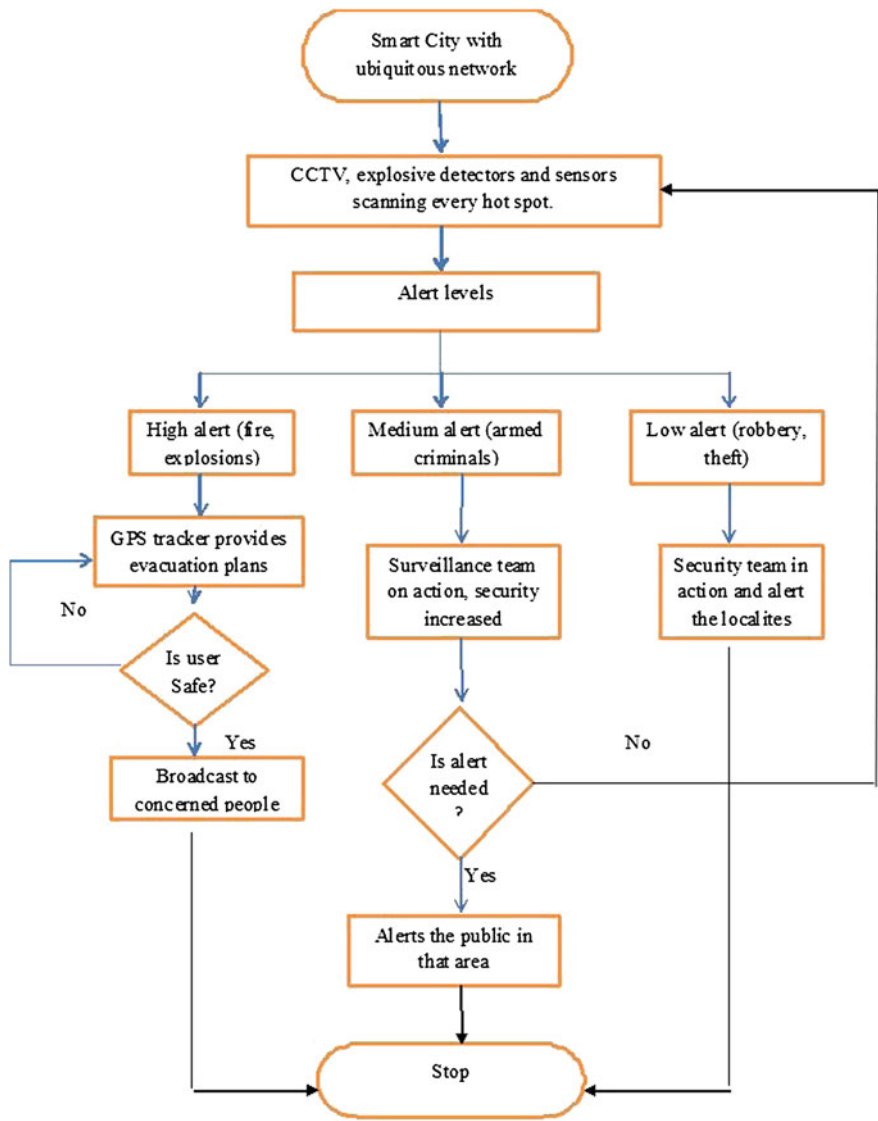


Fig. 25.4 REDEP strategy: safety implementation in Smart City

out if the person is in a safe zone or not through the interactive API without causing any network congestion.

If a medium level alert is generated, the surveillance team is brought into verify the alert. If the team confirms the alert, the evacuation plan for the high level alert is followed.

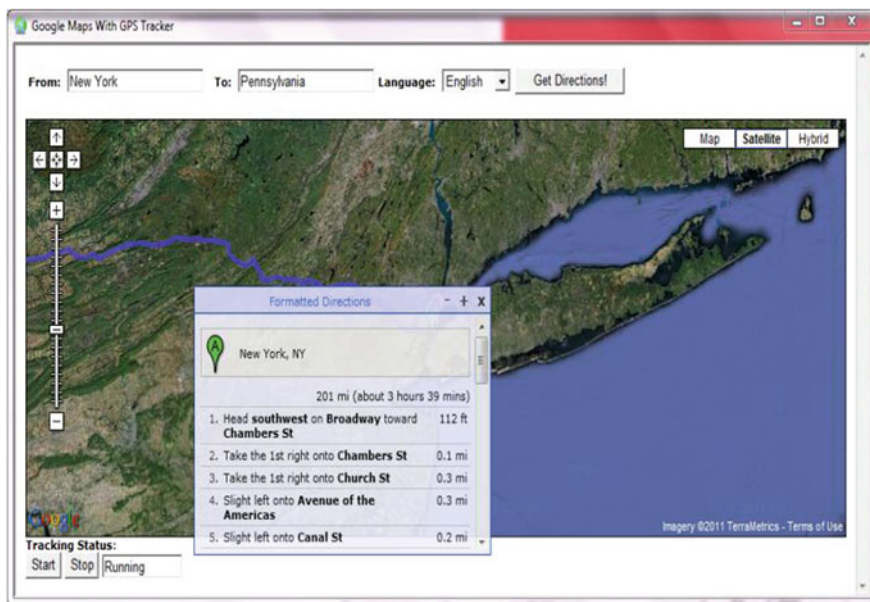


Fig. 25.5 Route generated through Google maps with GPS tracker

If a low level alert is generated, the local security team deals with it to alert the localities.

25.4 Challenging Issues

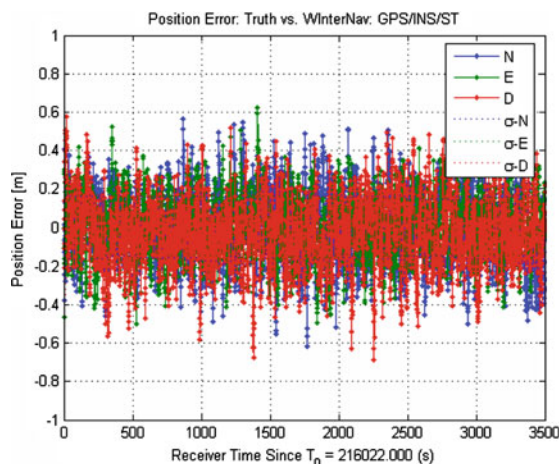
The deployment of such a large infrastructure faces many challenges which includes the financial sustainability as a lot of funding is required to cover the expenses.

Another challenging issues faced in the implementation of the safety architecture apart from the power and cost factors are false positives, location errors, and failure of GPS. These are described below.

25.4.1 False Positives

Verification of false positives is a challenging issue. This issue can be solved by the use of explosive detectors with minimum false positives. Example of such detector is the Quantum Sniffer QS-BTS Benchtop explosive detector, which has less than 1 % false positives [17].

Fig. 25.6 GPS/INS/ST position estimation error (simulated GPS, INS, and star tracker data were generated using NAVSYS GPS toolbox for MATLAB, N—North, E—East, D—Down)



25.4.2 Location Errors

Location errors are present in Google Map API due to the environmental effects on satellite signals, causing signal interference [18]. These errors need to be resolved. According to the simulation carried by NAVSYS Corporation [19] for GPS positioning (see Fig. 25.5), there are position errors of 0.7 m. Conventional methods such as Kalman filter based methods [20, 21] are applied to detect the error. Another new proposed method estimates camera parameters by minimizing an energy function that is defined by using the re-projection error and considers GPS positioning accuracy [22].

25.4.3 Failure of GPS

Another challenge is overcoming the threats posed on the disruption of GPS services. The performance of GPS may be hampered due to some of the potential threats like: solar disturbances and LOS between GPS satellites and receivers. The terrorists may try to jam the signals using radio interference. Spoofing of signals is yet another concern (Fig. 25.6).

25.5 Countermeasures for GPS Failure

There are some countermeasures to overcome the flaw of GPS failure. In a report from TISN [23], one of the effective method introduced is “Controlled Reception Pattern Antenna” or CRPA which determines the direction of a jamming source and modifies its antenna reception pattern to ignore signals from that direction. Another method is to employ a jamming signal to thermal noise ‘powermeter’

which provides a source of integrity check on the reliability of the receiver. It measures the total amount of power received by the antenna, and knowing the amount of power expected from thermal noise, the receiver can measure the amount of received jamming power.

We provide an alternate solution to the condition when GPS fails or the positioning is inaccurate. Mobile navigation techniques by using other sensors like WLAN, Bluetooth, GSM, etc. based on Wi-Fi signals or ultrasound can be utilized to determine the user position in case of GPS failure. The integration of different services provides more positioning accuracy and low rate of failure. In [24] the authors have described about various positioning techniques in case of GPS failure and have provided a solution of integrating the Wi-Fi positioning with GPS to provide more accuracy.

25.6 Conclusions

Using the mechanism described in this paper, it is possible to safeguard the people against attacks and criminal activities. Technology is used to promote safety of individuals. Thus smart cities can become the cornerstone of safety of individuals by using the urban wireless infrastructures and integrating it with cloud services. The next step we intend to pursue is to find solutions to potential problems like privacy and false positives. With the aid of large-scale automation and high-quality technology permeating every sphere of life, a safe and sound city can be built.

References

1. South Asia Terrorism Portal Database. <http://www.satp.org/satporgtp/countries/india/index.html>
2. <http://www.wired.com/dangerroom/2010/09/4-years-later-pentagon-lets-allies-onto-anti-bomb-website/#more-31707>
3. Ojala T (2011) Experiences inside the Ubiquitous Oulu Smart City. *IEEE Comput Soc* 44(6):48–55
4. Lee K, Murray D, Hughes D, Joosen W (2010) Extending sensor networks into the cloud using Amazon Web Services. In: *IEEE international conference on networked embedded systems for enterprise applications (NESEA)*
5. Bonas Project. <http://cordis.europa.eu>, <https://sites.google.com/a/tekever.com/bonas/home>
6. http://ftp.cordis.europa.eu/pub/fp7/security/docs/hamlet_en.pdf
7. The Engineer. <http://www.theengineer.co.uk/sectors/military-and-defence/news/cctv-project-aims-to-make-it-easier-to-track-down-criminals/1009710.article>
8. The Engineer. <http://www.theengineer.co.uk/sectors/military-and-defence/news/surveillance-system-tags-and-tracks-suspicious-individuals/1008855.article>
9. Rosy SS, Grace SS (2009) Bomb detection using wireless sensor and neural network. *IJCSI* 2
10. Hariharan B, Sasidharan A (2011) iWEDS—an intelligent explosive detection and terrorist tracking system using wireless sensor network. *IJCSI* 8(4):245–250

11. Senesac L, Thundat TG (2008) Nanosensors for trace explosive detection. *Materials Today* 11:28–36
12. GPS. <http://www.gps.gov>
13. Remote bomb detection sensors. <http://www.homelandsecuritynewswire.com/remote-bomb-detection-sensors>
14. Yang M (2011) Face detection. In: Li SZ (ed) *Encyclopedia of biometrics*, 2nd edn. Springer
15. Kuikkaniemi K, Jacucci G, Turpeinen M, Hoggan E (2012) From space to stage: how interactive screens will change urban life. In: *Proceedings of the designing interactive systems conference*, NY, pp 458–467
16. Pelman K, Robinson A (2011) An interactive mapping application for rapid evacuation planning. *ISCRAM 8th international conference on information systems for crisis response and management*, pp 257–259
17. Quantum Sniffer QS-BTS Benchtop Detectors. <http://www.oconnors.earnes.com>
18. GPS/Map Position Coordinate Issues: GPS Position Accuracy. <http://newyorksearchandrescue.org>
19. Brown A (NAVSYS Corporation, Colorado Springs, Colorado), Integrated Gps/Ins/Star tracker space navigation system using a software defined radio
20. Kong F, Dai G, Cai L, The composed correcting Kalman filtering method for integrated SINS/GPS navigation system. In: *2010 IEEE international conference on intelligent computing and intelligent systems (ICIS)*
21. Malleswari BL, MuraliKrishna IV, Lalkishore K, Seetha M, Nagaratna, Hegde P (2009) The role of Kalman filter in the modelling of GPS errors. *J Theor Appl Inf Technol* 5:15–24
22. Kume H, Taketomi T, Sato T, Yokoya N (2010) Extrinsic camera parameter estimation using video images and GPS considering GPS positioning accuracy. In: *ICPR*, pp 3923–3926, 2010 20th international conference on pattern recognition
23. GPS Vulnerability—Information for CIOs. Australian Global Navigation Satellite Systems Coordination Committee (AGCC), TISN
24. Binghao L, Dempster AG, Rizos C (2010) Positioning in environments where standard GPS fails. *TS 2C—low cost GNSS and new positioning techniques*, FIG congress 2010 facing the challenges—building the capacity Sydney, Australia, 11–16 April 2010

Chapter 26

A Versioning Subsystem of Metamodeling System

Rünno Sgirka

Abstract A metamodeling system is a system for creating modeling systems. In this paper, we propose a versioning subsystem for a web-based and database-based metamodeling system as well as for the modeling systems which are created by this metamodeling system. We describe briefly the metamodeling system, then present a design of the proposed versioning subsystem, and finally discuss whether the proposed design fits our needs for version management in our metamodeling system.

26.1 Introduction

We have proposed a web-based and database-based metamodeling system [1] with a query subsystem [2]. We call it WebMeta in this paper. WebMeta uses an object-relational database system (ORDBMS) as its *enabling technology* and has a web-based user interface. It is possible to use this system to create web-based and database-based modeling systems.

Metamodeling systems are important in the context of model driven development because they facilitate the creation of modeling systems that support the use of domain specific languages and generation of code based on these languages. The use of such languages and modeling systems improves communication between developers and users of the developed systems, increases productivity of developers, and helps developers to improve the quality of applications [3].

R. Sgirka (✉)
Department of Informatics, Tallinn University of Technology,
Raja 15, 12618 Tallinn, Estonia
e-mail: runno.sgirka@gmail.com

Collaboration between developers also often requires version management, whether for program code or the models the code is based on. In this paper, we focus on version management of the metamodels, as well as the artifacts created based on the metamodels. We use the more general concept *artifact* to denote the different outputs a metamodeling system can produce—models, patterns, plans etc.

The *first goal* of the paper is to propose a design for the version subsystem of our metamodeling system WebMeta. We propose a design for both the metamodeling system and for the modeling systems created by the metamodeling system. The *second goal* of the paper is to analyze whether the proposed design satisfies our needs for version management.

The rest of the paper is organized as follows. In Sect. 26.2, we describe briefly the design and working principles of our metamodeling system WebMeta. In Sect. 26.3, we propose the design of the versioning subsystem for the metamodeling system and the modeling systems created by the metamodeling system. In Sect. 26.4, we briefly discuss whether the proposed design meets our needs for version management. Finally, we conclude and describe further work with the current topic.

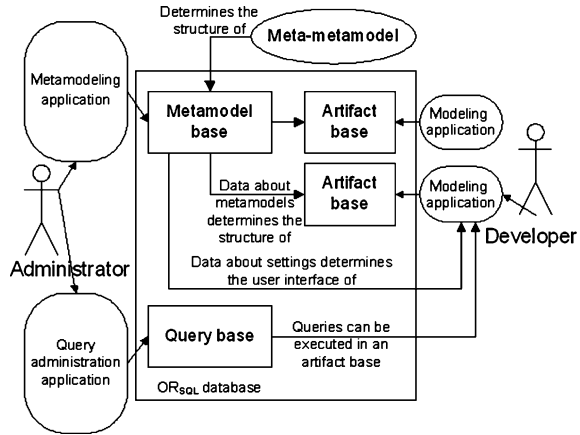
26.2 A Web-Based and Database-Based Metamodeling System

In this section, we describe briefly our web-based and database-based metamodeling system, WebMeta. Both the metamodeling system and the modeling systems that are created by using the metamodeling system use an ORDBMS as their enabling technology.

Figure 26.1 presents the general architecture of the system [1, 2]. The database allows us to integrate different parts of the system. At the logical level the database consists of exactly one *metamodel base*, exactly one *query base*, and zero or more *artifact bases*. We implement the metamodel base and the query base in a database as a single *SQL schema*, which is “a persistent, named collection of descriptors” [4]. In addition, each artifact base is implemented as a separate SQL schema. We have to use different schemas in order to prevent possible name conflicts.

The metamodeling system allows *administrators* to create web-based modeling systems by using a web-based user interface. *Developers (end-users)* use the modeling systems in order to manage (create, read, update, and delete) artifacts. The administrators and developers do not have to use Java Applets or to install additional plug-ins to their computer in order to use the system. Specifications of modeling systems (*metamodels*) as well as *artifacts*, which have been created by using these systems, are recorded in one database. The metamodeling and modeling systems provide *form-based* user interface. The system does not allow developers to create diagrams.

Fig. 26.1 Architecture of the WebMeta system



Each modeling system, which is created by using our metamodeling system, allows developers to create artifacts by using exactly one *software language*. The administrators of each new modeling system have to specify the abstract syntax of its underlying language in terms of a metamodel. In addition, an administrator has to register settings of the user interface and user identification of the modeling system as well as manage completeness and consistency (C&C) queries. The system records the queries in the query base. Developers can execute these queries based on artifacts.

Meta Object Facility (MOF) Specification [5] describes four-layer metadata architecture. We can characterize the proposed system in terms of this architecture (see Fig. 26.2). A meta-metamodel specifies the abstract syntax of a language that one can use to specify new metamodels. In our system the meta-metamodel layer is implemented as a set of base tables (tables) that together form the metamodel base. All these tables are in exactly one SQL schema.

Our metamodeling system allows administrators to use exactly one meta-metamodeling language to create metamodels. We have tried to keep the language as simple as possible. In our system administrators have to specify metamodels in terms of objects and sub-objects. Therefore, the metamodel base contains tables *Object* and *Sub_object* among others.

Our system records metamodels in the tables of the metamodel base. In addition, each metamodel that is recorded in the metamodel base has exactly one corresponding artifact base, which is implemented as a separate SQL schema. It helps us to prevent name conflicts between the elements of different metamodels. If an administrator defines a metamodel m , then the system creates exactly one corresponding schema s in the database. If an administrator defines a new object that belongs to m , then the system creates corresponding base table (table) t in s . If an administrator defines a new sub-object so of o , then the system creates corresponding column c in t . The type of so determines the type of c . If the sub-object is used to specify a relationship between o and o' (objects which both belong to m), then the system also creates a foreign key constraint to c .

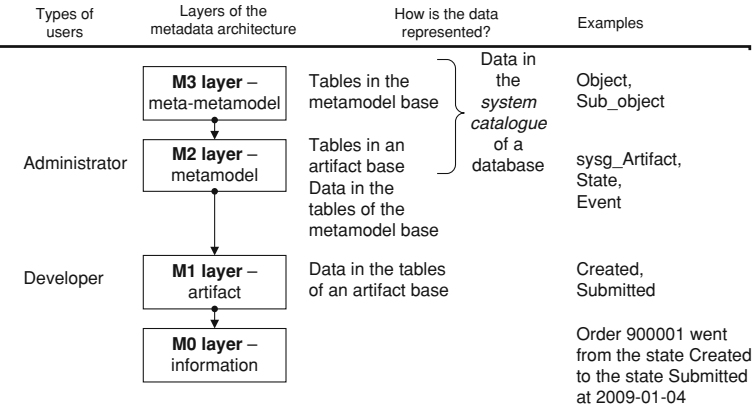


Fig. 26.2 Layers of the WebMeta system

In case of each metamodel in the metamodel base, the system creates automatically object *sysg_artifact* with sub-object *name*. Prefix “sysg_” denotes that this object is system generated. This object is necessary because it allows developers to create different artifacts, which are all recorded in the same artifact base and have the same metamodel. The system also automatically ensures that each main object (except *sysg_artifact*) o_i or object o_i that represents a relationship has sub-object *sysg_artifact_id* that specifies a relationship between o_i and *sysg_artifact*, ensuring that the tables in each artifact base contain identifiers of artifacts. It simplifies the creation of C&C queries about a single artifact—the queries must perform the restriction operation based on column *sysg_artifact_id*. It also simplifies the creation of some integrity constraints.

Each metamodel m has a set of associated well-formedness rules (constraints) [6] that give information about the semantics of m . Each well-formed artifact, which is created according to m , must satisfy all these constraints. In case of some constraints, administrators can decide whether these constraints must be *invariants*. In this case the system ensures that all the artifacts which are created according to m and stored in the database of the modeling system conform always to the constraints. The metamodeling system generates data definition language statements to implement this kind of constraints as integrity constraints in the database. In addition, there could be constraints that an artifact can initially violate. Administrators can create C&C queries that allow developers at any time to check artifacts in terms of these constraints. It makes possible gradual improvement of artifacts.

If an administrator changes the specification of a modeling system (metamodel), then he/she does not have to upload modified files to the server. If a developer requests a particular page, then the system determines the structure and behavior of the page based on data in the metamodel base. The content that is presented on this page comes from an artifact base. It is similar to the extreme extending (X^2) approach [7] according to which “major parts of the presentation layer (GUI) reside

within the DB Server”. It means that a DBMS generates dynamically HTML pages used for user interaction. In our system, PHP engine generates the pages. However, the specification of the structure and behavior of the pages is recorded in the database.

During the creation of each modeling system an administrator has to specify a hierarchy of main objects. It determines the order, based on which developers can see and modify different elements of artifacts. Object *sysg_artifact* is always automatically at the highest level of the hierarchy. For instance, an administrator can specify that object *State* belongs to the second highest level. Therefore, developers firstly have to create or select an artifact. After that they can manage states that belong to the selected artifact.

26.3 The Versioning Subsystem of the Metamodeling System

In this section, we propose a design for the versioning (or version management) subsystem of the metamodeling system. This subsystem can be divided into two parts—the versioning subsystem of the metamodeling system itself and the versioning subsystem of modeling systems which are created by the metamodeling system.

The versioning subsystem would allow administrators and developers to *commit* (store) a snapshot of the metamodel or an artifact, respectively, which they work on at a particular moment. This commit operation creates a numbered *revision*, which contains the changes done between the previous revision and the current one. The most current revision is also known as the *head revision*. If the head revision is the first one for that particular metamodel or artifact, then all the changes done until the moment of commit are stored in that particular revision.

We have decided to keep the versioning subsystem as light and simple as possible and therefore we have chosen not to use automatic commit. The user has to manually click the commit button which will then store the revision of the metamodel or artifact he/she works on.

Figure 26.3 represents the class diagram of the versioning subsystem of the metamodeling system (the dark grey classes). Our metamodeling system WebMeta is database-based. The versioning subsystem will also be developed as a database-based system, with all the revision data stored into database tables. In addition, the new versioning subsystem can make use of the powerful query mechanism of the DBMS.

Let us assume that an administrator has created a simple metamodel *ElementModel*, with two objects *Element* and *Relationship*. Object *Element* has sub-objects *name* and *description*, object *Relationship* has sub-objects *first_element*, *second_element* and *description*, with the first two as foreign objects referencing object *Element*. Figure 26.4 presents the class diagram of *ElementModel*.

Fig. 26.3 Class diagram of the versioning subsystem of metamodeling system

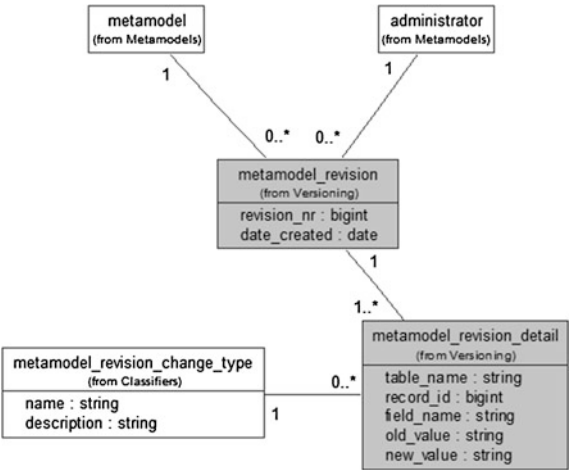
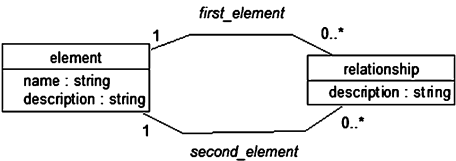


Fig. 26.4 Class diagram of the ElementModel metamodel



After an administrator clicks the commit button for metamodel *ElementModel*, the information about revision date, creator and the metamodel this particular revision is about will be stored into table *Metamodel_revision* (see Fig. 26.3). An unique revision number will be created as well. Since there is no previous revision available for that particular metamodel, all the changes done with this metamodel will be stored. The versioning subsystem will scan the metamodeling system tables *Metamodel*, *Object* and *Sub_object*, as well as the metamodel main object hierarchy settings and user rights settings tables, and store the following information into table *Metamodel_revision_detail*: table name, where the information is found; record ID in question from that table; field name in that table which contains the changed value; old field value from previous revision (empty if it is the first revision for that particular metamodel) and new field value (empty if the value was removed). Each revision detail has a revision change type, which can be one of the following: *Added*, *Updated*, *Deleted*. For every field change, a new record is created into *Metamodel_revision_detail*; deleting a metamodel element or metamodel itself will be an exception. In that case, only one record with revision change type *Deleted* will be added for the deleted element or metamodel and only its table name and record ID are stored.

Let us now assume that an administrator has changed object *Element*'s sub-object *name* to *element_name*, deleted sub-object *description*, and added another object called *Element_type* with sub-objects *name* and *description*. Also, a foreign object called *element_type* is added to object *Element*, which references object *Element_type*. If the administrator commits those changes, all field values not

changed will be skipped and only the created or updated values will be recorded to table *Metamodel_revision_detail*.

Table 26.1 represents a simplified view (some records are omitted) of the records in table *Metamodel_revision_detail*, which correspond to the initial revision (in normal text) and the head revision (second revision for that metamodel; in italic) of the changes in metamodel *ElementModel*.

Figure 26.5 represents the versioning subsystem (the dark grey classes) of the modeling systems which are created by using the metamodeling system. For every created metamodel, there will be a corresponding database schema. In these schemas, there will be several system generated tables in addition to the tables corresponding to the objects. References to those tables and the fields they contain are also added to tables *Object* and *Sub-object*, respectively. One of the tables already mentioned above is *Sysg_artifact*, where artifacts created by using the modeling system will be stored. Besides *Sysg_artifact*, there are other system generated tables, such as *Sysg_user*, where registered end-user (developer) data will be stored. If user rights settings specify that end-users can see and manage only the artifacts created by themselves, then *Sysg_user_artifact* table is used for connecting end-users with artifacts.

The tables for versioning subsystem are also created automatically to each schema corresponding to a metamodel. Table *Sysg_revision* holds information about the artifact the revision is based on and the user who committed the revision. Also, a unique revision number is also determined and stored there. Table *Sysg_revision_detail* holds information about the object and sub-object, which correspond to the database table and table field, respectively, where the changed value is located. In addition, the ID of the record the changed value belongs to is also stored into *Sysg_revision_detail*, as are the value before the change (empty, if the value was just created) and the new or updated value. Finally, there is table *Sysg_revision_change_type*, which, like the metamodeling system counterpart *Metamodel_revision_change_type*, holds classifiers for revision detail types *Added*, *Updated* and *Deleted*.

The tables of the versioning subsystem of the modeling system are very similar to those in the versioning subsystem of the metamodeling system. The main difference is, that when with metamodeling system we had to link the revision with metamodel and find the table and field name for the change in question by using the system catalog of the DBMS, then for modeling system, instead of metamodel, the revision is linked to the artifact the change took place in, and, as already mentioned, instead of the table and field names, the corresponding object and sub-object reference is provided. Because a revision is associated with one artifact at a time, there is no possibility to commit a revision together for all of the artifacts, which are located in the same modeling system.

Since the metamodeling system and the modeling systems created by the metamodeling system are all database-based and so are the versioning subsystems of those systems, the revision logging is much more easier than it would be, when, for example, the metamodels and artifacts would be stored into files, but revisions in are stored in database. We can use database queries to determine specific changes

Table 26.1 Records in table Metamodel_revision_detail corresponding to the change revisions of metamodel ElementModel

| revision_nr | table_name | record_id | field_name | old_value | new_value | change_type |
|-------------|------------|-------------------------------------|-------------------|-----------|------------------------------------|--------------------|
| 1 | Metamodel | 1 (ID for ElementModel) | name | <NULL> | ElementModel | 1 (ID for Added) |
| 1 | Metamodel | 1 | description | <NULL> | Test model for simple elements | 1 |
| 1 | Object | 8 (ID for Element) | name | <NULL> | Element | 1 |
| 1 | Object | 8 | description | <NULL> | Generic element | 1 |
| 1 | Object | 8 | metamodel_id | <NULL> | 1 | 1 |
| 1 | Object | 9 (ID for Relationship) | name | <NULL> | Relationship | 1 |
| ... | ... | ... | ... | ... | ... | ... |
| 1 | Sub_object | 15 (ID for name in Element) | name | <NULL> | name | 1 |
| 1 | Sub_object | 15 | description | <NULL> | Represents element name | 1 |
| 1 | Sub_object | 15 | sub_object_type | <NULL> | string | 1 |
| 1 | Sub_object | 15 | object_id | <NULL> | 8 | 1 |
| ... | ... | ... | ... | ... | ... | ... |
| 2 | Sub_object | 15 | name | name | element_name | 2 (ID for Updated) |
| 2 | Sub_object | 16 (ID for description in Element) | <NULL> | <NULL> | <NULL> | 3 (ID for Deleted) |
| 2 | Object | 10 (ID for ElementType) | name | <NULL> | Element_type | 1 |
| 2 | Object | 10 | description | <NULL> | Represents the type of the element | 1 |
| 2 | Object | 10 | metamodel_id | <NULL> | 1 | 1 |
| 2 | Sub_object | 20 (ID for name in ElementType) | name | <NULL> | name | 1 |
| 2 | Sub_object | 20 | description | <NULL> | Represents element type name | 1 |
| 2 | Sub_object | 20 | sub_object_type | <NULL> | string | 1 |
| 2 | Sub_object | 20 | object_id | <NULL> | 10 | 1 |
| ... | ... | ... | ... | ... | ... | ... |
| 2 | Sub_object | 21 (ID for element_type in Element) | name | <NULL> | element_type | 1 |
| 2 | Sub_object | 21 | description | <NULL> | Represents element type in Element | 1 |
| 2 | Sub_object | 21 | sub_object_type | <NULL> | integer | 1 |
| 2 | Sub_object | 21 | object_id | <NULL> | 8 | 1 |
| 2 | Sub_object | 21 | foreign_object_id | <NULL> | 10 (ID for ElementType) | 1 |

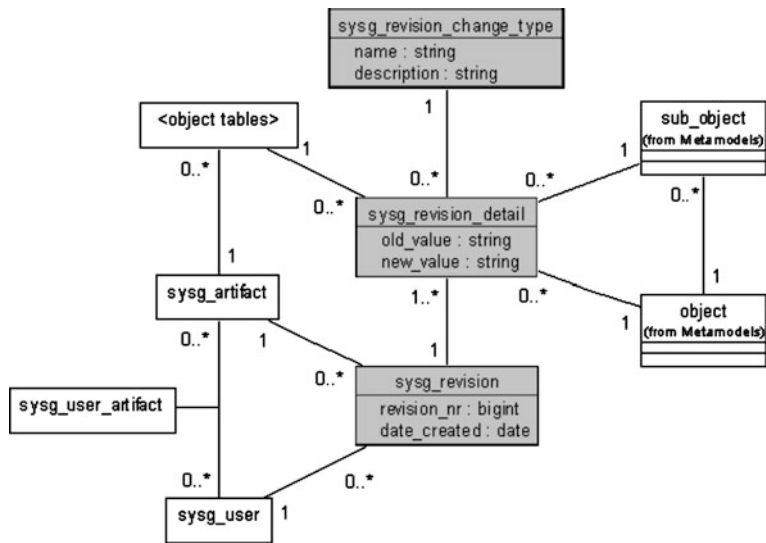


Fig. 26.5 Class diagram of the versioning subsystem of modeling systems

done in the head revision, compared to the previous one. Also, we can use queries to determine the changes needed to restore a revision of a metamodel or an artifact.

Let us assume that a developer is using the modeling system based on the initial metamodel of *ElementModel* (see Fig. 26.4). He/she wants to model the statement “a dog is an animal”. For that, the developer creates an artifact *DogStatement*, with two *Element* instances *Dog* and *Animal* and a *Relationship* instance linking those two *Element* instances, with description value “is”.

Table 26.2 represents a view of the records in table *Sysg_revision_detail*, which correspond to the initial revision of the changes in artifact *DogStatement*, created with the *ElementModel* modeling system.

26.4 Discussion

We have created a design for versioning subsystems of metamodeling system and modeling systems which are created using the metamodeling system. The design has not yet been implemented and there is no working proof-of-concept prototype. In this section, we will try to analyze (strictly on design level), whether the proposed design satisfies our needs for version management and what are the drawbacks of this particular design.

As already mentioned above, we can make use of database queries, e.g. to find differences between revisions, to find the exact changes needed to roll back (restore) a revision etc. Since both the metamodeling system and the modeling system created by the metamodeling system are database-based, there is no real

Table 26.2 Records in table Sysg_revision_detail corresponding to the change revision of artifact DogStatement

| revision_nr | object_id | record_id | sub_object_id | old_value | new_value | change_type |
|-------------|--------------------------|-----------|--|-----------|--------------------------------|------------------|
| 1 | 1 (ID for Sysg_artifact) | 1 | 1 (ID for name in Sysg_artifact) | <NULL> | DogStatement | 1 (ID for Added) |
| 1 | 8 (ID for Element) | 1 | 15 (ID for name in Element) | <NULL> | Dog | 1 |
| 1 | 8 | 1 | 16 (ID for description in Element) | <NULL> | Element representing a dog | 1 |
| 1 | 8 (ID for Element) | 2 | 15 | <NULL> | Animal | 1 |
| 1 | 8 | 2 | 16 | <NULL> | Element representing an animal | 1 |
| 1 | 9 (ID for Relationship) | 1 | 17 (ID for first_element in Relationship) | <NULL> | 1 (ID for Dog) | 1 |
| 1 | 9 | 1 | 18 (ID for second_element in Relationship) | <NULL> | 2 (ID for Animal) | 1 |
| 1 | 9 | 1 | 19 (ID for description in Relationship) | <NULL> | is | 1 |

benefit for using a file-based version management system like Apache Subversion (SVN) [8] or Concurrent Versions System (CVS) [9].

Since we have tried to keep the versioning subsystem simple and light, it is possible to extend the use of it to any subsystems of the metamodeling system, as long as the target subsystem itself is database-based. For example, in addition to the metamodeling subsystem of our metamodeling system WebMeta (covered in this paper), we can use the version management provided by the versioning subsystem also on the query subsystem [2] or the metamodel settings subsystem [1] of WebMeta.

One of the main features of a versioning management system is so-called *branching*. Branching is the duplication of an object under revision control (such as a source code file, or a directory tree) so that modifications can happen in parallel along both branches [10]. The originating branch is sometimes called the *parent branch*. *Child branches* are branches that have a parent; a branch without a parent is referred to as the *trunk* or the *mainline* [10]. In our versioning subsystem, branching could be accomplished by creating a query, which selects the latest changes of all elements of a metamodel or an artifact and then uses the data to create another metamodel or artifact, respectively. That way the newly created metamodel or artifact is the child branch of the original (parent) one. One possible extension of the versioning subsystem could be to create the functionality which allows administrators and developers to merge the changes made in branch metamodel or artifact into the original (trunk) metamodel or artifact, respectively.

One of the main drawbacks of our proposed design is that in case there is a large number of metamodels or artifacts created and the changes are committed frequently by administrators or developers, tables *Metamodel_revision_detail* or *Sysg_revision_detail* can grow very fast and the queries for selecting revision differences or other data can get very slow. Possible solution for this might be to set up table partitioning [11].

Another drawback is that this design cannot be extended to program code or graphical representation of metamodels or artifacts (e.g. diagrams). Since our metamodeling system WebMeta has been intentionally form-based for research purposes, the latter is not that important. As with the former, then one of the main purposes of a metamodeling system is to allow developers to create program code based on artifacts (models). If there is a need for a change in the code, then the artifact, not the code itself, is changed and code is re-created based on the updated artifacts [3]. Therefore, we do not really need a versioning subsystem in our metamodeling system, which would support storing the changes done in program code.

We have chosen this kind of approach to our versioning subsystem, where only the latest changes are recorded into a new revision. While this might be an advantage, because the latest changes will most probably take much less table space than the whole state of the metamodel or artifact (unless it is the first revision of the metamodel or artifact), it could be also a possible drawback. Taking a snapshot of the whole state of a metamodel or an artifact would allow us to branch and merge different revisions more quickly, because we do not have look for the latest changes throughout the revisions table.

Branching has also a drawback. If we branch both a metamodel and an artifact made based on that metamodel, we would experience a situation, where metamodel might get changed and the revisions of artifacts would not be compatible with the metamodel change. A possible solution would be to lock the metamodel and allow us to branch as another metamodel, which can then be changed.

In our experience, there is no other metamodeling system, which makes use of a version management subsystem, especially on a database level. There are, however, examples of version management systems used on modeling tools, for example AMOR system [12].

We strongly support to use version management in terms of metamodeling and modeling. Version management improves user collaboration in multi-user systems and environments. It also stores the development history, which can be useful for finding bugs or restoring a previous state of a metamodel or an artifact.

26.5 Conclusions

In this paper, we presented a design of versioning subsystems for our metamodeling system WebMeta and the modeling systems which are created by using this metamodeling system. Because both the metamodeling system and the modeling systems are database-based, we have also designed the new subsystem as such. We have described how the subsystem works on the database level, as well as discussed whether the proposed design is suitable for our needs of version management. We have pointed out the advantages and possible drawbacks of the design and investigated, whether there is a solution available for the latter and what it might be. We have reached a conclusion that this design is practicable in terms of our database-based metamodeling system and the modeling systems created by using the metamodeling system.

Future work regarding this topic must include creating a proof-of-concept prototype of the versioning subsystems for both metamodeling and modeling systems. This includes user interfaces to commit revisions for both metamodeling and modeling level, as well as to display differences between revisions, to branch a metamodel or an artifact and to merge the changes made in the child branch back to the parent (trunk) branch.

Acknowledgments This research was supported by the Estonian Doctoral School in Information and Communication Technology.

References

1. Eessaar E, Sgirka R (2010) A database-based and web-based meta-CASE system. In: Advanced techniques in computing sciences and software engineering: international

- conference on systems, computing sciences and software engineering (SCSS 08). Springer, Dordrecht, pp 379–384
2. Eessaar E, Sgirka R (2010) A SQL database based meta-CASE system and its query subsystem. In: Innovations in computing sciences and software engineering: international conference on systems, computing sciences and software engineering (SCSS 09). Springer, Dordrecht, pp 57–62
 3. Kelly S, Tolvanen J-P (2008) Domain specific modeling. Enabling full code generation. Wiley-Interscience Publication, Hoboken
 4. Melton J (2003) ISO/IEC 9075-2:2003 (E) information technology—database languages—SQL—Part 2: foundation (SQL/foundation), August
 5. MetaObjectFacility(MOF) Specification, version 1.4.1, formal/05-05-05
 6. Greenfield J, Short K, Cook S, Kent S (2004) Software factories: assembling applications with patterns, models, frameworks, and tools. Wiley Publishing, Chichester
 7. Mahnke W, Ritter N (2002) The ORDB-based SFB-501-reuse-repository. In: Jensen CS, Jeffery KG, Pokorny J, Saltenis S, Bertionio E, Böhm K, Jarke M (eds) EDBT’2002. LNCS, vol 2287. Springer, Heidelberg, pp 745–748
 8. “Apache Subversion” [online]. <http://subversion.apache.org>. Accessed 18 Oct 2011
 9. “Concurrent Versions System” [online]. <http://savannah.nongnu.org/projects/cvs>. Accessed 18 Oct 2011
 10. Berczuk S, Appleton B (2003) Software configuration management patterns: effective teamwork, practical integration. Addison-Wesley, Boston
 11. “PostgreSQL: Documentation” [online]. <http://www.postgresql.org/docs/9.1/static/ddl-partitioning.html>. Accessed 18 Oct 2011
 12. Altmanninger K, Kappel G, Kusel A, Retschitzegger W, Seidl M, Schwinger W, Wimmer M (2008) AMOR—towards adaptable model versioning. In: Proceedings of the 1st international workshop on model co-evolution and consistency management

Chapter 27

Difficulties in Understanding Object Oriented Programming Concepts

Soly Mathew Biju

Abstract Understanding object oriented concepts is always a difficult task for students. It is equally challenging for lecturers to teach these concepts. Over the years teachers have used various methods to teach these concepts. The result of a class test was analysed to identify the various areas students have difficulty in understanding. The result will help in designing course material that would focus on these areas of object oriented programming.

27.1 Introduction

Existing research shows that there have been a number of problems faced by teachers teaching programming to undergraduate students. Students find it very difficult to understand object oriented concepts like classes, constructor invocation, overloaded constructors, friend functions and other object oriented concepts [2]. Students who have been exposed to procedural programming find it a little difficult to move towards object oriented programming. It takes some time for them to understand Object Oriented concepts. Object oriented concepts consist of a number of classes within a single program.

This paper discusses the concepts of object oriented programming which are difficult for students to understand.

S. M. Biju (✉)

Tallinn University of Technology, Raja 15, Tallinn, Estonia
e-mail: SolyMathewBiju@uowdubai.ac.ae

27.2 Teaching Methods Deployed

Student centred learning methods for programming classes are very effective. Student-centred learning consists of facilitating understanding and conceptual change or intellectual development [9]. As a teaching tool, programming assignments are constructed to encourage students' development of analytical, programming writing skills. Watkins [8] argues that a student/learning-centred concept of teaching is one where high quality learning is viewed as "requiring active construction of meaning and the possibility of conceptual change on the part of the learners".

In case of programming languages, this approach focuses on teaching concepts by encouraging students to write and implement a program based on a concept to solve the given problem.

Learning can be made possible within the lecture by making students do things which are called 'active learning' [11]. For example, while teaching the concept of classes, one of the examples used in an introductory class is as follows.

```
class distance1
{
private:
    int feet;
    float inches;

public:
    void getdata();

    void showdata();
};
void distance1::getdata()
{
    cout<<"Enter the feet "<<endl;
    cin>>feet;
    cout<<"enter the inches
" <<endl;
    cin>>inches;
}

void distance1::showdata()
{cout<<feet<<"\n"
" <<inches<<"\n" <<endl;
}
```

```

int main()
{
    distance1 d1;
        d1.getdata();           //calling member
functions of the class with
        d1.showdata();//member access operator
        system ("pause");

        return 0;
}

```

The concept of encapsulation is taught. The fact that `getdata ()` and `showdata()` are member functions of the class and therefore could access all the variable members is clearly explained to the students.

When a program was given to the students to work on many of them questioned how the functions could access the variable members without passing them as parameters to the functions.

Though the students claimed to understand the concept of data encapsulation and the fact that the functions of a class can access the data members of that class, they do not really understand it unless they start writing codes based on that concept. To help them understand, these students are given home works and assignments based on the concept taught in the class.

Ramsden's [7] theory of Teaching as making learning possible seems to be appropriate in case of programming classes. Teaching is comprehended as a process of working cooperatively with learners to help them change their understanding. He also states that teaching involves finding out about students' misunderstandings, intervening to change them and create an environment for learning. In programming classes conducted by me, this is implemented by conducting quizzes and tests and providing appropriate feedback. Sometimes these tests bring to light certain concepts that have been misunderstood, in such cases; I make it a point to dedicating a part of the next session to reinstate the concept correctly.

Another method implemented by me, is using pictorial representation to describe different concepts or even use Unified Modelling Language diagrams wherever applicable.

27.3 The Research

A research was conducted in order to identify the basic object oriented concepts students have difficulty in understanding in a programming course. The students in the course have a basic procedural programming background and this course is an introduction to object oriented concepts.

27.3.1 Research Questions

The research questions were:

1. To what extent do students understand the underlying process which takes place when creating an object, specifically when the constructors are overloaded?
2. To what extent do students understand how and which constructors are implicitly called when a parameter is passed which objects construction.
3. To what extent do students understand how copy constructors are used?
4. To what extent do students understand how friend constructors are called in a program?
5. Do the students understand the concept of data hiding and data encapsulation?
6. Do the students understand the working of a friend function?

27.3.2 Research Population

The research population had 30 students who took the test in the course “Applied programming “ in C++ in the summer semester in 2008 and 2009 after some basic concepts of OOP were introduced.

27.3.3 Research Instruments

The research instrument was the class test which consisted of six questions covering one of the introduction chapters. The questions check whether the students have understood the concepts of data encapsulation, data hiding, constructors, and friend functions.

The test questions are as given below

```
class A
{
private :
int n=0;
public:
A(){
this->n=0;
cout<<"A constructor1";
}
A(int n){
this->n=n;
cout<<"A constructor2";
}
A(A &Obj){
this->n=Obj.n;
cout<<"A constructor3";
}
```

- 1 What will be printed as a result of the execution of the following statement?A a;
- 2 What will be printed as a result of the execution of the following statement?A b(9);
- 3 What will be printed as a result of the execution of the following statement?A b(a); where a is an object of the class A that has already been created.

```
using namespace std;
class Student
{
    int Studentid;
    string StudentName;
    double marks[5];
    double average;
    static int count;
public:
    Student()
    {
        Studentid=0;
        StudentName="";
        average=0.0;
        count++;
    }
    void calAvg();
    void getInfo();
    void displayInfo();
    static int get_count()
    {
        return count;
    }
    friend double
highestAvg(Student[], int);
    friend void
highestDetail(Student[], int);
};
int main()
{
    Student *stdArr;
    int count;
    cout<<"How many students do
you wish to add: ";
    cin>>count;
    stdArr=new Student[count];
    for (int i=0;i<count;i++)
    {
        stdArr[i].getInfo();
        stdArr[i].calAvg();
    }
}
```


4 Add necessary lines of code to call the friend functions.

5 Consider the class student given above

```
1.  int main()
2.  {
3.    Student s;
4.    s.id=99;
5.    s.getInfo();
6.  }
```

which line will cause an error while compiling
?why?

6 Is there an error in the friend function of the class Student defined below? If so correct it.

```
1.  void display (Student s)
2.  {
3.    cout<<average;
4.  }
```

27.4 Discussion

The result of the test is given below

In question 1, the students were required to exhibit understanding of the underlying process of creating an object. The constructor invoked by the object 'a' created in this question will initialize n to 0 and prints A constructor 1.

From Table 27.1 we can see that 30 % of the students did not thoroughly understand the process of implicit innovation of constructors. These students were confused and thought that the question had an error.

In question 2 the students were required to exhibit understanding of overriding of constructors.

The match is found with the second constructor with signature A(int).The output would be A constructor 2.

From the result above we can see that 30 % of the students did not understand this concept. They wrote down both A constructor1 and A constructor 2 in the answer sheet.

In question 3, the students were supposed to exhibit understanding of a constructor with the signature A(&Obj) that is the copy constructor is called.

We can see that 40 % of the students did not understand this.

In question 4, students were supposed to give the correct syntax for calling a friend function. A friend function does not belong to any class but is the friend of the class hence can access the data members of the class. A friend function can access data members of the class only using the dot operator.

Table 1 Test result

| Section | Mean | STDEV |
|---------|------|---------|
| 1 | 70 | 0.48305 |
| 2 | 70 | 0.48305 |
| 3 | 60 | 0.5164 |
| 4 | 70 | 0.48305 |
| 5 | 80 | 0.42164 |
| 6 | 70 | 0.48305 |

30 % of the students have not understood the concept of friend functions.

In question 5, students were supposed to exhibit their understanding of data encapsulation. The data member *id* is private hence cannot be accessed from the *main()*. 20 % of the students did not understand this concept.

Question 6 is also related to friend function. the friend function accepts a parameter of the object *s* of class *student*.

30 % of the students did not understand the concept of writing friend functions.

27.5 Conclusion and Future Research

To conclude, it is apparent from the above results that the students have a good understanding of the following concepts:

- Data hiding can be done by declaring the data member as private. A private data member cannot be accessed by functions outside the class.

On the other hand, the students had more difficulty with understanding the following concepts:

The benefits and use of friend functions were not clearly understood by the students.

The use of constructors and how overloaded constructors are invoked implicitly as soon as the object is created.

Overall, the above results are relatively good. Students were expected to do better and exhibit a better understanding of these concepts. The concept of encapsulation, classes and objects are little difficult for students to follow. This paper points out specific areas of difficulty which educators must be aware of so that they can plan the learning process accordingly.

Many researchers in this area believe in introductory level of programming, visualization for presenting new concepts for the students [3].

Academicians feel that there is a need to support the learning of OOP concepts using software tools [4–6, 11].

As a result of this study, I am planning to propose a change in the way the subject is taught.

For this subject, I plan to use the software Alice to introduce students to object oriented concepts. Alice is an Open education Resource made available to the teaching community by Carnegie Mellon. Alice was aimed at teaching computer programming to students in an interesting 3D environment.

Alice is used for teaching students to program rather than teaching students a specific programming language. It is a teaching tool for introductory computing. It uses 3D graphics and a drag-and-drop interface to facilitate a more engaging, less frustrating first programming experience [12].

Animated views can help the students in three central learning activities: Understand programs; Evaluate existing programs; Develop new programs [1].

In Alice's interactive interface, students drag and drop graphic tiles to create a program, where the instructions correspond to standard statements in a production oriented programming language, such as Java, C++, and C#. Alice allows students to understand Object Oriented Concepts better through animation programs. By manipulating the objects in their virtual world, students gain experience with all the programming constructs typically taught in an introductory programming course.

A study will be conducted to measure the understanding of object oriented concepts after the student have been introduced to Alice.

References

1. Tango SJ (1990) A framework and system for algorithm animation. *IEEE Computer* 23(9):27–39
2. Holland S, Griffiths R, Woodman M (1997) Avoiding object misconceptions. In: *Proceedings of the 28th SIGCSE*, pp 131–134
3. Lahtinen E, Ahoniemi T (2005) Visualizations to support different levels of cognitive development. In: *Proceedings of the fifth Finnish/Baltic sea conference on computer science education*, November 2005.
4. Ragonis N, Ben-Ari M (2005) On understanding the static's and dynamics of object-oriented programs. In: *ACM SIGCSE*, pp 226–230
5. Murray KA, Heines JM, Kolling M, Moore T, Wagner PJ, Schaller NC, Trono JA (2003) Experiences with IDEs and Java teaching: what works and what doesn't. *ACM SIGCSE Bulletin* 35(3):215–216
6. Roberts E (2001) An overview of MiniJava. *ACM SIGCSE Bulletin* 33(1):1–5
7. Martin E, Ramsden P (1993) An expanding awareness: How lecturers change their understanding of teaching. *Research and Development in Higher Education* 15, 148-155.
8. Watkins D (1998) A cross-cultural look at perceptions of good teaching: Asia and the West. In: Forest JJF (eds) *University teaching: International perspectives*. Garland Publishing Inc, New York, pp 19–34
9. Devlin M (2006) Challenging accepted wisdom about the place of conceptions of teaching in university teaching. *Improv Int J Teach Learn Higher Educ* 18(2):112–119 <http://www.isetl.org/jtlhe/ISSN> 1812-9129.
10. Race P, Brown S (1998) Refreshing your lecturing. *The Lecturer's Toolkit*. Kogan Page Ltd, pp 19–49
11. Ramsden P (2003) *Theory of teaching in higher education. Learning to Teach in Higher Education*. 2nd edn, RoutledgeFalmer, London, pp 06–116
12. Alice (1999) An educational software that teaches students computer programming in 3D environment. <http://www.alice.or>

Chapter 28

Real-Time System for Monitoring and Analyzing Electrocardiogram on Cell Phone

O. Muñoz-Ramos, O. Starostenko, V. Alarcon-Aquino
and C. Cruz-Perez

Abstract A novel framework on cell phone for recollecting, processing and interpretation of patient's electrocardiograms ECG as part of development of health care and assisted living environments is presented in this paper. The proposed architecture and algorithm provide continuous detection of the QRS complex during real time ECG monitoring and interaction between doctor and patient expanding coverage of medical services. The developed procedure for heart activity monitoring uses a set of filters for image noise reduction and computes ECG signal gradient for identification of the components with the greatest slope. To highlight the steepest parts of ECG, the absolute value of gradient is averaged over a moving window of 80 ms considered as the minimum duration of QRS complex. In the decision phase, a peak detector is applied. The height of detected peaks is compared to the threshold determined as the signal-to-noise ratio for final definition of heart rate. The designed prototype has been tested using standard MIT-BIH Arrhythmia Database and evaluated confirming that system has good compromise between high transmission and processing speed and satisfactory accuracy, which does not fall below the precision of commercial equipment for heart monitoring.

O. Muñoz-Ramos (✉) · O. Starostenko · V. Alarcon-Aquino · C. Cruz-Perez
Research Center CENTIA, Department of Computing, Electronics and Mechatronics,
University de las Américas-Puebla, 72820, Cholula, Mexico
e-mail: orlando.munozrs@udlap.mx

O. Starostenko
e-mail: oleg.starostenko@udlap.mx

V. Alarcon-Aquino
e-mail: vicente.alarcon@udlap.mx

C. Cruz-Perez
e-mail: claudia.cruzpz@udlap.mx

28.1 Introduction

In recent years the computing power of mobile phones and their capabilities have increased considerably with the addition of new processors, video cameras, sensors and actuators. Development of operating systems, such as Apple iOS, Google Android, Windows Phone 7 from Microsoft and RIM BlackBerry make cell phones today capable to perform complex tasks previously restricted to these devices.

The heart disease control on mobile devices is very important area because they have high incidence among the population. There are many commercial systems used to support health care and assisted living environments. As usually, personal mobile health monitoring systems support the rehabilitation process of patients after heart surgery or recovering from a heart attack [1, 2], recollecting patient's electrocardiograms ECG, heart rate and oxygen level in the blood [3], remote control of cardio implants [4], tracking of patients by GPS in case of emergency [5], expanding coverage of medical services to rural communities [6], real time interaction and communication between doctors and patient [7], and others used for diagnosis assistance [8–10].

The widely used algorithms for analysis of heart activity are based on interpretation of ECG, detection of the QRS complex, heart rate as well as other vital signs [7, 8, 10]. The well-known approaches include neural networks, genetic algorithms, wavelet transforms, heuristic methods, etc. [11, 12]. However, the complexity of these approaches and large number of floating point operations during ECG analysis make little feasible the utilization of cell phones.

Therefore, instead of mentioned data processing approaches the simplified algorithms are suggested to apply. They use filters, analyze amplitude and slope of ECG signal, operate with thresholds and gradient to detect the QRS complex in systems with limited resources, such as a cell phone [13]. For example, the most used approaches for detection of the QRS complex in real time are the Pan-Tompkins and Hamilton algorithms based on signal filtering and computing derivatives [14, 15]. Table 28.1 resumes some high performance systems for heart activity monitoring.

The paper has the following structure. In the second section, the proposed architecture and algorithm are described. The Sect. 28.3 presents the designed framework for heart monitoring and its implementation. In the Sect. 28.4 the evaluation of system is discussed. Finally, the contributions of the paper are presented in conclusions.

28.2 The Proposed Architecture and Algorithms

After analysis of well-known systems it was detected that the key parameters of heart monitoring are the sampling frequency and number of ECG leads sent to cell phone. Some systems report sampling frequency between 100 and 1,000 samples

Table 28.1 Mobile healthcare systems and their characteristics

| System, device | Data acquisition | Output parameters | Used algorithms | Platform |
|---|--|-------------------------------------|--------------------------|----------------|
| Personal health monitoring [1, 2] | ECG via bluetooth | Heart rate, arrhythmia detection | Hamilton | Windows mobile |
| Wireless children monitoring [3] | ECG via wireless | Heart rate, arrhythmia detection | Hamilton | Windows mobile |
| AliveECG [8] | ECG via wireless | Heart rate | Variable sampling | iOS, |
| Android | | | | |
| H'andy Sana 210 [9] | ECG, embedded | Heart rate, blood pressure, glucose | Retransmission to server | – |
| Mobile heart monitoring [10] | ECG via bluetooth | Heart rate | Retransmission to server | Android |
| Mobile biotelemetry system [7] | ECG via bluetooth | Heart rate, arrhythmia detection | 100–1,000 sps | Windows mobile |
| HeartToGo [6] | ECG via bluetooth | Heart rate, arrhythmia detection | Pan Tompkins, 300 sps | Windows mobile |
| Mobile personal ECG monitoring System [5] | Temperature, ECG, blood pressure via bluetooth | ECG samples | Retransmission to server | – |

per second (sps) as recommended value [6, 7] considering that high sampling frequency provides better signal quality but increments data quantity reducing battery life of cell phone due to continuous work of processor.

28.2.1 Generalized Architecture of Heart Monitoring on Mobile Device

The health-monitoring system on mobile devices may be introduced by three-level architecture containing physical layer of data acquisition, mobile device-server communication layer and application layer for data processing shown in Fig. 28.1.

The input data acquired by sensors represent patient's vital and physiological signs, for example, ECG, heart rate, blood pressure, temperature, volume of oxygen in the blood, etc. A mobile device operates as communication node for connected sensors, preprocessing recollected data before transmission them to remote monitoring or emergency centers. The preprocessing is used for data sampling, A–D conversion, compression, storage, and visualization on mobile device. Additionally, a mobile device receives and visualizes notifications, alerts



Fig. 28.1 Block diagram of the system for heart monitoring on mobile device

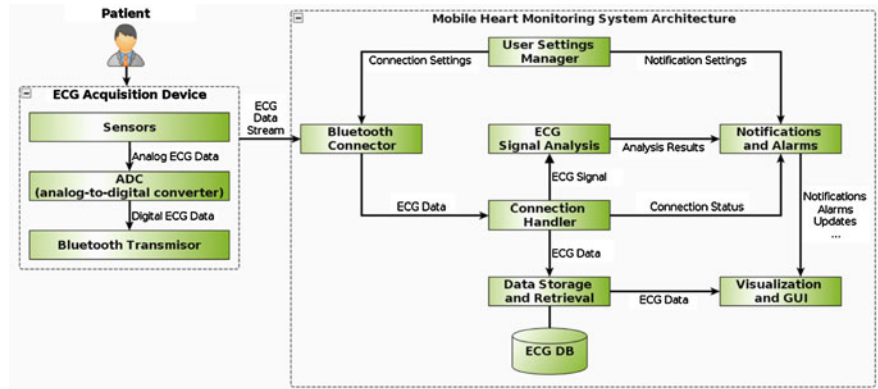


Fig. 28.2 Architecture of system for ECG acquisition and preprocessing on mobile device

and recommendation as result of applying different healthcare services. The proposed system for ECG monitoring is presented in Fig. 28.2. It has been designed with some modifications: all operation, such as for ECG reception, processing and visualization are implemented on cell phone, which does not delegate any data processing task to server. It is possible if a cell phone has enough processing power and sufficient amount of memory. The interaction of cell phone with a server is provided for emergency messages or alerts including situations when a doctor requests data form mobile device of monitored patient.

The *Bluetooth Connector* establishes communication with the ECG acquisition device, handles address acquisition and user preferences for transmission by that device. The *Connection Handler* is responsible for receiving data sent by ECG acquisition device preparing them for analysis, storage and monitoring changes in the state of connection and generation of notifications or alerts for user. The *ECG*

Signal Analysis module implements the proposed algorithm to detect QRS complex and patient's heart rate.

The obtained results are processed by module of *Notification and Alarm*, which generates and distributes the corresponding messages about events that may occur in analyzed ECG, or because of changes in connection status, for example, loss of connection, interruption, changes in heart rate, etc.

Data Storage and Retrieval module is responsible for persistent and secure storage of ECG data and ensures their availability for heart monitoring applications. The control of system is provided according to preferences defined by *User Settings Manager* (record, store and share information, such as user identification, duration of analysis session, definition of a default acquisition device, etc.).

28.2.2 The Proposed Algorithm for Detection of QRS Complex

The proposed algorithm for ECG analysis is based on detection of QRS complex, which is the most notable representation of the cardiac cycle. The QRS detection algorithm has been implemented using as base the Pan-Tompkins QRS detector improved by Hamilton [14, 15].

Data flow chart of the algorithm and results of applying filters to ECG signal are shown in Figs. 28.3 and 28.4 respectively.

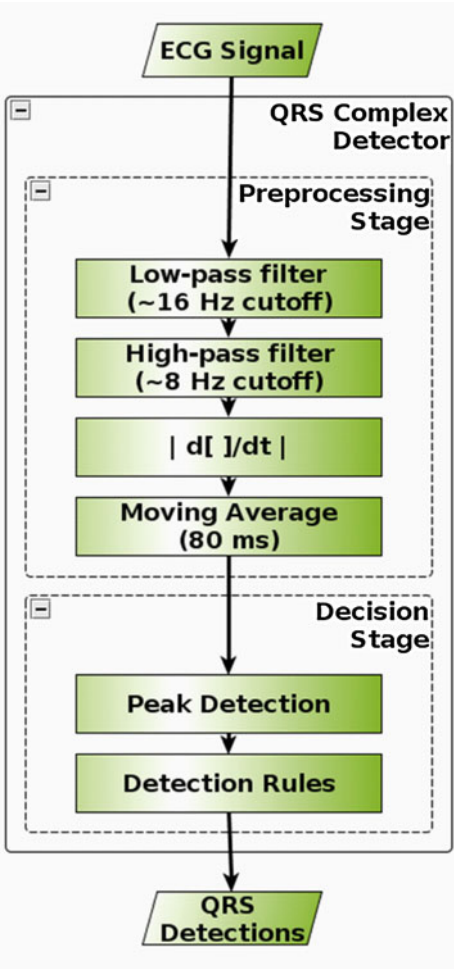
On the preprocessing stage, the ECG signal passes through a series of filters: Low pass filter is used for reduction of signal noise. The high pass filter and first derivative are applied for detection of signal gradient and identification of the components with the greatest slope. To highlight the steepest parts, the absolute value of gradient is averaged over a moving window of 80 ms considered as the minimum duration of the QRS complex. In the decision phase, a peak detector is applied. The height of detected peaks is compared to the threshold determined as the S/N signal-to-noise ratio. If the height of peak is higher than threshold, it is considered as a signal peak, otherwise it is interpreted as a noise peak. Finally, the detected signal peaks corresponding to QRS complex are selected and used for computing heart frequency.

The advantage of this algorithm is its adaptability to any heart rate. In addition, the implementation of filters is provided by simple and fast shift operations with integer numbers that permits to obtain satisfactory results on cell phone with limited speed, memory resources, and computing power.

28.3 Designed Framework for Heart Monitoring

In order to evaluate the performance and efficiency of the proposed architecture and algorithm the framework for ECG analysis on cell phone has been designed. It consists of ECG acquisition module, heart activity monitoring module, and ECG

Fig. 28.3 QRS complex detector: flow chart of the proposed algorithm



storage and management module shown in Fig. 28.5. Electrocardiograms used by this framework are obtained from *MIT-BIH Arrhythmia Database* available in [16]. This database contains 48 records of 30 min two-channel ECGs with expert annotations indicating occurrence of the QRS complex.

ECG acquisition module via *BluetoothConnector* supports communication between ECG tool and phone by RFCOMM protocol, which allows emulating a serial port on both devices to send the connection request and desired recording time of ECG. Before ECG acquiring the *ConnectionHandler* defines signal characteristics, such as sample rate, calibration value produced by ADC corresponding to 0 volts (*ADC zero*), and gain factor of ADC (*ADC gain*). During data recovery, ECG samples are converted to corresponding values in millivolts (mV) according to Eq. 28.1 used then for correct ECG displaying.

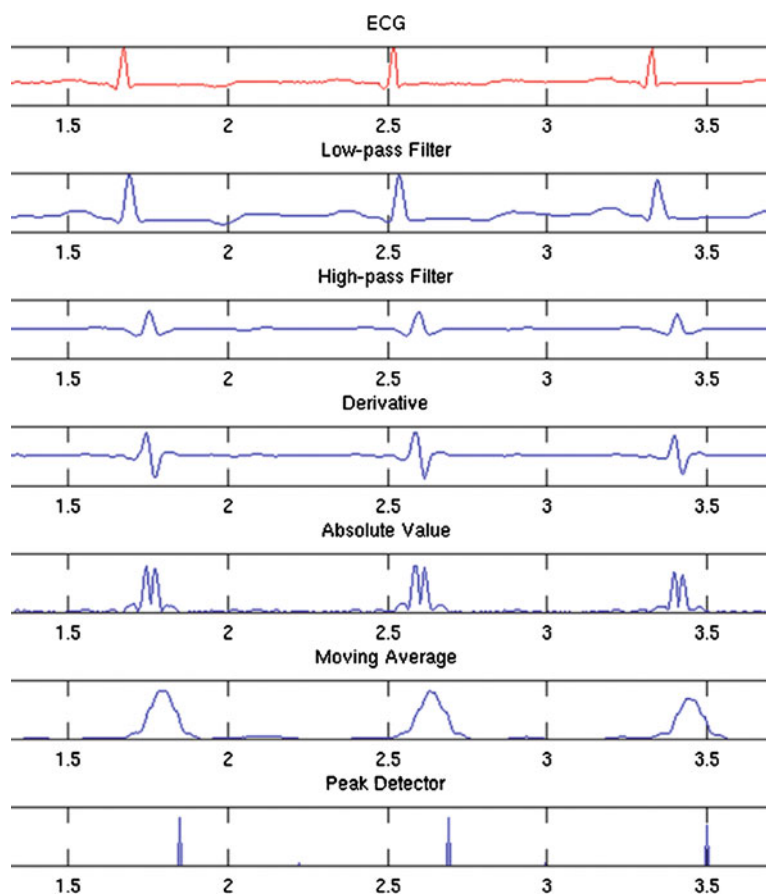


Fig. 28.4 Application of filters in QRS complex detector

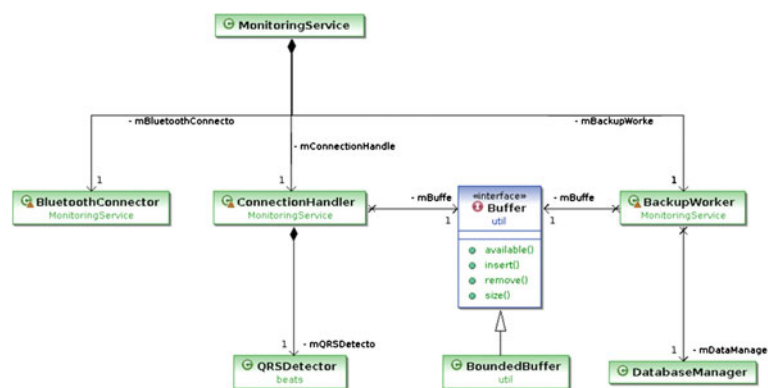


Fig. 28.5 Class diagram of the framework for heart activity monitoring

$$mV = \frac{\text{sample value} - \text{ADC zero}}{\text{ADC gain}} \quad (28.1)$$

MonitoringService supports ECG data acquisition, processing and analysis, management of database, user interaction interface and system testing. Additionally, it delegates responsibilities to associated classes and defines a set of methods to describe the changes in state of connection (on hold, online, offline, etc.) and state of heart activity.

The *BackupWorker* receives data from *Buffer* and feeds *DatabaseManager*. *QRSDetector* determines the time of event occurrence and updates the value of heart rate using time interval between the last two QRS complexes. *BoundedBuffer* supports a graphical interface that allows users to control the system, select data acquisition device, define user preferences, start or stop transmission, display the ECG, and others.

The prototype of system for heart activity monitoring has been implemented on Linux operating system, particularly, on the Ubuntu version 10.10 for 64-bit processors, using Eclipse development environment 3.6 and Toolkit WFDB library functions [16]. Cell phone is Samsung Galaxy S, IG9000 with ARM Cortex A8 processor at 1 GHz, with 515 MB RAM, 5 GB of internal memory and Bluetooth radio.

One of the most important requirements of heart monitoring system is its ability to store, management, and retrieval of ECG records over long periods. The storage capacity for ECG signals is limited by used sampling frequency. On cell phone SQLite database manager has been used, which supports various operating systems for mobile devices like Symbian, iOS, BlackBerry, Android. The stored ECG records include parameters of data acquisition device (*frequency*, *adc_zero*, *adc_gai*, *sample_number* and *sample_value*), identification number, date of created record (*record_id* and *created_at*) and time of occurrence of QRS complex (*beat_time*). As result, system with SQLite database provides simple and fast management of ECG data on cell phone.

Designed interface provides *Menu of Options* that allows user to define an ECG acquisition device, analyze a list of records stored in database, specify desired transmission time and visualize real-time ECG. Some examples of applications for continuous heart monitoring are shown in Fig. 28.6.

28.4 Experiments and Discussion

To determine the feasibility of the prototype and evaluate performance of the proposed algorithm some tests have been done. They include analysis of battery behavior during continuous heart monitoring, required time for transmission of ECG and precision of heart rate interpretation by the algorithm. The power consumption defines the time of ECG recording without recharging the battery. With the battery of Samsung Galaxy S IG9000 cell phone with 100 % of capacity

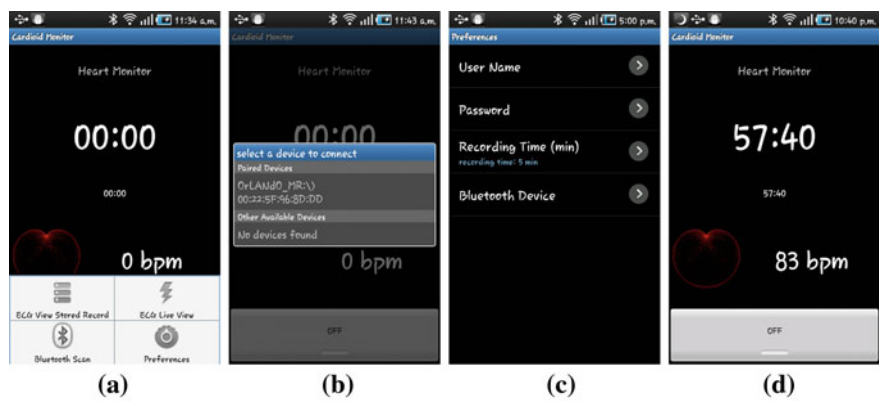


Fig. 28.6 **a** Principal menu of options. **b** Selection of ECG acquisition device. **c** User preferences interface. **d** Main GUI for continuous monitoring

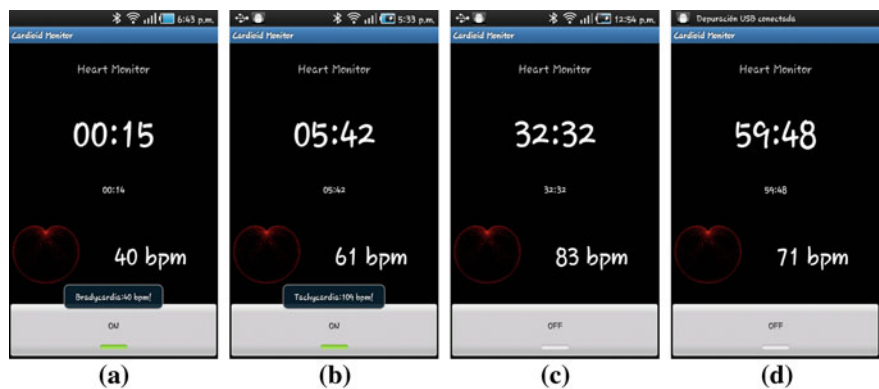


Fig. 28.7 Alarm notifications **a** bradycardia and **b** tachycardia. Transmission and storage time measurements in tests with **c** 200 sps and **d** 360 sps

(1,500 mAh) the operational time was 6 h, 14 min for sequential transmission of 12 complete (30 min each) and one partial records supporting monitoring until the battery is completely discharged.

For example, even though the battery is low, system processes the input signals and detects emergency events when the value of heart rate is less than 60 (bradycardia) or greater than 100 (tachycardia) beats per minute. Cell phone generates audible alarm, which is maintained until the frequency heart returns to normal range of 60–100 beats per minute. Some examples of these notifications and are depicted in Fig. 28.7 a, b. In Fig. 28.7 b the system detects and notifies occurrence of rapid heartbeats while the interface still presents the previous heartbeat.

For performance evaluation of our prototype with respect to ECG transmission, processing and storage time, the 30-min ECG records have been transmitted from *MIT-BIH Normal Sinus Rhythm Database* and *MIT-BIH Arrhythmia Database*

using different sampling frequency of 128, 200 and 360 sps. For registration of record transmission and processing-storage time, the prototype uses two timers shown in principal interface. The first big size timer measures the time of ECG transmission to phone, while the second small size timer measures the time from start of transmission until the ECG is stored in the database.

The results of this test are illustrated in Fig. 28.7c and d. In the cases of 128 and 200 sps, the transmission time approaches the duration of 30-min ECG, such as 31:42 and 32:32 (in Fig. 28.7c) minutes respectively. However, in the case of 360 sps, the transmission time is increased almost in 50 % (59:48 min in Fig. 28.7d). The values of small size timer in all three cases were similar to the transmission time indicating that the time of data processing and storage takes less than one second.

In Fig. 28.7d the heart rate value (71 bpm) obtained with 360 sps is different comparing with the values obtained in the other cases (83 bps for 200 sps). It means that the computed result with 360 sps is more accurate due to higher quality of received signal. It is because the heart rate precision depends on the sampling frequency.

After statistical analysis of multiple tests with different records and sampling frequencies the relative error $\delta(\%)$ of computing the heart rate obtained according following equation

$$\delta(\%) = \frac{|N_{\max/\min} - N_{real}|}{N_{real}} * 100\% \quad (28.2)$$

lies in the range of $\pm 10\%$ for 128 samples per seconds, $\pm 8\%$ for 200 sps and $\pm 4\%$ for 360 sps. $N_{\max/\min} - N_{real}$ is the absolute error of real value N_{real} of heart rate measurement with respect to maximum or minimum measured values $N_{\max/\min}$. However, because of significant increment (almost twice) of transmission time for high sampling rate, it is not recommended to use sampling with more than 200 sps despite using high technology phone like Samsung Galaxy S IG9000. The increment of transmission time for 128 and 200 sps does not exceed 5 and 8 % respectively. This is a good compromise between high speed and satisfactory accuracy, which does not fall below the precision of commercial equipments for heart monitoring.

Visualization of ECG in real time is a very useful feature of the proposed framework for heart monitoring. However, some concurrent operations, such as ECG acquisition by cell phone, ECG retrieval from database and plot of ECG on display have a significant impact on system performance. For example, in Fig. 28.8 the ECG plot of 4:11 min is shown even though cell phone has completed 5:14 min of data reception and storage in database. Additionally, the computed heart rate is more precise after 5:14 min (73 bps in Fig. 28.8b) than after 4:11 min presented in plot (71 bps in Fig. 28.8 a). Evaluating low cell phone performance during graphical visualization of ECG plot, the heart rate relative error does not exceed $\pm 3\%$ in the case of ECG received signal encoded by 200 sps.

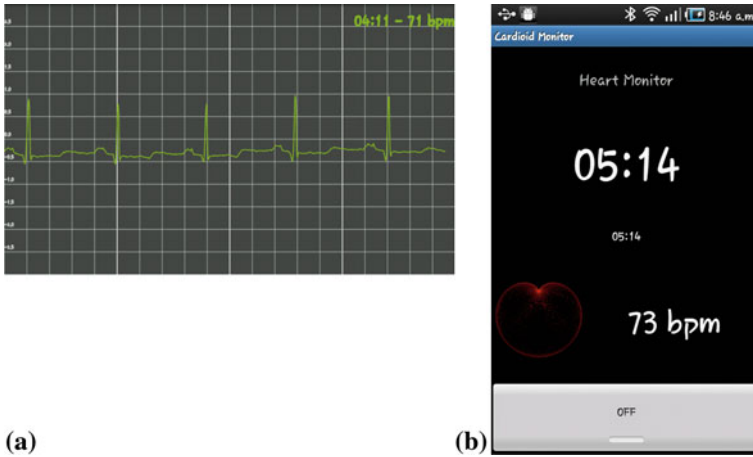


Fig. 28.8 **a** ECG plot during real time monitoring and **b** corresponding heart monitor

This problem may be solved by using mobile phone with high speed and sufficient processing power as well as by extension of bandwidth of data communication channel with ECG transmission tool.

28.5 Conclusion

The proposed architecture, which provides continuous data transmission and processing on mobile device with limited computing resources, may be considered as the conceptual contributions of this paper. This approach has sufficient merit to be used as a reference in development of applications, such as personal health monitoring system, mobile learning assistant, remote detection and control, etc. Additionally, the proposed and implemented algorithm for ECG signal receiving, processing, storage and visualization provides successful monitoring of heart activity.

The relative error of computing the heart rate is less than $\pm 10\%$ in the worst case when sampling frequency of ECG signal is 128 sps, and it is less than $\pm 4\%$ for 360 sps.

In terms of practical contributions, the proposed architecture integrates various emerging technologies, which with minimum computational resources has operational performance enough to be used as heart monitoring system on cell phone with the ability to process and visualize in real time ECG signals, detect and manage situations of risk and provide the interaction between doctor and patient.

It is important to note that the prototype does not exclude diagnosis by a doctor. Our intention is to propose applications that serve as the approach towards

development of more complex health assistance systems expanding coverage of medical services.

Acknowledgments This research is sponsored by Mexican National Council of Science and Technology, CONACyT, Projects: #154438, and #156228.

References

1. Gay V (2009) A mobile rehabilitation application for the remote monitoring of cardiac patients after a heart attack or coronary bypass surgery. In: Proceedings of the international conference on pervasive technologies, Greece, pp 235–238
2. Leijdekkers P (2008) A self-test to detect a heart attack using a mobile phone and wearable sensors. In: International symposium on computer-based medical systems (CBMS), Finland, pp 93–98
3. Kyriacou E (2009) System for monitoring of children with arrhythmias. In: International conference on pervasive technologies Greece, 2009, p 668
4. Biotronik (2010) BIOTRONIK—excellence for life. Dec 2010. <http://www.biotronik.com/es/lam/2246>
5. Belgacem N, Boumerdassi S (2009) Mobile personal electrocardiogram monitoring system with patient location. In: ACM international workshop on medical-grade wireless networks, Louisiana, pp 69–72
6. Kulkarni S (2008) Smartphone driven healthcare system for rural communities in developing countries. In: International workshop on systems and networking support for health care and assisted living environment, Colorado, pp 1–3
7. CardioNet. Mobile Cardiac Outpatient Telemetry (MCOT) – Cardiac Telemetry CardioNet Event Monitors. December, 2010, available at <http://www.cardionet.com/>
8. AliveCor (2011) AliveCor—mHealth for iHumans. Jan 2011. <http://alivecor.com/>
9. H'andy Sana (2010) The doctor in the pocket. Dec 2010. <http://handysana.com/>
10. IMEC (2010) Monitoring your health with your mobile phone. Oct 2010. http://www2.imec.be/be_en/press/imec-news/wirelesshealthnecklaceinterface.html
11. Alarcon-Aquino V, Starostenko O (2009) Detection of microcalcifications in digital mammograms using the dual-TREE complex wavelet transform. *J. Eng Intell Syst* 17(1):49–63
12. Gonzalez R, Woods R (2007) Digital image processing. Prentice Hall, New Jersey
13. Starostenko O, Alarcon-Aquino V (2010) Computational approaches to support image-based language learning within mobile environments. *J Mobile Learn Organ* 4(2):150–171
14. Pan J, Tompkins WJ (1985) A real-time QRS detection ALGORITHM. *IEEE Trans Biomed Eng* 32:230–236
15. Hamilton P (2002) Open source ECG analysis. *J Comput Cardiol* 29:101–104
16. Goldberger AL et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101(23):e215–e220. <http://www.physionet.org/physiobank/>

Chapter 29

Research of Camera Track Based on Image Matching

Yuan Wang

Abstract Description of camera motion is one of the critical issues to implement automatic return for camera. This paper employs the Affine Transform Model to describe global motion, thus reconstructing movement of the camera according to information extracted from features of adjacent frames. Algorithm for feature point extraction is mainly discussed. In combination with secondary matching, SIFT (Scale Invariant Feature Transform) is improved by taking the density of points into consideration. The experimental results show that this method enhances the precision of matching with good real-time performance.

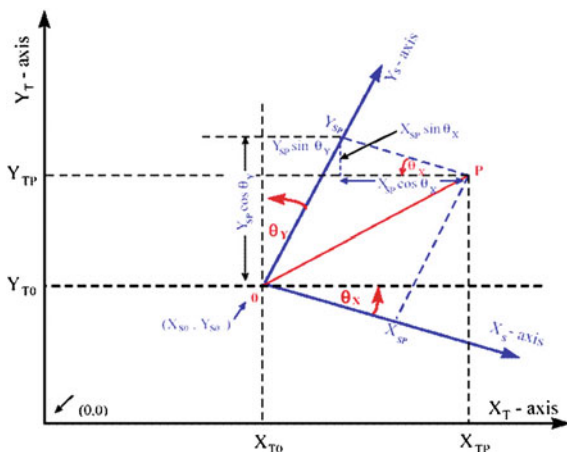
29.1 Introduction

Computer vision, a scientific discipline which thrives since the middle of 1960s, is concerned with the theory and practice for building artificial intelligence systems that extract information from visual data. In [1], it is described as a complement of biological vision as mentioned. In [2], we are informed that computer vision is widely used in many areas. A lot of research work has been carried out on image-based 3D modeling, multi-view geometry, structure-from-motion, etc. In [3], computer vision is applied in UAVs to improve their autonomies both in flight control and perception of environment around them.

Image matching is one of the key points of computer vision. In [4], it indicates that image matching is widely applied in medical image analysis, remote sensing data analysis, computer vision and pattern recognition, image segmentation, etc. In

Y. Wang (✉)

Northeastern University, Shenyang 110819, Liaoning, China
e-mail: oleg.starostenko@udlap.mx

Fig. 29.1 Affine translation

[5], imaging matching is divided into three classifications: Gray Image-based, feature-based and explanation-based. In [6], we learn that it is the high precision, strong robustness, fast matching speed and matching automation that be pursued by researchers. Finding corresponding image points precisely with good real-time performance is a very active field of research. In [7], Quan Wang takes an approach which treats the image matching problem as a recognition problem of spatially related image patch sets. In [8], Alhwarin. F proposes two modification of the popular SIFT algorithm to accelerate features matching, which involves splitting the SIFT features into two types and extending them by a new attribute. In [9], Grishin. V. A. puts forward image processing in two channels with substantially different resolution as a method of computational costs reduction.

In [10], for enhanced matching precision of algorithm based on SIFT, Mortensen adopts a global context vector of 60 dimensions similar to shape contexts. But in [11], the performance of the descriptor is impaired through PCA dimensions descending. In [12], Jieyu Zhang proposes a method of correcting SIFT mismatching based on spatial distribution descriptor which involves too much computational cost. In this paper, improvement for the method of feature points extraction, selection and matching is made.

29.2 Description of Motion Based on Parameter Model

Relative movement that occurs between adjacent images taken on the same plane, for example translation, rotation or the blending of them, can be treated as affine translation between two coordinate axes. In this paper, affine Transform Model of six parameters is employed. Affine translation is shown in Fig. 29.1.

As is shown in Fig. 29.1, coordinates of P in the two coordinate system are (X_{SP}, Y_{SP}) , (X_{TP}, Y_{TP}) respectively. The transformation between them is expressed as:

$$\begin{pmatrix} X_{TP} \\ Y_{TP} \end{pmatrix} = \begin{pmatrix} X_{T0} \\ Y_{T0} \end{pmatrix} + \begin{pmatrix} \cos \theta_X & \sin \theta_Y \\ -\sin \theta_X & \sin \theta_Y \end{pmatrix} * \begin{pmatrix} X_{SP} \\ Y_{SP} \end{pmatrix}$$

It can be further simplified as:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a & b \\ d & e \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} c \\ f \end{pmatrix} = R \begin{pmatrix} x \\ y \end{pmatrix} + T$$

where R and T represent:

$$\text{Rotation matrix: } R = \begin{pmatrix} a & b \\ d & e \end{pmatrix} = \begin{pmatrix} \cos \theta_X & \sin \theta_Y \\ -\sin \theta_X & \sin \theta_Y \end{pmatrix}$$

$$\text{Translation matrix: } T = \begin{pmatrix} c \\ f \end{pmatrix} = \begin{pmatrix} X_{T0} \\ Y_{T0} \end{pmatrix}$$

Each parameter can be solved in formulae as follows:

(1) Rotation angle

We assume:

$$a = (y'_2 - y'_1) * (x_2 - x_1) - (x'_2 - x'_1) * (y_2 - y_1)$$

$$b = (x'_2 - x'_1) * (x_2 - x_1) - (y'_2 - y'_1) * (y_2 - y_1)$$

When $|b| > 0$, rotation angle = $\arctan(a/b)$, otherwise it is 0.

(2) Scale factor

We assume:

$$a = (x'_2 - x'_1)$$

$$b = (x_2 - x_1) * \cos(\text{rotation}) - (y_2 - y_1) * \sin(\text{rotation})$$

When $|b| > 0$, scale factor = a/b , otherwise it is 0.

(3) Deviation along the x axis and y axis

$$x_{\text{translate}} = x'_1 - \text{scale} * (x_1 * \cos(\text{rotation}) - y_1 * \sin(\text{rotation}))$$

$$y_{\text{translate}} = y'_1 - \text{scale} * (x_1 * \sin(\text{rotation}) + y_1 * \cos(\text{rotation}))$$

Then, in the affine Transform Model of six parameters, each parameter can be expressed as:

$$a = e = \text{scale} * \cos(\text{rotation})$$

$$b = -d = -\text{scale} * \sin(\text{rotation})$$

$$c = \text{xtranslate}$$

$$f = \text{ytranslate}$$

In this paper, two-dimensional motion is described by resolving parameters. First, read the image sequence to acquire several corresponding points $[x \ y]^T \rightarrow [x' \ y']^T$ between adjacent ones. Second, resolve motion parameters according to the formulae. Finally, reconstruct camera movement in the light of the parameters obtained.

29.3 Extraction, Selection and Matching Algorithm

In [13], it is maintained that matching algorithm based on gray level fails to meet real-time demand, while feature-based algorithm can do a quick work. In addition, it possesses strong distortion-proof ability from gray level and shape.

In SIFT the feature points are extracted in multi-scale space. First, the image is first convolved with Gaussian-blurs at different scales. Then the Difference-of-Gaussian images are taken from adjacent Gaussian-blurred images per octave. Once DoG images have been obtained, feature points are identified as local minima/maxima of the DoG images across scales. If the pixel value is the maximum or minimum among all compared pixels, it is selected as a candidate feature point and its scale is recorded. Next, fit the feature points with a three-dimensional quadratic function to specify their location and scale, while discarding low-contrast ones and eliminating edge responses, which substantially improves matching performance and stability.

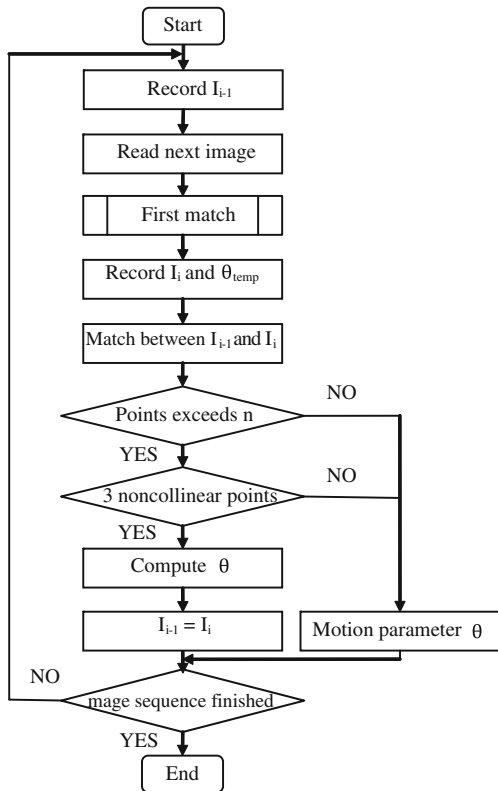
In this paper, middle area is taken as ROI (region of interest) and produces feature point solely in the light of [14]. The entire feature points obtained through matching participates in computation, with a comprehensive consideration of which delivers the final motion parameters.

A further constraint is put forward in order to ensure feature points being identified precisely by reducing cumulative error, thus corresponding to the continuity of motion. This is done by matching between results of first matches. Record common information, which includes motion parameters θ , between image $i-1$ and i as I_{i-1} and get I_i in the same way. Then make a match between I_{i-1} and I_i for D_i , which reflects common information of all. This is called secondary matching and can track the feature points.

The algorithm is shown in Fig. 29.2.

Unmatched points that scatter in similar structure mismatch from time to time because only local gradient information in neighborhood counts in SIFT descriptor. So some more distinctive descriptors are expected. Edge feature is taken into consideration as a new constraint to remove false points.

Fig. 29.2 The flow chart of algorithm based on secondary matching



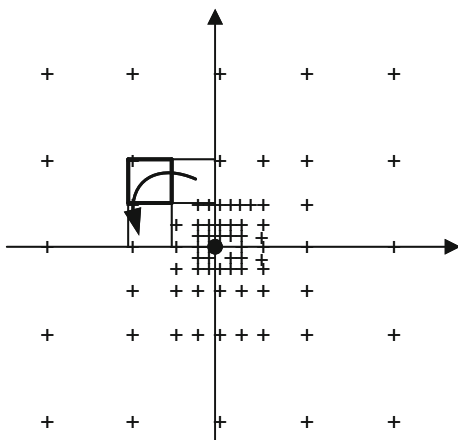
Jieyu Zhang proposes a mismatch correcting algorithm based on spatial distribution descriptor which achieves a relatively high match rate. The partitioning method Jieyu Zhang employs is shown in Fig. 29.3. However, the calculation procedures seems rather complicated due to the large number of feature points that SIFT gets. So in this paper, image is partitioned in this way: center on a specific feature point, we get segments in upper left, lower left, upper right and lower right in turn. Side-length of the square is set as L which is the minimum distance the point reaches image edge. Thus, global information is preserved while iteration is greatly reduced.

When too much feature points are extracted, real-time performance should be guaranteed by further constraint. Feature points that located in area where points are sparse can be identified more easily with less mismatching, and vice versa. On this point, I advance a density-based method to wipe off candidates that is falsely report.

The expression of density of points is of the form:

$$C_{x,y} = \sum_{x+m/2}^{x+m/2} \sum_{y+n/2}^{y+n/2} B_{i,j}$$

Fig. 29.3 The method of partitioning image



$C_{x,y}$ denotes the density of point (x,y) in a certain rectangular area $(m \times n)$. In this paper I choose square as the computational area, which means $m = n$. $B_{i,j}$ indicates whether or not (i,j) is a feature point. $B_{i,j}$ is given the value of 1 when $B_{i,j}$ is a feature point; otherwise it will be 0.

According to center-biased distribution characteristic of motion vector, uniform extraction can be achieved by doing this: when feature points exceed the threshold n , partition the middle area of the image into m regions in accordance with cross center-biased distribution model, while feature points being divided into m groups. If feature points within a certain region exceed n/m , put them in increasing order after computing density for each feature point. Then take the first n/m points. Total number will be kept within n after traversing the whole area.

By now, participants are greatly reduced in parameter calculation. Separated point pairs that gain through selection into several groups. Each group is called coefficient matrix and contributes to deliver a set of global motion parameter. Then, establish a histogram to obtain motion parameter with the highest frequency, namely $p_{\max} = p(\theta)$. Take the comprehension of all parameters that appear within $[\theta - d, \theta + d]$ (where the value of d is concerned with the number of point pairs and the size of groups) as the final result. Flow gram of this algorithm is shown in Fig. 29.4.

Whole workflow:

- (1) Feature points selection: Utilize partition and density to select qualified points.
- (2) First match: Record related information.
- (3) Traverse all of the images.

29.4 Experimental Result

Original images in this experiment are shown in Fig. 29.5.

Fig. 29.4 The flow chart of sifting algorithm based on density

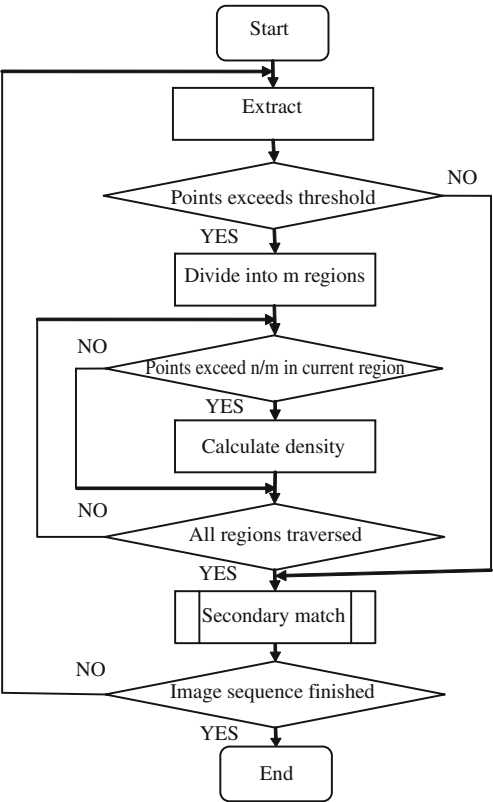


Fig. 29.5 Original image



Matching result of SIFT is shown in Fig. 29.6, from which we can see that mismatch happens.

Figure 29.7 shows matching result of the correcting algorithm based on spatial distribution descriptor.

Fig. 29.6 Matching result of SIFT

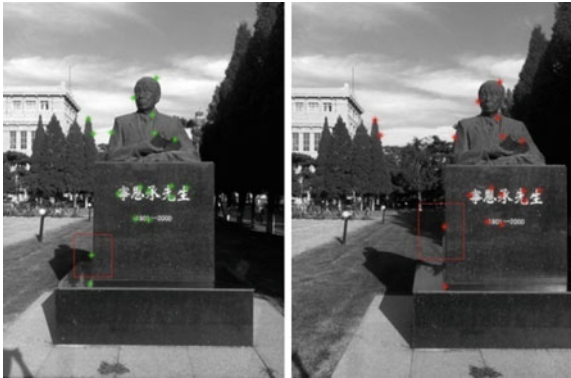


Fig. 29.7 Matching result of the correcting algorithm based on spatial distribution descriptor



Table 29.1 Comparison of the performance index of two methods (S)

| Index | Correcting algorithm | Improved method |
|-------|----------------------|-----------------|
| Pairs | 185 | 185 |
| Time | 4 | 0.0144 |

But the computational cost fails to meet the demand of real-time performance. Table 29.1 manifests that the method put forward in this paper does a better job.

Figure 29.8 demonstrates effect images that have been treated by the improved method.

Tests for algorithm based on partitioning and density are done. According to cross center-biased distribution model, I test 3-region, 5-region and 9-region respectively.

Experimental results in Table 29.2 reveal that 5-region practice excels in convergence along x axis, while movement distance per frame along y axis is relatively stable. Moreover, it is superior to 9-region practice for its easy operation.

Figure 29.9 demonstrates the schematic of partitioning.

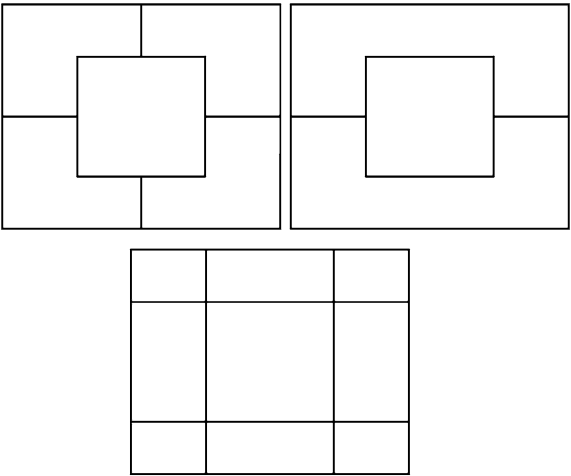
Fig. 29.8 Matching result of improved method



Table 29.2 Comparison of four partition scheme

| Partitioning | \bar{x} | $\text{var}(x)$ | $\bar{\Delta y}$ | $\text{var}(y)$ | Time(s) |
|--------------|-----------|-----------------|------------------|-----------------|---------|
| Without | -0.2101 | 0.0624 | 7.51 | 6.72 | 360 |
| 3-region | -0.0989 | 0.0401 | 7.72 | 7.23 | 0.0396 |
| 5-region | -0.0841 | 0.0362 | 7.68 | 6.79 | 0.0198 |
| 9-region | 0.0931 | 0.0258 | 7.71 | 6.74 | 0.0258 |

Fig. 29.9 The schematic of partitioning



5-region practice is put into effect. I make comparisons between algorithms based on secondary matching and density for matching precision and running time. Results are shown in Tables 29.3 and 29.4.

Table 29.3 The precision comparison of two methods (MM)

| Algorithm | \bar{x} | $\text{var}(x)$ |
|-----------------|-----------|-----------------|
| Secondary match | −0.2101 | 0.0624 |
| Density-based | −0.0841 | 0.0362 |

Table 29.4 Performance comparison of two methods (S)

| Performance index | Secondary match | | Density-based | |
|------------------------|-----------------|--------|---------------|----------|
| | Pairs | Time | Paris | Time |
| Extraction | 185 | 0.0144 | 60 | 0.01494 |
| Matching | 113 | 0.0015 | 16 | 0.000072 |
| Participate in compute | 113 | 360 | 16 | 0.0048 |

From Tables 29.3 and 29.4, we can reach to a conclusion that density-based algorithm enjoys a good property with high precision and good real-time performance when selecting feature points and rejecting false candidates.

References

1. Luo J, Tang X, Xu D (2011) Computer vision. University of Science and Technology of China Press, Hefei, pp 1–3

2. Sturm P (2011) A historical survey of geometric computer vision. In: 14th international conference on computer analysis of images and patterns, Grenoble, pp 1–8

3. Liu Y (2010) A survey of computer vision applied in aerial robotic vehicles. In: 2nd international conference on optics, photonics and energy engineering, Wuhan, pp 277–280

4. Frew E, McGee T, Kim Z, Xiao X, Jackson S et al (2004) Vision-based road- following using a small autonomous aircraft. In: IEEE aerospace conference, pp 3006–3015

5. Fanyan B (2010) Research on digital image matching. Hefei Industry University, Hefei

6. Zhou Y (2008) Research on image matching. Xidian University, Xian

7. Wang Q, Guan W, You S (2011) Augment distinctive feature for efficient image matching. In: IEEE workshop on application of computer vision, Kona, pp 15–22

8. Alhwarin F, Ristic Durrant D, Graser A (2010) Speeded up image matching using split and extended sift features. In: International conference on computer vision theory and applications, Angers, vol 5, pp 17–21

9. Grishin VA (2010) Two-channel algorithm of match making in computer vision systems. Sens Syst 65–68

10. Mortensen EN, Deng H, Shapiro L (2005) A SIFT descriptor with global context. In: IEEE computer society conference on computer vision and pattern recognition, pp 184–190

11. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. In: IEEE transactions on pattern analysis and machine intelligence, pp 1615–1630

12. Zhang J, Xiaojing B, Xu L (2009) A method of correcting SIFT mismatching based on spatial distribution descriptor. J Image Graphics 14(7):1369–1377

13. Baumberg A (2000) Reliable feature matching across widely separated views. In: IEEE conference on computer vision and pattern recognition, pp 774–781

14. Li R, Zeng B, Liou ML (1994) A new three-step search algorithm for block motion estimation. In: IEEE transactions on circuits and systems for video technology, pp 438–442

Chapter 30

Curriculum Design Change of the Industrial Engineering BA Program

Eszter Bogdány, Ágnes Balogh, Gabriella Cerháti,
Tibor Csizmadia and Réka Polák-Weldon

Abstract Higher education institutions are complex adaptive systems that need to respond to environmental pressures in a flexible way in today's dynamically changing environment. Resource dependency theory provides a framework to understand the process of adaption required to meet challenges. Based on the theory this paper introduces a case study presenting curriculum design change of Industrial Engineering BA program at the University of Pannonia, Hungary. The case study followed four steps of curriculum design change: reasons, design and development, teacher preparation and course design, and course evaluation. As a result of external and internal constraints the foreign language education and learning modules of the Industrial Engineering BA program had to be improved in 2009. After almost 3 years of offering a new foreign language introductory course the evaluation of the efficiency became necessary. Results show that learning module change has a positive effect on student activities of foreign language course enrolments and efficiency.

30.1 Introduction

This paper focuses on the question of how a higher education institution responds to governmental expectations, in the form of changes in curriculum design. Hungary as a medium-sized structurally open economy is increasingly exposed to impacts of global social and economic trends. So, foreign language education and learning is increasingly relevant in Hungary. It is a peculiar, hard to learn language, spoken by only a small number of people. Consequently, economic policy

E. Bogdány (✉) · Á. Balogh · G. Cerháti · T. Csizmadia · R. Polák-Weldon
University of Pannonia, Veszprém, Hungary
e-mail: wendywangyuan@163.com

making efforts of the present government aim at developing higher education programs with a special focus on language education and natural sciences such as Industrial Engineering.

The interesting point here is to what extent this situation has affected the course results and the number of courses students enrolled in. Thus, in an international perspective, the Hungarian case could provide an interesting contrast to those countries where policy-making concerning curriculum design change of university programs has been more comprehensive. As such it can be expected to add to the growing body of knowledge on how higher education institutions are reacting in response to contextually different external demands.

The paper is divided into three parts. The first part of the paper begins with a theoretical chapter that elaborates the theoretical concepts touched upon in this paper. The second part of the paper concentrates on the case study analysis. The final part of the study contains a summary of the paper and presents some further applications of the development process.

30.2 Theoretical Background

In this chapter the theoretical framework is discussed. The chapter provides an overview of the resource dependency theory that will be used in this paper.

30.2.1 Resource Dependency

The point of departure for discussion is to understand the way an organization responds to environmental pressures at the organizational level. The theoretical framework developed by Pfeffer and Salancik [1] serves this purpose emphasizing that to understand organizations one must understand how they relate to other actors in their environment. The resource dependency approach is constructed on the basis of the fundamental assumption that all organizational action is ultimately directed at securing its survival.

Some organizations might be more important to an organization than others with respect to resource acquisition. When the dependency is low, resistance represents minimal risk to organizational interests because it “is no longer held captive by a single or limited number of sources of social support, resources or legitimacy” [2]. Thus, in sum, the resource dependence theory implies that an organization’s responses to external requirements can be predicted from the situation of resource dependencies confronting it.

In addition, organizational response to demands does not necessarily mean passive adaptation, but rather a strategic choice to cope with external pressures [3]; dependency is not a simple one-way concept, it involves more. It gives the strategic repertory a focus [1]. It excludes the possibility that organizations contribute

consciously to their own demise. The action or strategy chosen depends on the motivations and preferences of organizational actors, the characteristics of the exchange relation and the structure of the network. In addition, organizations also focus on “altering the system of constraints and dependencies confronting the organization” [1]. Thus they retain the ability to have some flexibility in the responses to deal with particular issues.

Looking at the dependency relation from the governmental point of view, this opens the possibility of University of Pannonia to change in accordance with Hungarian Accreditation Committee priorities. As the aim of this study is to explain to what extent University of Pannonia indeed responds to governmental ‘demands’, the resource dependency is a crucial approach. Higher education institutions implement governmental initiatives in order to appear legitimate in the eyes of government agencies, which control vital resources [4].

30.2.2 Applications of the Framework

The Higher Education Act requires all Hungarian higher education institutions and their programs to be accredited regularly. In addition to institutional accreditation, the Hungarian Accreditation Committee (HAC) also conducts separate program accreditation under a variety of schemes required by law. Furthermore, degree programs are evaluated in the five-yearly institutional accreditation process. The HAC accredits national qualification requirements and all new programs launched at a university. The application for launching a degree program, in which there already are accredited national qualification requirements, focuses on the local context in which the proposed program will run, such as the teaching staff and infrastructure, as well as the curriculum.

All higher education institutions are expected to fulfill a set of requirements stated by HAC which can derive from the institution and program accreditation process. Specific recommendations have been made related to the improvement of the Industrial Engineering program which was completed by the HAC at the University of Pannonia in 2008. Regarding the improvement of foreign language courses the HAC pointed out that it is advisable to separate preparing students to language exams and the knowledge level of engineering and management disciplines in foreign languages. The evaluation of HAC did not imply directly applicable actions so the leadership had the freedom to implement strategic actions specific to the Industrial Engineering program. Regarding strategic actions the leadership considered the views of various stakeholders such as companies, students, lecturers and guest lecturers [5, 6].

30.3 Industrial Engineering Case Study

The case study explains and describes the reasons, process and results of the improvement of foreign language education and learning at the Industrial Engineering program. It allows investigating a phenomenon of interest within their broad context. It also provides the opportunity to appreciate these impacts and to explore to what extent the proposed concepts and theorized relationships are viable. Additionally, the case study facilitates the appraisal of some variables central to this inquiry that are not directly observable. In order to appraise these concepts, the complex social, organizational settings in which they emerge should be captured. The improvement process of foreign language education and learning activities followed the steps developed by Richards [7].

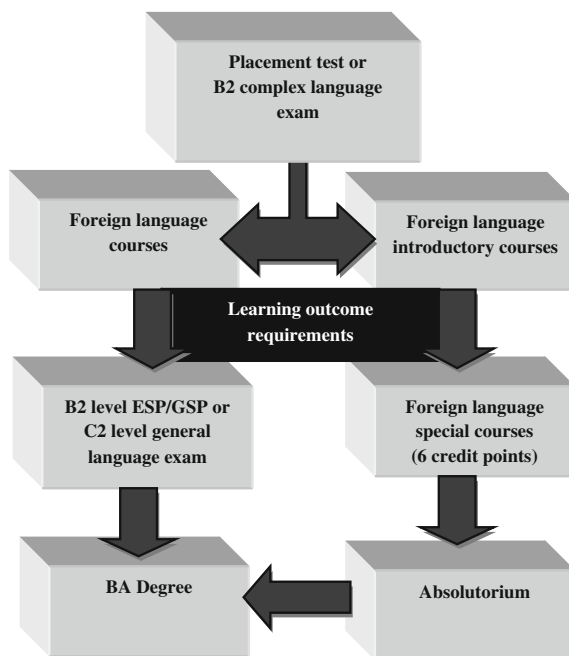
30.3.1 *Reasons*

The improvement of the Industrial Engineering program started with the outlining of the reasons behind the need for change. The main concerns regarding the improvement of curriculum design of foreign language courses were the followings:

- The HAC advised several foreign language course improvement approaches.
- Quality gap between the levels of language exam knowledge (English and German for Special Purposes hereafter ESP/GSP language exam) and the language skills required to complete special university courses—for a total of 6 credit points—offered in foreign languages.
- Often students passed ESP/GSP language exams after completing foreign language special modules.
- The initiation of the Bologna Process (also called Bologna Accords) indicated that the cycles of higher education qualification should be harmonized which means that BA level qualification has dual function—offering ESP/GSP language courses and foreign language special modules—and master level qualification has single function and that is offering only foreign language special modules.

As a result of the above explained propositions the structure of the university curriculum design framework regarding foreign language special courses had to be improved. Based on that a foreign language business introductory course was initiated which provides basic professional foreign language knowledge which can effectively foster the successful completion of foreign language special modules later in the course of studies.

Fig. 30.1 The steps of preparing students for fulfilling language and BA degree requirements



30.3.2 Design

To further the efficiency of foreign language competencies it was considered to be purposeful to separate foreign language exam preparation and teaching foreign language special modules. The foreign language courses are offered by the Centre of Foreign Language Education and the foreign language business introductory and special courses are offered by the Faculty of Economics.

The development process of foreign language competencies can be divided into two steps (Fig. 30.1):

- (1) preparing students for the B2 level ESP/GSP language exam:
 - a Students are expected to take a language placement test or have to have a B2 level complex language exam in order to participate in the foreign language course
 - b After either passing the placement test or proving the required language knowledge by a language exam certificate students are allowed to sign in for a two-hour foreign language course.
 - c By completing the foreign language course students are able to pass ESP/GSP language exam in order to fulfill the learning output requirement
- (2) In case of completing foreign language special courses, lectures and seminars students are expected to take the above mentioned language placement test or have to have a B2 level complex language exam in order to participate in the

foreign language introductory courses called ‘Comprehensive Business Studies I and II’(CBS I; CBS II) which includes relevant engineering and management topics. Upon passing CBS I and CBS II students fulfill the prerequisite of foreign language special courses for a total of 6 credit points.

30.3.3 Teacher Preparation and Course Development

The core element of the improvement process was to launch a new foreign language introductory course. The CBS I and II modules had to be fitted in the foreign language curriculum because by participating in the learning process of the introductory business modules student become able to succeed in the completion of the foreign language special modules.

In the beginning of 2009 resulting from the need for the improvement of curriculum design of the Industrial Engineering program the dean of the Faculty of Economics along with the Head of Management and Leadership Department initiated the addition of an introductory business module in two foreign languages, English and German to the existing program curriculum. The idea was thoroughly discussed with teachers of foreign language special modules in order to assure that the newly introduced module has some real advantages both for students and other university stakeholders. After representing the idea of the introduction of a new module those involved directly in the course design process formed a working group focusing on specific issues related to the new module. Two of the foreign language special module teachers were involved in the working group throughout the whole process of discussion, course design and implementation. The argument behind opting for those two teachers was that their qualifications and experience allowed them to effectively partake in the decision-making and design process. The working group had 6 months to design the foreign language introductory module. The Head of Management and Leadership Department acted as an advisor as well as a coordinator.

Firstly, all foreign language special module syllabi had been acquired for detailed study. The principles followed in the study process of the collected syllabi had to be agreed on. The members of the working group together with the advisor and coordinator outlined the principles as the following:

- group syllabi according to their field of study such as engineering and management,
- cluster syllabi based on similarity of their field of study,
- individual and group discussions with foreign language special module teachers regarding the focal points of their teaching material.

The created clusters indicated the main topics that were covered in the foreign language introductory course material and as the development process was finalized the teaching of the introductory course could begin. Teachers of the foreign language introductory course recognized the undeniable benefits of creating a student centered quality online course system. E-learning served and serves

effectively as specific media to implement the learning and teaching process for our students. It also provided platform for continuous course material development. The Moodle architecture (an acronym for Modular Object-Oriented Dynamic Learning Environment) had already been used at the university at the time and it became an effective mode of critical knowledge transfer. The lack of course specific textbook also indicated the need of adopting an effective student centered course management system. Additionally, the challenges of reaching large number of students and enabling a knowledge network of students could be met.

On the one hand, essential learning material is uploaded regularly and on the other hand teachers and students form real time study networks relying on the various Moodle functions to enhance learning results.

The main functions used by teachers and CBS I and II course participants are:

- forums serving as message boards allowing students to post messages and exchange ideas or problems related to course material,
- assignments for task completion and online feedback,
- questionnaires used to help students in their self-assessment process,
- news forum include all activities carried out by all course participants, teachers and students.

E-learning tools have been evaluated by one of the students as “useful because the CBS online course materials are available real-time and can be used as reference when preparing for foreign language special course exams”.

One of the lecturers teaching the CBS courses believe that since “the introductory courses are not language courses they specifically focus on professional language and provide competencies and self-confidence that allow students to perform well at foreign language special courses”.

A native guest lecturer added that “students’ communication abilities enhanced in terms of their involvement in debates, they became more confident in using the special terms of engineering and management fields”.

30.3.4 CBS Course Evaluation

From 2009 the foreign language education and learning at the Industrial Engineering program changed according to the above described process. After 2 years of teaching experience the effectiveness of the new foreign language introductory courses (CBS I and II) have been evaluated.

The positive effects of the new introductory course could be first seen in the academic year 2010/2011 (Fig. 30.2). Up to that point the average course enrolment was around 15 students per year. Following the completion of the introductory course students, relying more on knowledge acquired during their introductory course studies became more willing to participate in foreign language special courses.

Fig. 30.2 Total number of course enrolments

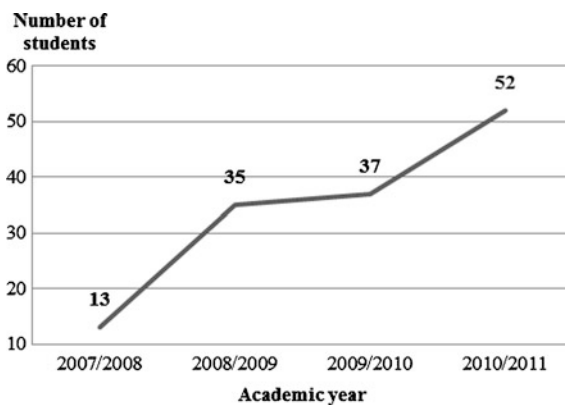
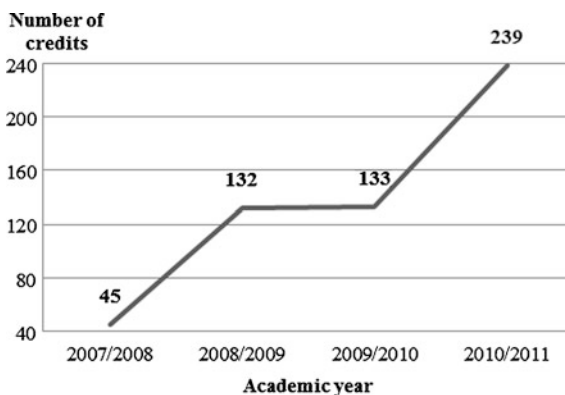


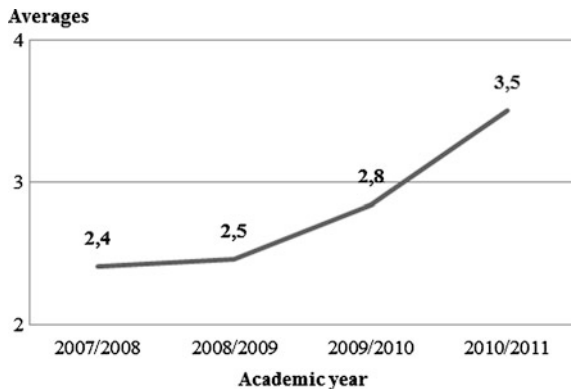
Fig. 30.3 Total number of credit points acquired



According to one of the students “CBS courses gave me the basic vocabulary of engineering in English that made me feel positive about taking courses of native guest lecturers”.

The growing number of course enrolments indicates that the total number of credit points acquired also increased from 2010 onwards (Fig. 30.3). The sharp increase in the number of credit points does not only suggest that more students enrolled in foreign language courses in general but also that the variety of courses students enrolled in widened. It means that students became more open to enroll in more difficult courses.

Despite the fact that the number of enrolments and credit points acquired increased the average of foreign language course results improved only slightly (Fig. 30.4). It shows that although the introductory course supports student preparation to undertake foreign language special courses there is room for improvement. The course results reached the average of 3.5 which is considerably higher than results prior to the introduction of the CBS courses.

Fig. 30.4 Average course results

30.4 Conclusion

The main purpose of the curriculum design change of the Industrial Engineering BA program was to improve the efficiency of foreign language course enrolment and results. As it was detailed in our paper the process itself began as a response to both external and internal pressures and heavily relied on resource dependency theory. The case study methodology allowed us to present the main advantages and disadvantages of the curriculum design development process which included four steps: reasoning the planned changes, designing, preparing teachers and developing course material and finally evaluating the newly introduced CBS course.

The thoroughly executed preparation and course design stage of the process lead to fast and flexible respond to environmental pressures. Moreover, continuous e-learning course material development enhances our ability to respond to future internal pressures. Considering the environmental constraints the strategic choice and action to change foreign language education and learning modules of the Industrial Engineering BA program have been considered effective by all stakeholders and can serve as an example to other higher education institutions facing similar challenges.

Acknowledgments This paper was made under the project TÁMOP-4.2.2/B-10/1-2010-0025. The project is supported by the European Union and co-financed by the European Social Fund.

References

1. Pfeffer J, Salancik GR (1978) The external control of organizations: a resource dependence perspective. Harper and Row, New York
2. Oliver C (1991) Strategic responses to institutional processes. *Acad Manag Rev* 16:145–179
3. Rhoades G (1992) Organization theory. In: Clark B, Neave G (Eds.), *The encyclopedia of higher education*, vol 2. Oxford Pergamon, pp 1884–1897

4. Maassen P, Gornitzka Å (1999) Integrating two theoretical perspectives on organisational adaptation. In: Jongbloed B, Maassen P, Neave G (eds) *From the eye of the storm: higher education's changing institution*. Kluwer Academic Publishers, Dordrecht
5. Gaál, Z Szabó L, Kovács Z (2007) Culture, competence, competitiveness. Managing diversity at individual and community Level. *Int J Divers Organ, Commun Nations* 7(5)131–141
6. Böcskei, E Vevői igények kielégítése (2009) A TQM és a controlling kapcsolata, A Controller, Ecovit Kft. 1:6–8
7. Long MH, Richards JC (1987) *Methodology in TESOL: a reader*. Rowley, Mass: Newbury House

Chapter 31

Comparing Two Methods of Sound Spatialization: Vector-Based Amplitude Panning (VBAP) Versus Linear Panning (LP)

Jonathan Cofino, Armando Barreto and Malek Adjouadi

Abstract There is an interest in presenting the sound from “Screen Reader” programs used by blind individuals to access the World Wide Web in multiple virtual “auditory columns”, aiming to restore the perception of 2-dimensional placement of items on contemporary web pages. As a prelude to that application, this paper reports the experimental comparison of two forms of virtual placement of sounds over 5 positions, in front of a computer user. The results indicate that there is not a statistically significant difference in accuracy achieved by application of the Vector-Based Amplitude Panning (VBAP) or the Linear Panning (LP) methods.

31.1 Introduction

For navigation of the World Wide Web (WWW), blind computer users rely on programs called “Screen Readers” to vocalize text (performing text-to-speech conversion of the contents of the web pages), identify forms, tables, and graphics, and guide them while browsing. Screen readers operate by identifying a focus (link, text box, menu item, etc.) and announcing relevant information about the element of focus [1]. Although this technology is invaluable to those who use it, navigation of a modern web site is often complicated by the complex two-dimensional layout of many web pages. When forced to linearize screen content, that is to output it as a single stream of sounds, screen readers must arbitrarily choose the path of content read aloud while losing a sense of spatial orientation within the web page. This usually means that the screen reader will follow the

J. Cofino (✉) · A. Barreto · M. Adjouadi
Electrical and Computer Engineering Department, Florida International University,
Miami, 33174 FL, USA
e-mail: bogdany.eszter@gtk.uni-pannon.hu

elements in the order that they have been coded in HTML without regard to how these elements may actually be arranged on the webpage.

This study proposes to create a finite number of “auditory columns” that can be distinguished by blind users in order to, eventually (in future work), build an audio-navigable web page environment that aurally preserves the spatial structure of the information contained in the web pages. Our focus in this study is on the presentation of information in 5 distinct “auditory columns” which would require the user to distinguish which of 5 locations was used for the virtual placement of a sound. This auditory virtual placement can be accomplished by a variety of methods, and this paper reports on the comparative study we performed, involving two of those methods.

Our experiment compares the vector-based amplitude panning (VBAP) method proposed by Pulkki [2] and the linear panning method. This study is a preliminary step in determining whether the method used for panning has a significant influence in restoring the sense of space and orientation to the browsing of a blind user.

31.2 Methods

A Japanese study by Ohuchi et al. [3] has demonstrated that the sound localization acuity of blind people is superior, on average, to that of sighted people. Those researchers found this result by placing a listening test subject in the center of a circular array of 12 (physical) speakers which were evenly spaced at 30-degree intervals along the horizontal plane. No amplitude panning was used, as each sound was actually played from one of 12 different physical locations (speakers) at a time.

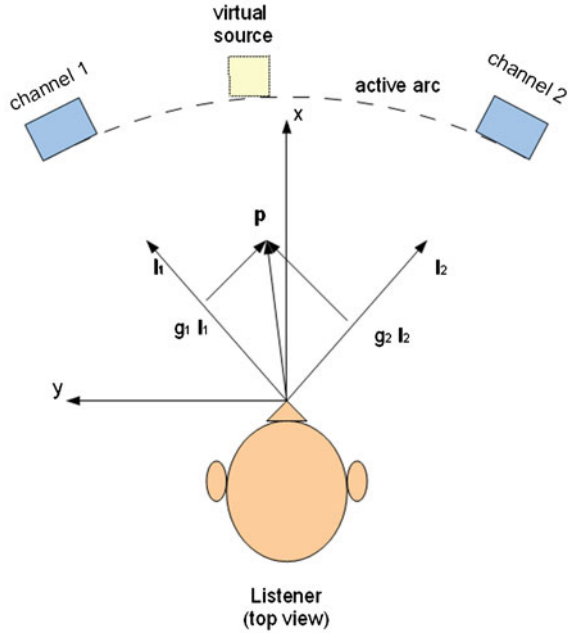
In a 1998 paper [4], Pullki and Lokki discussed using the VBAP panning method with a multichannel array of speakers to create an immersive three-dimensional auditory display. For our intended application, it would be impractical to require a typical computer user to sit at the center of a multichannel speaker array; therefore, our experiment utilized two stereo speakers since this setup best reflects a typical user’s audio playback situation. This necessitates stereo panning to represent the five auditory columns mentioned previously.

Sound spatialization can be achieved through stereo panning. Each of the two stereo speakers emits coherent signals with varying amplitudes. This difference in amplitude creates the perception of a single auditory source emanating from a virtual position located between the actual speakers [4]. By varying the respective amplitudes, the virtual source can be ‘positioned’ anywhere between the two speakers.

31.2.1 Stereophonic VBAP Method

In the VBAP method, two speakers will be used to create an “active arc” as shown in Fig. 31.1.

Fig. 31.1 Stereophonic configuration formulated with vectors (as proposed in [2])



Each speaker location can be expressed as a unit vector:

- $l_1 = [l_{11} l_{12}]^T$
- $l_2 = [l_{21} l_{22}]^T$

The vector corresponding to the virtual source, p , can be expressed as a linear combination of the speaker vectors l_1 and l_2 :

$$p = g_1 l_1 + g_2 l_2 \quad (31.1)$$

where g_1 and g_2 correspond to the gain factors of channels 1 and 2, respectively.

Equation (1) can now be expressed in a matrix form:

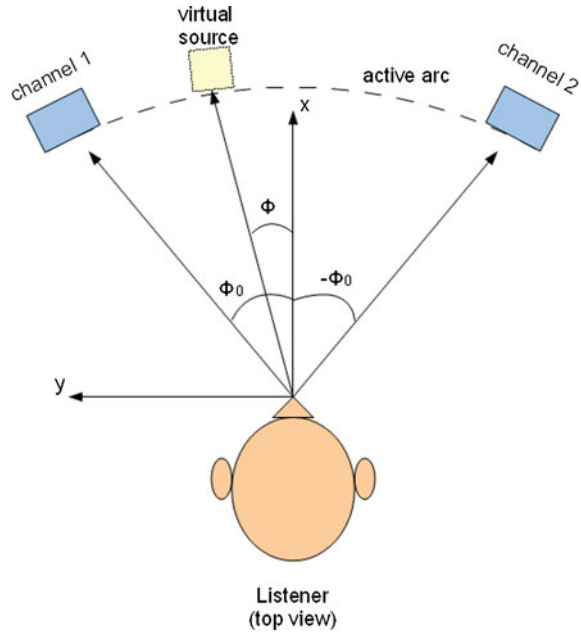
$$p^T = g L_{12} \quad (31.2)$$

$$\text{where } g = [g_1 g_2] \text{ and } L_{12} = [l_1 l_2]^T$$

An inverse for matrix L_{12} exists as long as the speakers are not placed collinearly, i.e. both on the horizontal or vertical axis. The gain vector can now be found as shown:

$$g = p^T L_{12}^{-1} = [p_1 p_2] \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix}^{-1} \quad (31.3)$$

Fig. 31.2 Stereophonic panning (angular formulation as proposed in [2])



To perceive all virtual sources as emanating from equidistant points, an active arc of constant radius is required. To achieve this constant sound power level, g is normalized to maintain:

$$g^{\text{scaled}} = \frac{g}{\sqrt{g_1^2 + g_2^2}} \quad (31.4)$$

To simplify the implementation of the VBAP method in two dimensions, an angular approach is desirable as shown:

In Fig. 31.2, the virtual source is represented by an angle Φ , which is constrained by the two channel speakers spaced at equiangular positions ($-\Phi_0, \Phi_0$) on either side of the x -axis.

In order to reconcile the angular approach with the need for channel-specific gain constants, Bauer [5] reformulated Blumlein's stereophonic law of sines in phasor form as shown:

$$\frac{\sin \Phi}{\sin \Phi_0} = \frac{g_1 - g_2}{g_1 + g_2} \quad (31.5)$$

where $0^\circ < \Phi_0 < 90^\circ$,

$-\Phi_0 \leq \Phi \leq \Phi_0$,

and $g_1, g_2 \in [0, 1]$.

The equations above are only valid for when a listener's head is held still and forwardly oriented along the x -axis.

As noted by Ohuchi et al. [3] and also by Pulkki and Karjalainen [6], if a listener's head is allowed to swivel, a stereophonic tangent law is more accurate:

$$\frac{\tan \Phi}{\tan \Phi_0} = \frac{g_1 - g_2}{g_1 + g_2} \quad (31.6)$$

where $0^\circ < \Phi_0 < 90^\circ$,

$-\Phi_0 \leq \Phi \leq \Phi_0$,

and $g_1, g_2 \in [0, 1]$.

It can be further shown that the gain vector \mathbf{g} satisfies the tangent law above. The two-channel stereophonic loudspeaker configuration matrix (L) components and virtual source position components can be observed from Fig. 31.2.

$$l_{11} = l_{12} = \cos \Phi_0$$

$$l_{12} = l_{22} = \sin \Phi_0$$

$$p_1 = \cos \Phi$$

$$p_2 = \sin \Phi$$

The inverse of the L_{12} matrix can be determined:

$$L_{12}^{-1} = \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix}^{-1} = \frac{\begin{bmatrix} l_{22} & -l_{12} \\ -l_{21} & l_{11} \end{bmatrix}}{l_{11}l_{22} - l_{21}l_{12}} \quad (31.7)$$

Equation (3) can now be reformulated:

$$\mathbf{g} = \frac{[p_1 l_{22} - p_2 l_{21} p_2 l_{11} - p_1 l_{12}]}{l_{11}l_{22} - l_{21}l_{12}} \quad (31.8)$$

$$g_1 = \frac{\cos \Phi \sin \Phi_0 + \sin \Phi \cos \Phi_0}{2 \cos \Phi_0 \sin \Phi_0} \quad (31.9)$$

$$g_2 = \frac{\cos \Phi \sin \Phi_0 - \sin \Phi \cos \Phi_0}{2 \cos \Phi_0 \sin \Phi_0} \quad (31.10)$$

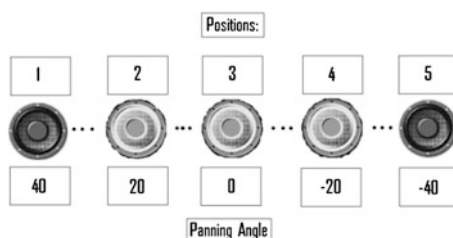
$$\frac{g_1 - g_2}{g_1 + g_2} = \frac{2 \sin \Phi \cos \Phi_0}{2 \cos \Phi \sin \Phi_0} = \frac{\tan \Phi}{\tan \Phi_0} \quad (31.11)$$

It is now apparent that the VBAP method satisfies the stereophonic tangent law.

The five VBAP positions used in the experiment as well as the plotted gain values used to implement them are pictured in Figs. 31.3 and 31.4.

It can be noted from the previous figure that the VBAP method produces a concave pattern of speaker gains to implement panning.

Fig. 31.3 VBAP panning angles



31.2.2 Linear Panning Method

To assess the benefits of the VBAP method, a simple linear panning approach will be compared. Linear panning does not account for a constant sound power level nor does it take into consideration the perceived angle of localization.

The graph in Fig. 31.6 was generated by the linear equations:

- $\text{left_gain} = (\frac{1}{2})(1 - \text{pan})$
- $\text{right_gain} = (\frac{1}{2})(1 + \text{pan})$

where the parameter ‘pan’ corresponds to the desired horizontal panning position.

The five linear panning positions used in the experiment as well as the plotted gain values used to implement them are pictured in Figs. 31.5 and 31.6.

31.2.3 Experiment Design Methods

The objective of our experiment was to compare the level of accuracy with which a typical computer user can identify the virtual placement of a sound (out of 5 possible locations) when VBAP and linear panning are used.

This experiment prompted each of the 20 (sighted) participants to locate a virtual sound source in auditory space. Borrowing from the methods used by Ohuchi et al. [3], the sound sources consisted of white noise and lasted 2 s in duration. White noise was chosen for its flat power spectral density. Having equal power across any given bandwidth, white noise is not subject to the directionality commonly associated with single-frequency tones. Each stereophonic panning method was used twice for each one of the five virtual positions, yielding 20 trials of sound playback that the subject needed to identify as emerging from one of the 5 pre-set virtual source locations. The order of these virtual source locations and the spatialization method used were randomized for each participant.

In each trial, the subject was given the indication to press the “enter” key to make the trial sound play and then asked to enter a number: 1, 2, 3, 4, or 5 to identify the spatial location that was perceived by him/her as the origin of the sound. Since the intended spatialized locations were determined in advance, each

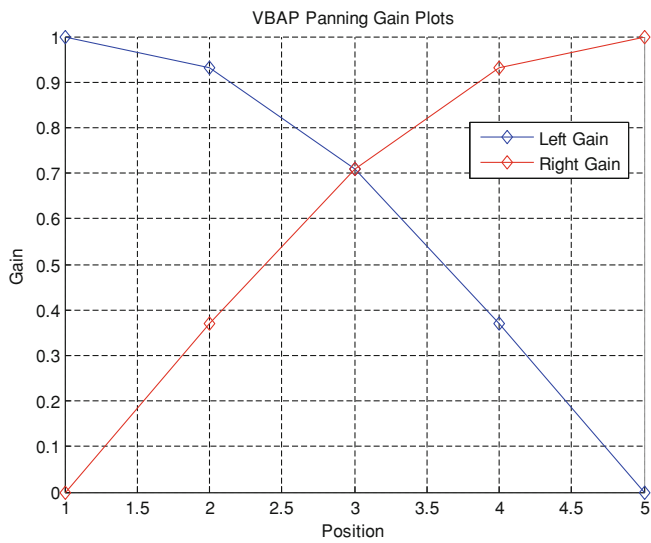
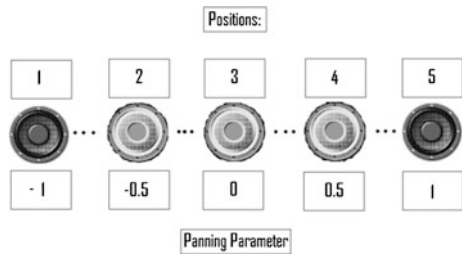


Fig. 31.4 VBAP panning (5 positions)

Fig. 31.5 Linear panning parameters



trial was scored as correct (if the subject’s numerical answer exactly matched the numeric identifier of the virtual sound placement) or incorrect (in any other case). Accordingly, a percentage of accuracy could be derived for each subject: when the spatialization was performed with VBAP, when the spatialization was performed with linear panning, and also a percentage of overall accuracy (using the results obtained under both spatialization methods).

As can be seen in Figs. 31.1 and 31.2, the listening subject and the two speakers form an isosceles triangle. Each speaker was placed 30 inches from the subject’s chest making the lateral distance between the speakers 38.57 inches. This implies that using the linear panning technique the listening subject should perceive that the five virtual sound sources are placed at 9.64-inch intervals. It was important to find a compromise between sufficient angles, distances, and space limitations (such as the need for the two speakers to fit on a typical desk). Pulkki [2] states that in two-dimensional amplitude panning the Φ_0 angle typically chosen is 30° . In this experiment, we extended the horizontal distances to make the sources more

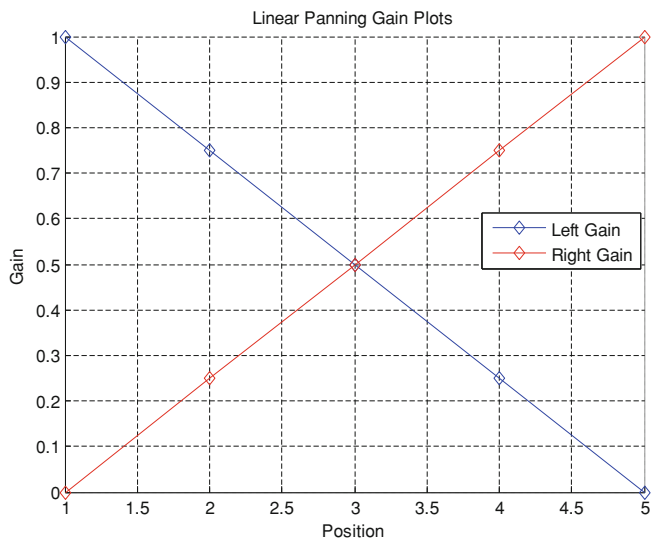


Fig. 31.6 Linear panning (5 positions)

distinct in the linear panning method. This was accomplished by increasing the Φ_0 angle from 30 to 40°.

31.3 Results

| Localization accuracy percentages table | | | | |
|---|---------|------|--------|------|
| Subject # | Overall | VBAP | Linear | Diff |
| 1 | 90 | 90 | 90 | 0 |
| 2 | 70 | 50 | 90 | −40 |
| 3 | 95 | 100 | 90 | 10 |
| 4 | 70 | 70 | 70 | 0 |
| 5 | 95 | 100 | 90 | 10 |
| 6 | 100 | 100 | 100 | 0 |
| 7 | 85 | 90 | 80 | 10 |
| 8 | 75 | 90 | 60 | 30 |
| 9 | 95 | 90 | 100 | −10 |
| 10 | 95 | 100 | 90 | 10 |
| 11 | 75 | 60 | 90 | −30 |
| 12 | 60 | 70 | 50 | 20 |
| 13 | 85 | 80 | 90 | −10 |
| 14 | 95 | 100 | 90 | 10 |
| 15 | 80 | 60 | 100 | −40 |

(continued)

(continued)

| Subject # | Overall | VBAP | Linear | Diff |
|-----------|---------|--------|--------|--------|
| 16 | 100 | 100 | 100 | 0 |
| 17 | 95 | 100 | 90 | 10 |
| 18 | 100 | 100 | 100 | 0 |
| 19 | 95 | 100 | 90 | 10 |
| 20 | 65 | 50 | 80 | -30 |
| Mean | 86.000 | 85.000 | 87.000 | -2.000 |
| St. Dev. | 12.732 | 18.209 | 13.416 | 19.358 |

31.4 Statistical Analysis

Statistical hypothesis testing was utilized for interpreting the data. The “null hypothesis” was stated: the mean accuracy under the two methods is assumed to be equal. An appropriate statistical test must be chosen to ascertain the validity of the null hypothesis. A paired-samples dependent *t* test is desirable for its simplicity and calculability, but it requires a normal distribution.

The Kolmogorov–Smirnov (K–S) and Shapiro–Wilk (S–W) tests are used to determine whether the difference between the samples is normally distributed. A significance value less than the typically chosen 0.05 indicates that a given set of sample differences significantly deviates from a normal distribution. Using the output from the PASW 18 statistical analysis software package, the K–S test indicates that the significance is much lower than the threshold ($0.003 < 0.05$) and the S–W test displays a similar result ($0.018 < 0.05$). Both tests indicate a significant deviation from a normal distribution.

Tests of normality

| | Kolmogorov–Smirnov ^a | | | Shapiro–Wilk | | |
|------------|---------------------------------|----|-------|--------------|----|-------|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| Difference | 0.241 | 20 | 0.003 | 0.881 | 20 | 0.018 |

^a Lillifors significance correction

Since the sampled differences are non-normally distributed, the Wilcoxon signed-rank test will be used as it is non-parametric, meaning that it does not assume normality as a prerequisite for hypothesis testing. To summarize, this test gathers the sign and magnitude of the paired differences. The magnitudes of the differences are then ranked from least to greatest, excluding the differences of zero. Tied ranks are averaged, i.e., if the difference value ‘7’ is repeated as the 3rd, 4 and 5th ordered difference value then the value ‘7’ will have a repeated rank of

$(3 + 4 + 5)/3 = 4$. The positive and negative differences are summed separately, and the lesser sum is declared the test statistic W . This test statistic can then be correlated with a significance value.

Using PASW 18, the output of the Wilcoxon signed-ranks test (Z-statistic) is shown:

Wilcoxon Signed of Rank Test

| Ranks | | N | Mean rank | Sum of ranks |
|--|----------------|----------------|-----------|---------------------|
| VBAP-Linear | Negative ranks | 6 ^a | 10.50 | 63.00 |
| | Positive ranks | 9 ^b | 6.33 | 57.00 |
| | Ties | 5 ^c | | |
| | Total | 20 | | |
| <div><div>^a VBAP < Linear</div><div>^b VBAP > Linear</div><div>^c VBAP = Linear</div></div> | | | | |
| Test statistics ^b | | | | VBAP-linear |
| Z | | | | −0.175 ^a |
| Asymp. Sig. (2-tailed) | | | | 0.861 |

^a Based on positive ranks

As shown above, the Wilcoxon Test resulted in a significance value of 0.861, much greater than the typical threshold 0.05. This result indicates that the null hypothesis should not be rejected. In practical terms, this indicates that there is not a statistically significant difference in the localization performance of our experimental subjects when the spatialization was performed using the VBAP method or the linear panning method, for this experimental configuration.

31.5 Discussion

In this preliminary experiment only 20 listening subjects were recruited (mean age = 26.9 years, std. dev. = 6.215 years.). All of these 20 subjects were sighted individuals with normal hearing. Using the findings from Ohuchi et al. it can be hypothesized that blind and low-vision individuals would outperform their sighted counterparts in terms of localization accuracy. This may also lead to a reduced average differential between methods.

Some test subjects noted that they could perceive a reduced intensity of some sounds. This is most likely a result of the lack of gain coefficient normalization inherent in the linear panning method. There may also be a psychoacoustic phenomenon of perceived unequal loudness with respect to virtual source placement and localization.

As this was only a preliminary study, certain experiment variables could be altered to benefit future experiments. The angle of placement for the speakers could be widened or narrowed, and the number of positions could be expanded. It is possible that increasing the number of subjects as well as widening the range of ages would yield more normally distributed data, making the paired-samples *t* test a possibility for statistical analysis.

31.6 Conclusion

The localization accuracy table shows that the mean accuracy percentage for both methods was quite close. This observation can be summarized by noting that the average difference of accuracy was only 2 percentage points. The VBAP method was found to have a wider dispersion about the mean, as indicated by the greater standard deviation as compared to the linear panning method. In general, the overall average of 86 % indicates that the listening subjects were able to correctly localize the audio in the experiment.

Based on the statistical analysis presented, this preliminary study suggests that there is not a statistically significant difference between the VBAP and linear panning methods for a 5-position array of sound sources like the one tested here. Most application programming interfaces (APIs) for audio implement the linear panning method. Therefore, this result justifies the prospective use of the linear panning method to create five “virtual auditory columns” for the enhancement of screen reader technology to help blind users navigate contemporary web pages.

Acknowledgments This work was sponsored by NSF grants HRD-0833093, and CNS-0959985.

We wish to thank all of the listening subjects who volunteered their time towards this study. All subjects were recruited during the 2011 Fall semester at the Florida International University Engineering Center, located at 10555 West Flagler Street, Miami FL 33174.

References

1. Hersh MA, Johnson MA (2008) Screen readers, Assistive technology for visually impaired and blind people. Springer, London
2. Pulkki V (1997) Virtual sound source positioning using vector base amplitude panning. *J Audio Eng Soc* 45(6):456–466
3. Ohuchi M, Iwaya Y, Suzuki Y, Muneakata T (2006) A comparative study of sound localization acuity of congenital blind and sighted people. *Acoust Sci Technol* 27(5):290–293

4. Pulkki V, Lokki T (1998) Creating auditory displays with multiple loudspeakers using VBAP: a case study with DIVA project. In: Proceedings of the international conference on auditory display (ICAD'98), Glasgow, 1–4 Nov 1998
5. Bauer BB (1961) Phasor analysis of some stereophonic phenomena. *J Acoust Soc Am* 33:1536–1539
6. Pulkki V, Matti K (2008) Multichannel audio rendering using amplitude panning [DSP applications]. *IEEE Signal Proc Mag* 25(3):118–122. Print

Chapter 32

Contrast Enhancement in Image Pre-Compensation for Computer Users with Visual Aberrations

Jian Huang, Armando Barreto, Malek Adjouadi and Miguel Alonso

Abstract Pre-compensation of display images in computers can enhance the interaction of computer users with visual impairments and computers. However, the pre-compensation process reduces the contrast of the images perceived by the user, when the pre-compensated images are presented in display devices with limited intensity levels. Therefore, the helpfulness of the pre-compensation process is reduced by the accompanying contrast loss. This paper proposes a side-trim histogram correction method aiming to improve the contrast of the image perceived by the user after pre-compensation has been performed. The side-trim method is based on the analysis of the histogram of the pre-compensated image, in which the trimming process is performed automatically, without manual intervention.

32.1 Introduction

Many computer users have difficulties identifying and recognizing pictures, icons and text displayed on computer screens, which are the basic forms of information presentation and interaction used by many software products, within the graphical users interface (GUI) utilized in most computers. This is more obvious for those computer users with severe low vision. Without the necessary vision correction,

J. Huang (✉) · A. Barreto · M. Adjouadi
Department of Electrical and Computer Engineering, Florida International University,
Miami, 33174 FL, USA
e-mail: jcofi001@fiu.edu

M. Alonso
School of Computer and Engineering Technology,
Miami-Dade College, Miami, 33176 FL, USA

effective interaction with computers for these users could be quite limited. Traditionally, vision correction is mainly achieved by spectacles and contact lenses. Laser vision correction, such as LASIK and LASEK, has also become popular recently.

On the other hand, methods based on image processing have also been used to help computer users with low vision to achieve better visual performance. Peli et al. [1, 2] proposed a conceptual pre-emphasis model of image enhancement for the visually impaired based on the contrast sensitivity function. Peli et al. [3] also used a wide-band enhancement method to enhance the television images for the people with visual impairments. Alonso et al. [4–6] proposed a pre-compensation method to improve the intended display image for computer users with impaired vision. In contrast with the traditional optical correction through lenses which modify the images of external objects before reaching the eye, this pre-compensation approach modifies computer images at their source, before displaying them to the users. This pre-compensation method is based on the a priori knowledge of the visual aberration of the user's eye, which can be measured by a wavefront analyzer. The pre-compensation method has been verified to be able to help computer users with visual impairments under controlled experimental conditions. However, the implementation of the pre-compensation method results in contrast loss that limits its benefits. This paper describes the essential reason for the contrast loss generated during the process of pre-compensation and proposes a method to reduce this undesirable side effect. Although the contrast enhancement process also comes with a degradation of the pre-compensation effect, the side-trim method in this paper is able to increase the contrast of pre-compensated images while keeping the degradation under control.

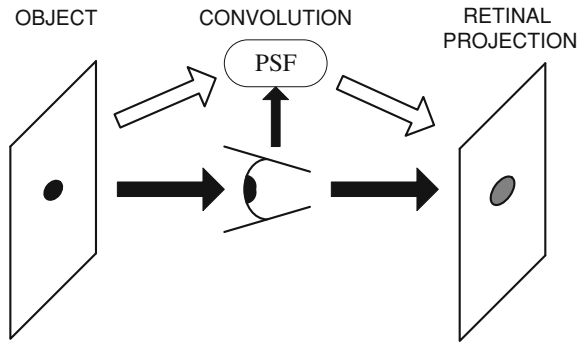
32.2 Background

The human visual system is considered as the combination of the optical system of the eye and the neural processing of the light information that begins at the retina. Like other optical systems, the imaging process in the human eye is described as the linear mapping of the light distribution of external objects to the light distribution on the retina, which is the light sensitive portion of the human eye. This makes it possible to apply the theory of linear systems to the imaging system in the eye.

32.2.1 Visual Aberration

Visual degradation of the human eye is caused by visual aberration in the eye's optical system. The eye's visual aberration can be modeled and quantified by the wavefront aberration function (WAF). The wavefront aberration function, used to describe the refraction characteristic of the eye, is defined as the difference between the actual aberrated wavefront and the ideal spherical wavefront of light coming into the eye [7]. Any wavefront aberration will degrade the resulting retinal images.

Fig. 32.1 Imaging of the human eye as a convolution process



32.2.2 Point Spread Function

Any object, when viewed by any optical system, including the human eye, can be considered as a two-dimensional array of point sources with variable intensity [7]. The image mapped by the optical system for a point source is called the point spread function (PSF), which is analogous to the two-dimensional impulse response function of the imaging system. Therefore, the process of forming a retinal image can be represented by the convolution of the distribution of light intensities in the external object being viewed and the PSF of the human eye, as shown in Fig. 32.1.

32.3 Pre-Compensation Method

When pictures, text or icons are displayed on a computer screen and viewed by the eye of a user with visual aberrations, the image formed on the retina will be degraded. The pre-compensation method introduced here processes the images before they are displayed, compensating, in advance, for the degradation that will take place in the user's eye, due to the eye's visual aberration. This process is illustrated in Fig. 32.2.

The pre-compensation model, which generates a suitable pre-compensated image for any given original image (object), is built based on the a priori knowledge of the visual aberration of the computer user's eye, which could be measured by a wavefront analyzer. The imaging system of the human eye is considered as a linear system. Therefore, if the pre-compensation is correctly set, it should, ideally, counteract the image degradation introduced by the eye. Suppose an image with intensity $o(x,y)$ is displayed on screen to be viewed by computer users. Let us say the image is degraded due to the aberrations of the user's eye, resulting in a blurred image on the user's retina with intensity $i(x,y)$. The PSF of the user's eye is derived as $psf(x,y)$ based on the wavefront aberration function

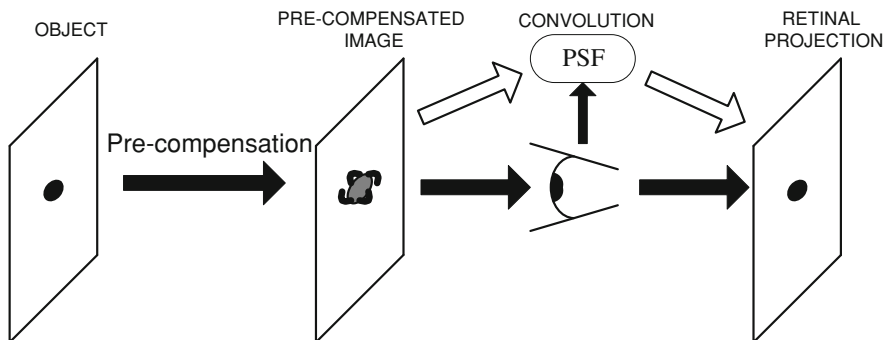


Fig. 32.2 Diagram of the pre-compensation process

measured in a wavefront analyzer. As mentioned before, the distortion can be described as a convolution process:

$$i(x, y) = o(x, y) * psf(x, y) \quad (32.1)$$

The optical transfer function (OTF), another useful function for describing visual performance, is calculated by applying the Fourier Transform to the PSF as:

$$OTF(u, v) = \mathcal{F}\{psf(x, y)\} \quad (32.2)$$

And the modulation transfer function (MTF) could be derived as the modulus or absolute value of the complex-valued OTF:

$$MTF(u, v) = |OTF(u, v)| \quad (32.3)$$

Therefore, in order to remove the degradation introduced by the PSF, the Fourier transform of the pre-compensated image, $C(u, v)$, could be calculated as:

$$C(u, v) = \frac{O(u, v)}{OTF(u, v)} \quad (32.4)$$

Accordingly, the pre-compensated image to be displayed on the screen $c(x, y)$ can be obtained through inverse Fourier Transform:

$$c(x, y) = \mathcal{F}^{-1} \left\{ \frac{O(u, v)}{OTF(u, v)} \right\} \quad (32.5)$$

However, Eq. (32.5) is not practical since errors in the measurement of the wavefront aberration function will be greatly amplified when the value of $OTF(u, v)$ is close to zero. This is more significant for the low frequency components as the low frequency errors will pass through the filtering implemented by the PSF in the imaging process of eye. Thus, the Wiener filter is used to solve this problem as:

$$C(u, v) = \frac{O(u, v)}{OTF(u, v)} \frac{MTF(u, v)^2}{MTF(u, v)^2 + K} \quad (32.6)$$

and $c(x, y)$ could be calculated by inverse Fourier transform:

$$c(x, y) = \mathcal{F}^{-1}\{C(u, v)\} \quad (32.7)$$

In (32.6), K is the regularization parameter that limits the amplification of unknown noise components. Therefore, the intensity of pre-compensated image could be generated based on (32.6) and (32.7).

32.4 Contrast Loss

Simply speaking, the PSF of human eye behaves primarily as a low pass filter, which is also the reason why human eyes have limited ability to perceive high frequency information [8]. From (32.6), it is not difficult to infer that the pre-compensation process has characteristics of high pass filter. A low pass filter allows the perception of flat areas in the images viewed, while inhibiting the high frequency components. The pre-compensation process reduces the distortion caused by the eye's PSF, but it generates a pre-compensated image $c(x, y)$ that has a wider range than the original image, possibly even involving negative values. On the other hand, display devices (e.g., an LCD) have limited intensity scales. Therefore, $c(x, y)$ needs to be shifted and scaled before display. Suppose the range of $c(x, y)$ is $[c_{\min}, c_{\max}]$ and the intensity range of display device is $[d_{\min}, d_{\max}]$ the shifted and scaled image is given by:

$$d(x, y) = \frac{c(x, y)(d_{\max} - d_{\min})}{c_{\max} - c_{\min}} - \frac{c_{\min}(d_{\max} - d_{\min})}{c_{\max} - c_{\min}} + d_{\min} \quad (32.8)$$

Thus, the Fourier transform of $d(x, y)$ is given by:

$$D(u, v) = \alpha C(u, v) + \beta \delta(u, v) \quad (32.9)$$

in which δ is the Dirac delta function, α is defined as:

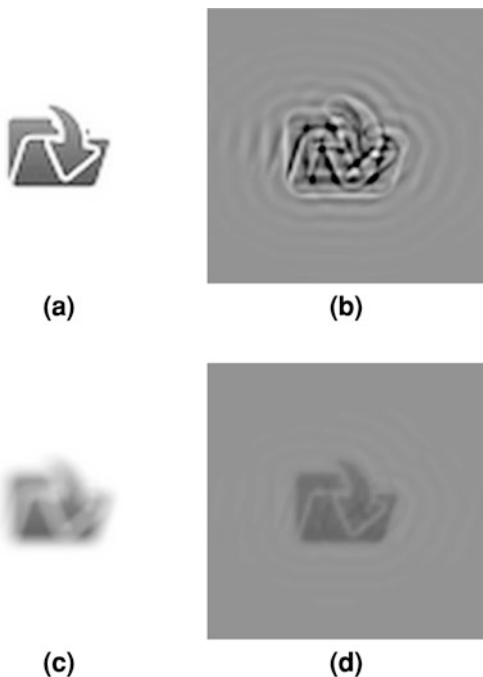
$$\alpha = \frac{d_{\max} - d_{\min}}{c_{\max} - c_{\min}}, \quad (32.10)$$

and β is defined as:

$$\beta = d_{\min} - \frac{c_{\min}(d_{\max} - d_{\min})}{c_{\max} - c_{\min}} \quad (32.11)$$

Therefore, the Fourier transform of the image expected to form on the retina after pre-compensation could be predicted as:

Fig. 32.3 **a** Original image intended for display. **b** Pre-compensated image; **c** Simulation of image perceived when viewing panel (a); **d** Simulation of image perceived when viewing panel (b)



$$P(u, v) = \alpha C(u, v)OTF(u, v) + \beta \delta(u, v)OTF(u, v) \quad (32.12)$$

and $p(x, y)$ could be derived by inverse Fourier transform of $C(u, v)$. Since practically the range $[c_{\min}, c_{\max}]$ is always wider than $[d_{\min}, d_{\max}]$, α is smaller than 1. From (32.12), we can find that the image displayed loses part of its contrast during the scaling process. The smaller α is, the more contrast is lost.

The contrast loss can be appreciated clearly in the simulation results show in Fig. 32.3. In this case, one icon (open file folder), which is used frequently in GUI design, is selected as the source image displayed on the computer screen. It is shown in Fig. 32.3a. A human eye with -4D defocus aberration is selected to view the icon in the simulation. Without pre-compensation, the image that will be perceived is degraded as shown in Fig. 32.3c, due to the visual aberration of the eye. In order to remove or reduce the degradation, pre-compensation is performed on the image of Fig. 32.3a before display, resulting in the pre-compensated image shown in Fig. 32.3b. The simulation of the retinal image that the user will perceive when viewing Fig. 32.3b is shown in Fig. 32.3d. Its icon shape is sharper than Fig. 32.3c. However, the contrast of Fig. 32.3d is evidently lower than that of Fig. 32.3a. This is consistent with the mathematical derivation above. Note that the image shown in Fig. 32.3a has extreme high contrast since its background is white. This makes the relative contrast reduction after pre-compensation even larger. In a

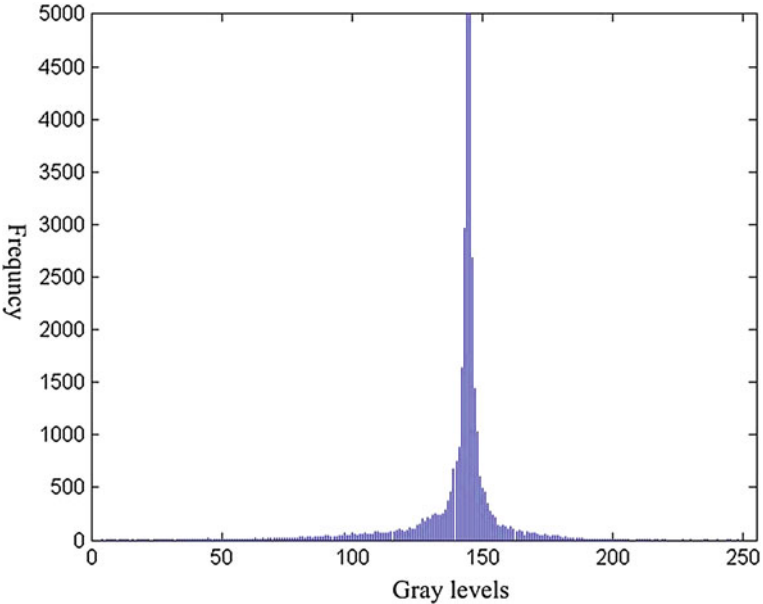


Fig. 32.4 Histogram of the pre-compensated image shown in Fig. 32.3b. Note that most of the intensities concentrate on the center band

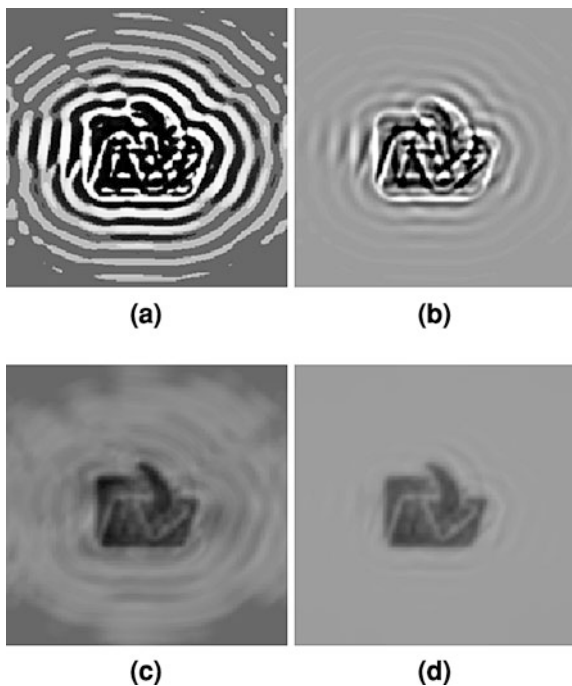
practical scenario, the pictures or icons usually do not have such high contrast because extreme high contrast makes computer users feel uncomfortable.

32.5 Methods of Contrast Improvement

The simulation results show that the pre-compensation method improves the visual performance of the subject by providing a clearer perceived image. However, the contrast, also an important factor influencing the computer user’s visual quality, is reduced in the pre-compensation process. It would be desirable to develop a method of increasing the perceived contrast to overcome this limitation. It is not difficult to note that most intensity values in the pre-compensated image, shown as Fig. 32.3b, concentrate on the narrow band near the value of the background. This is more evident from observation of its histogram, shown in Fig. 32.4, which shows a shape that is typical for images after pre-compensation.

In general, histogram equalization is an effective way to increase contrast in this kind of scenario. However, the histogram equalization is not linear since it distributes the packed intensities based on the cumulative distribution function (CDF). Therefore, the shape of the histogram of pre-compensated images could not be maintained, which may distort the effect of pre-compensation. In fact, the contrast enhancement process is somewhat opposite to the inhibition of low

Fig. 32.5 **a** Pre-compensated image after contrast improvement by histogram equalization; **b** Pre-compensated image after contrast improvement by side-trim method; **c** Simulation of image perceived when viewing panel (a); **d** Simulation of image perceived when viewing panel (b)



frequency degradation. With the method we propose we seek to enhance the contrast of the perceived image, while keeping the degradation of the pre-compensation process at an acceptable level. From observation of Fig. 32.4, there are few pixels located towards the two ends of the total range. Note that, the effect of pre-compensation is mainly determined by the intensities around the center band. If appropriate cutoff intensity levels are selected to trim the original histogram, the contrast of the pre-compensated image could be enhanced. In this side-trim method, the distortion of the pre-compensation model is isolated from the critical intensity levels in the center band of the histogram, and the basic shape of the histogram is kept. Thus, the pre-compensated image is transformed as:

$$c'(x,y) = \begin{cases} d_{\min}, & c(x,y) < c_L \\ \frac{[c(x,y)-c_L][d_{\max}-d_{\min}]}{c_H-c_L} + d_{\min}, & c_L \leq c(x,y) \leq c_H \\ d_{\max}, & c(x,y) > c_H \end{cases} \quad (32.13)$$

In (32.13), c_L denotes the low side cutoff intensity and c_H denotes the high side cutoff intensity. $c'(x,y)$ is the pre-compensated image after transformation. In practice, c_L and c_H can be set based on the statistical information of the histogram. For example, we could traverse the intensity levels of the histogram from the peak of highest frequency, to the left, to find the first intensity level with $<0.1\%$ of frequency and set this value to be c_L . Correspondingly, we could traverse right from the peak of highest frequency to the first intensity level with $<0.1\%$ of the

frequency and set this value to be c_H . This threshold definition process can easily be automated.

After applying this side-trim contrast enhancement method to it, the pre-compensated image to be displayed is as shown in Fig. 32.5b. Simulating the viewing of this image by the subject's eye yields the image in Fig. 32.5d. This simulated perceived image clearly exhibits better contrast than Fig. 32.3d. For an additional comparison, Fig. 32.5a presents the pre-compensated image after performing standard histogram equalization. The corresponding simulated perceived image is shown in Fig. 32.5c. The histogram equalization method also yields a perceived image with higher contrast, but it clearly aggravates the severity of the ringing effect introduced by the pre-compensation process, which is not the case for the side-trim approach. The icon shape in Fig. 32.5d also suffers from some distortion, but it is kept at a tolerable level.

32.6 Conclusion

In this paper, the contrast limitation of vision correction method for computer users by image pre-compensation is analyzed in detail. The contrast loss of image perception after pre-compensation is caused by intensity extension during inverse filtering and limited intensity levels of display devices. Overall, it seems that contrast improvement could only be achieved in a tradeoff with the introduction of distortion to the pre-compensation effect. This paper proposed a side-trim method to raise the contrast of the pre-compensated image while keeping the distortion to the pre-compensation effect at an acceptable level. While the results shown here are exclusively based on simulations, the method is suitable for application to real pre-compensation performed to facilitate computer access by users with visual impairment.

Acknowledgments This work was sponsored by NSF grants HRD-0833093, CNS-0959985 and CNS-0940575. Jian Huang is the recipient of Presidential Fellowship at Florida International University.

References

1. Peli E, Goldstein RB, Young GM, Trempe CL, Buzne SM (1991) Image enhancement for the visually impaired: Simulations and experimental results. *Invest Ophthalmol Vis Sci* 32:2337–2350
2. Fine EM, Peli E (1995) Enhancement of text for the visually impaired. *J Opt Soc Am A* 12:1439–1447
3. Peli E, Kim J, Yitzhaky Y, Goldstein RB, Woods RL (2004) Wideband enhancement of television images for people with visual impairments. *J Opt Soc Am A* 21:937–950
4. Alonso M, Barreto A, Jacko JA, Adjouadi M (2007) Evaluation of onscreen precompensation algorithms for computer users with visual aberrations, In: *Proceedings of the ASSETS 2007*,

- the 9th international ACM SIGACCESS conference on computers and accessibility, Tempe, Arizona, pp 219–220
5. Alonso M, Barreto A, Cremades JG, Jacko JA, Adjouadi M (2005) Image Pre-compensation to facilitate computer access for users with refractive errors. *Behaviour Inf Technol* 24(3):161–173
 6. Alonso M, Barreto A, Adjouadi M, Jacko JA (2006) HOWARD: high-order wavefront aberration regularized deconvolution for enhancing graphic displays for visually impaired computer users. *Lecture notes in computer science*, vol. LNCS 4061, pp 1163–1170
 7. Thibos LN (2000) Formation and sampling of the retinal image. In: Valois KKD (ed) *Seeing: handbook of perception and cognition*, 2nd edn. Academic Press, San Diego, pp 1–54
 8. Dai G (2008) *Wavefront optics for vision correction*. SPIE, Washington

Chapter 33

Interaction with 3D Environments Using Multi-Touch Screens

Francisco Ortego, Naphtali Rishe, Armando Barreto
and Melek Adjouadi

Abstract The increase in availability of multi-touch devices has motivated us to consider interaction approaches outside the limitations associated with the use of a mouse. The problem that we try to solve is how to interact in a 3D world using a 2D surface multi-touch display. Before showing our proposed solution, we briefly review previous work in related fields that provided a framework for the development of our approach. Finally, we propose a set of multi-touch gestures and outline an experiment design for the evaluation of these forms of interaction.

33.1 Introduction

This paper presents the initial development of our approach to work with 3D data environments using a multi-touch display. We introduce our emerging methods in the context of important previous work, resulting in our proposed gesture recognition approach and definition of translation and rotation gestures for multi-touch

F. Ortego (✉) · N. Rishe
School of Computing and Information Sciences, Florida International University,
Miami, FL, USA
e-mail: Forte007@fiu.edu

N. Rishe
e-mail: NDR@acm.org

A. Barreto · M. Adjouadi
Electrical and Computer Engineering Department, Florida International University,
Miami, FL, USA
e-mail: BarretoA@fiu.edu

M. Adjouadi
e-mail: Adjouadi@fiu.edu

interaction with 3D worlds. In this same process, we try to answer two important questions and provide an evaluation path to be implemented in the near future.

3D navigation and manipulation are not new problems in Human–Computer Interaction (e.g., [1, 2]) as will be shown in our brief review of previous work. However, with the availability of multi-touch devices such as the iPad, iPhone and desktop multi-touch monitors (e.g., 3M M2256PW 22" Multi-Touch Monitor) new concepts have developed in order to help the transition to a post-Windows-Icon-Menu-Pointer (WIMP) era. This gives rise to important questions such as: (1) What is the most appropriate mapping between the 2D interface surface and the 3D world? and (2) Can previous techniques used with other devices (e.g., joystick, keyboard and mouse) be used in 3D navigation?

To begin answering these questions, we first endeavor to understand touch interactions and previous multi-touch work. We believe that all those aspects create a foundation that is necessary for the development of a sound post-WIMP framework [3]. After the related work section, we cover our proposed solution, discussion and future work.

33.2 Background

33.2.1 *Understanding Touch Interactions*

A common option for multi-touch interaction is to use the set of points corresponding to n touches in a direct manner. However, to achieve a more natural interaction between the screen and the user, studies like [4–8] provide a different take on how touch information can be used. For example, [4] studies finger orientation for oblique touches which gives additional information without having extra sensors (e.g., left/right hand detection). In another example, Benko and Wilson [8] study dual finger interactions (e.g., dual finger selection, dual finger slider, etc.). Additional work dealing with contact shape and physics can be found in [6, 7]. In a very comprehensive review of finger input properties for multi-touch displays [5] provides suggestions that have been used already in [4].

One aspect that is important to have in mind is whether to keep rotations, translations and scaling separate [9] or combined [10]. If the latter is chosen, the user's ability to perform the operations separately may become a problem [9].

One very important point found in [11, 12] is that one-hand techniques are better for integral tasks (e.g., rotation) and two hands perform better with separable tasks. For our particular work, one can think of using one hand to perform common rotations and translations, and using two hands when special rotations need to be performed, utilizing the second hand to indicate a different point of reference.

33.2.2 Virtual Devices

In [1], Nielsen and Olsen used a triad mouse to emulate a 3D mouse. What is important about this work is how they perform 3D rotations, translations and scaling (one at a time). For example, in 3D rotation, they use point P_1 as the axis reference and points P_2 and P_3 to define the line forming the rotation angle to be applied to the object. In more recent work [10], one can find subtle similarities with [1], in the proposition of defining a point of reference to allow seamless rotation and translation.

The Virtual Sphere [2] is an important development in 3D rotation methods previously proposed, which was tested against other virtual control devices. It was found that the Virtual Sphere and the continuous XY + Z device behaved best for complex rotations (both devices behave similar with the exception that the XY + Z device does not allow for continuous rotations about all x, y, z axes). The Virtual Sphere simulates a real 3D trackball with the user moving left-right (x-axis), top-down (y-axis) and in circular fashion (z-axis) to control the rotation of the device. Similar work can be found in [13] with The Rolling Ball and in [14] with the Virtual Trackball. Another idea, similar to The Virtual Sphere [2] is the ARCBALL [15]. The ARCBALL “is based on the observation that there is a close connection between 3D rotations and spherical geometry” [16].

33.2.3 Gesture Recognition

Different methods for gesture recognition have been used in the past including Hidden Markov Models [17], finite state machines [18, 19], neural networks [20], featured-based classifiers [21], dynamic programming [22], template matching [23], *ad hoc* recognizer [24] and simple geometric recognizers [25–27]. An in-depth review can be found in [28]. For the purpose of this paper, we have concentrated in the simple geometric classifier, also known as geometric template matching.

The \$1 algorithm [25] provides a simple way to develop a basic gesture recognizer directed at people that do not have either the time or knowledge to implement more complicated algorithms. For example, algorithms based on Hidden Markov Models or neural networks. At the same time, \$1 provides a very fast solution to interactive gesture recognition with less than 100 lines of code [25]. The algorithm goes through four steps. The first step is to resample the points in the path. The idea is to make gestures comparable, by resampling each gesture at N points (e.g., $N = 64$) [25]. The second step is to find “the angle formed between the centroid of the gesture and gesture’s first point” [25], which is called the indicative angle. Then this step of the algorithm rotates the gesture so the indicative angle is equal to zero. The third step includes the scaling of the gesture

to a reference square which will help to rotate the gesture to its centroid. After scaling, the gesture is translated to a reference point so the centroid is at (0, 0). Finally, step 4 calculates the distance between the candidate gesture to each stored template and using a score formula which gives 0 to the closest gesture and 1 to farthest gesture [25]. The primary limitation of this algorithms are found in the incorrect processing of 1D gestures.

The \$N algorithm [26] extends the \$1 algorithm [25] primarily to allow single strokes to be recognized. The algorithm works by storing each multistroke as a unique permutation. This means that for each multistroke composed of two 2 strokes, the system creates 8 unistrokes. Another particular feature of this algorithm is that it allows the option for the stroke to be bounded by an arbitrary amount of rotation invariance. For example, to make a distinction between A and V, rotation must be bounded by less than $\pm 90^\circ$ [26]. This algorithm also supports automatic recognition between 1D and 2D gestures by using “the ratio of the sides of a gestured’s oriented bounding box (MIN-SIDE vs. MAX-SIDE)” [26]. In addition, to better optimize the code, \$N only recognizes a sub-set of the templates to process. This is done by determining if the start directions are similar, by computing the angle formed from the start point through the eight point. In general, this algorithm which contains 240 lines, was faster than \$1 when using 20 to 30 templates. Algorithms \$1 and \$N utilized the Golden Section Search [29]

Other methods similar to \$1 [25] and \$N [26] algorithms have been implemented. For example, the *Protractor Gesture Recognizer* algorithm [27] works by applying a nearest neighbor approach. This algorithm is very close to the \$1 algorithm [25] but attempts to remove different drawing speeds, different gesture locations on the screen and noise in gesture orientation.

33.2.4 Multi-Touch Techniques

We believe that all of the related work dealing with multi-touch, regardless whether it was designed for manipulation of objects or navigation of 3D graphical scenes, can contribute to the set of unifying ideas that serves as the basis for our approach.

Some of the work in 3D interactions has been specific for multi-touch, which is our focus as well. In [10], Hancock et al. provide algorithms for one, two and three-touches. This allows the user to have direct simultaneous rotation and translation. The values that are obtained from initial touches T_1 , T_2 and T_3 and final touches T'_1 , T'_2 and T'_3 are Δyaw , Δroll and Δpitch which are enough to perform the rotation in all three axes and Δx , Δy , Δz to perform the translation. A key part of their study showed that users prefer gestures that involve more simultaneous touches (except for translations). Using gestures involving three touches was always better for planar and spatial rotations [10].

A different approach is presented in Rotate 'N Translate (RNT) [30], which allows planar objects to be rotated and translated using opposing currents.

This particular algorithm is useful for planar objects and it has been used by 3D interaction methods (e.g., [31]).

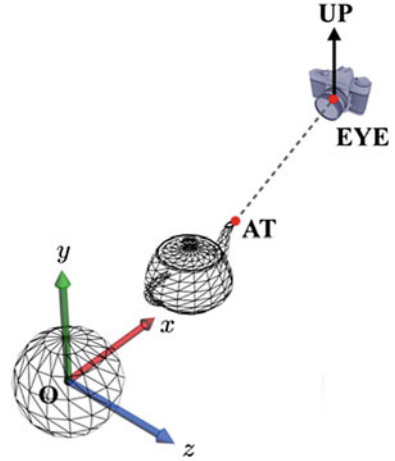
A problem that occurs when dealing with any type of 3D interaction in multi-touch displays, is the one of spatial separability [9]. To address this problem, in [9], the authors proposed different techniques that will allow the user to perform the correct combination of transformations (e.g., scaling, translation + rotation, scaling + rotation + translation, etc.). The two methods that were most successful were Magnitude Filtering and First Touch Gesture Matching. Magnitude Filtering works similarly to snap-and-go [9]. This method has some differences from normal snapping techniques because it does not snap to pre-selected values or objects. In addition, the authors introduce a catch-up zone allowing “continuous transition between the snap zone and the unconstrained zone.” [9]. The latter method, First Touch Gesture Matching, works by minimizing “the mean root square difference between the actual motion and a motion generated with a manipulation subset of each model” [9]. To select the most appropriate model, each prospective model creates two outputs, best-fit error and magnitude of the appropriate transformation. These outputs are given to an algorithm that decides which models to apply.

33.2.5 Design Guidelines

In [3], Jacob et al. proposed interfaces within a post-WIMP framework. In this framework, they try to find a balance between Reality Based Interactions (RBI) and artificial features. RBI includes Naïve Physics, Body Awareness and Skills, Environment Awareness and Skills and Social Awareness and Skills. To allow an application to balance itself, they proposed certain unrealistic features, providing a tradeoff between RBI and artificial methods. The characteristics that can be traded are: expressive power, efficiency, versatility, ergonomics, accessibility and practicality [3].

In another work [10], Hancock et al. also proposed some more specific guidelines when dealing with multi-touch rotations and translation. This includes the ability to provide more degrees of freedom than classical WIMP (ability to do rotation and scale independently or together), provide a constant connection between the visual feedback and the interaction, prevent cognitive disconnect by avoiding actions that the user may not be expecting, and provide realistic 3D visual feedback.

For a concise set of guides for 3D interaction, please review Bowman et al. [16].

Fig. 33.1 Camera view [33]

33.3 Proposed Solutions

33.3.1 Set Up

33.3.1.1 Camera

To test our work, we use OpenGL and perform the visualization through a virtual camera developed in [32] and described by [33], as shown in Fig. 33.1. One can see that the UP vector indicates which way is up, the EYE (or ORIGIN) vector indicates the position of the camera and the AT (or FORWARD) vector indicates the direction in which the camera is pointing.

33.3.1.2 Multi-Touch

We are using Windows 7 multi-touch technology [34] to test our proposed solutions with a 3M M2256PW Multi-touch Display. Windows 7 provides two ways of using the touches from the device. The first one is gesture-based, identifying simple pre-defined gestures and the second is the raw touch mode which provides the ability to develop any type of interaction. We chose the latter because our end goal is to create custom user interactions and for that we preferred to work at the lowest level possible that is available to us. Each touch has a unique identification (ID) that is given at the moment of TOUCHDOWN, to be used during the TOUCHMOVE, and to end when TOUCHUP has been activated. The ID gives us a very nice way to keep track of a trace, which is defined as the path of the touch from TOUCHDOWN to TOUCHUP.

33.3.1.3 Visual Display

As a test case, we have created a world with 64 by 64 by 64 spheres, in a cubic arrangement, drawn in perspective mode, where each sphere has a different color. This allows the user to test the 3D navigation provided by our gestures while having a visual feedback. It is important to note that we colored the spheres in an ordered fashion using the lowest RGB values in one corner and the highest values in the opposite corner of the cube of spheres.

33.3.2 Gesture Recognition

We decided to perform our own gesture recognition as opposed to using language oriented libraries [35]. The reason is that this gives us an ability to improve in current techniques, specially when performance is required. This also gives us the power to choose the best recognition techniques to use them in combination, while assuring that users do not notice performance degradation. Finally, we believe that working at a low level gives us control of the entire interaction environment. This will keep the process simple without using obfuscated libraries, that provide additional functionalities that may slow down the user experience.

For this particular study, we decided to use *ad hoc* recognition because of the few gestures we implemented. For example, for the swipe gestures shown in Fig. 33.2b, c, we determined if a given set of N points of the total M points from a given trace (path) increased either in x and/or y axes. If the value increases, then this qualifies as a swipe gesture.

However, this *ad hoc* method will be replaced with a template based matching [25–27], in particular an adaptation of the \$N\$ [26] algorithm. The reason that this was not selected at this point, is that the \$N\$ algorithm will always yield a gesture. This means that if a user performs an unknown gesture, the algorithm will still classify the closest one. Thus, yielding an incorrect interaction. In future work, we will try to find a proper threshold or a modification to the \$N\$ algorithm to remove this constraint.

33.3.3 Gestures Interaction

We have decided to develop separate gestures for translation and rotation to understand what combinations are more efficient for the user for 3D navigation. This means that when rotating, the user will be rotating by a specific axis and translations will be performed using two axes. We also decided to provide simple gestures for our initial design to see the interaction of the users. Once we have collected more data about the interaction, we can create more complex gestures, if needed.

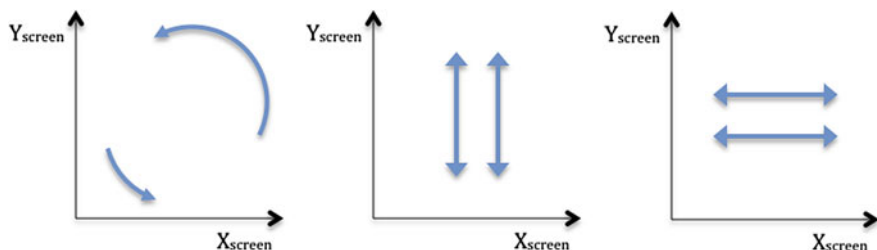


Fig. 33.2 Multi-touch rotation gestures. **a** About 3D x-axis; **b** About 3D z-axis; **c** About 3D y-axis

33.3.3.1 Translation

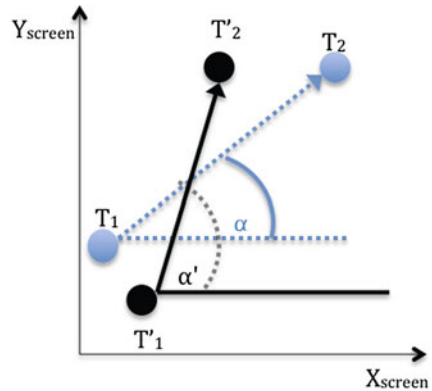
In order to translate the camera, we decided to combine the X- and Y-axes and leave Z by itself. The algorithm for the Y-axis is similar to Algorithm 1, replacing the variable X with the variable Y. The algorithm for the Z-axis is similar to the Y-axis with the exception that it uses 3 touches. In general, the user can perform simultaneous translation by the x and y axes using one finger or translate by the z-axis using three fingers. All the movements can be executed with a single hand.

33.3.3.2 Rotations

To address rotations, we have to think of rotation about x, y and z independently, given that is our belief that separating the rotation will demand a lower cognitive load from the user. This expectation is also supported by [9, 11, 12]. In addition, all the rotations are designed to use only one hand, which is preferable, as demonstrated in [9]. To keep the constraint of using only one hand, the algorithm checks that the touches are within a cluster.

The gesture for rotation about x, as shown in Fig. 33.3, merits to be described in more detail because the other two rotations about y and z use very similar algorithms to those already described for the translations. The only difference is that y and z rotations require two touches each. The gesture for rotation about x begins with T_1 and T_2 , which form an angle α with the horizontal axis. The user's final state is represented with T'_1 and T'_2 , forming an α' angle with the horizontal. Then, the difference between α' and α gives the rotation angle to be applied, about x.

Fig. 33.3 Rotation about the 3D x-axis (see Fig. 33.1)



33.4 Discussion

33.4.1 Gestures

We believe that the set of gestures that we are proposing based on the literature reviewed in the background section and our own preliminary testing will give a starting point to find the most natural gestures to interact in a 3D world using a multi-touch device. Even after finding the most natural gestures for 3D navigation, one will have to compare with other devices such as the ones found in Bowman et al. [16]. As we will outline in the next section, we suggest to make the comparison with a 3D mouse [36].

The first question asked in the introduction was: What is the most appropriate mapping between the 2D interface surface and the 3D world? We have proposed a simple solution to the problem, through the set of gestures described above. Defining and implementing the most natural mapping between this 2D multi-touch interface and the 3D world may still require additional work, but the concepts advanced in this paper may provide an interesting direction towards the solution of this ongoing challenge.

The other question asked in the introduction was: Can previous techniques used with other devices (e.g., Joystick, keyboard and mouse) be used in 3D navigation? We propose that the answer is yes. We can build upon existing work that was developed for the mouse or other interfaces, adapting it for use with multi-touch displays whenever possible. An example of this is The Virtual Sphere [2]. We could take The Virtual Sphere and create a similar device for use with multiple fingers to allow a pure 3D rotation and translation, even emulating a 3D mouse [36]. However, those considerations would be outside the scope of this paper.

In general, we find that multi-touch displays can work efficiently for achieving a more natural user 3D interaction and 3D manipulation.

33.4.2 Proposed Evaluation Technique

The considerations presented above inform our current process of planning the experimental protocol and data analysis methods we will use for evaluating our approach. To answer our research questions and test the proposed gestures, we will recruit at least 30 subjects from the college student population at our university. The reason for our choice of target population is that we believe that all students will have a good grasp of basic mouse interaction, which will facilitate the completion of the experimental tasks by the subjects.

The actual experiment, after allowing the user to become familiarized with the interface, will consist of a set of tasks to test translation and rotation gestures (independently) using our 3M 22" Multi Touch Monitor (Model M2256PW) and the 3D mouse made by 3DConnexion [36]. For each of the tasks, we will measure the time of execution to complete the task, and the accuracy of the movement. For the completion time, we will use an external game controller to start and stop the time, and for the accuracy of the movement, we will automatically record the initial and final positions. In addition to the automated recording of performance data, we will ask the subjects to complete a short usability questionnaire [37].

33.5 Conclusion

Recently, multi-touch displays have become more widely available and more affordable. Accordingly, the search for protocols that will simplify the use of these devices for interaction in 3D data environments has increased in importance. In this paper we have outlined some of the most valuable previous contributions to this area of research, highlighting some of the key past developments that have emerged in the 3D-interaction community. This review of pertinent literature provides a context for the presentation of the core elements of the solution we propose for the interaction in 3D environments through a multi-touch display.

Specifically, we proposed a set of multi-touch gestures that can be used to command translations and rotations in 3 axes, within a 3D environment. Our proposed solution has been implemented using a 3M M2256PW 22" Multi-Touch Monitor as the interaction device. This paper explained the proposed gestures and described how these gestures are to be captured using the information provided by the device. In our definition of the proposed multi-touch gesture set we have established independent gestures for each type of translation and also for each type of rotation. We decided to proceed in this way so that we can study how users prefer to combine or concatenate these elementary gestures.

The next step in the development of our approach is to evaluate its efficiency in a comparative study involving other 3D interaction mechanisms, such as a 3D mouse. The ongoing process of planning the experiments for evaluation takes into account the nature of the devices and general principles of design of experiments,

in an effort to minimize the presence of confounding effects, such as subject fatigue, etc. Our experiments may lead us to define alternative gestures to allow more innovative means of interaction.

Acknowledgments This work was sponsored by NSF grants HRD-0833093, and CNS-0959985. Mr. Francisco Ortega is the recipient of a GAANN fellowship, from the US Department of Education, at Florida International University.

References

1. Nielson G, Olsen D Jr (1987) Direct manipulation techniques for 3D objects using 2D locator devices. In: Proceedings of the 1986 workshop on Interactive 3D graphics, pp 175–182
2. Chen M, Mountford S, Sellen A (1988) A study in interactive 3-D rotation using 2-D control devices. *ACM SIGGRAPH Comput Graph* 22(4):129
3. Jacob R, Girouard A, Hirshfield L, Horn MS, Shaer O, Solovey ET, Zigelbaum J (2008) Reality-based interaction: a framework for post-WIMP interfaces. In: Proceeding of the twenty-sixth annual SIGCHI conference on human factors in computing systems (CHI '08), pp 201–210
4. Wang F, Cao X, Ren X, Irani P (2009) Detecting and leveraging finger orientation for interaction with direct-touch surfaces. In: Proceedings of the 22nd annual ACM symposium on user interface software and technology, pp 23–32
5. Wang F, Ren X (2009) Empirical evaluation for finger input properties in multi-touch interaction. In: Proceedings of the 27th international conference on human factors. ACM, Boston
6. Wilson A, Izadi S, Hilliges O, Garcia-Mendoza A, Kirk D (2008) Bringing physics to the surface. In: Proceedings of the 21st annual ACM symposium on user interface software and technology, pp 67–76
7. Cao X, Wilson A, Balakrishnan R, Hinckley K, Hudson S (2008) ShapeTouch: leveraging contact shape on interactive surfaces. In: TABLETOP 2008. 3rd IEEE international workshop on horizontal interactive human computer system 2008, pp 129–136
8. Benko H, Wilson A, Baudisch P (2006) Precise selection techniques for multi-touch screens. In: Proceedings of the SIGCHI conference on human factors in computing systems (CHI '06), pp 1263–1272
9. Nacenta MA, Baudisch P, Benko H, Wilson A (2009) Separability of spatial manipulations in multi-touch interfaces. In: GI '09 proceedings of graphics interface 2009. Canadian Information Processing Society, Toronto, May 2009
10. Hancock M, Carpendale S, Cockburn A (2007) Shallow-depth 3D interaction: design and evaluation of one-, two- and three-touch techniques. In: Proceedings of the SIGCHI conference on human factors in computing systems, p 1156
11. Kin K, Agrawala M, DeRose T (2009) Determining the benefits of direct-touch, bimanual, and multifinger input on a multitouch workstation. Canadian Information Processing Society, Toronto, May 2009
12. Moscovitch T, Hughes J (2008) Indirect mappings of multi-touch input using one and two hands. In: Proceeding of the twenty-sixth annual SIGCHI conference on human factors in computing systems (CHI '08), pp 1275–1284
13. Glassner AS (1993) Graphics gems. Morgan Kaufmann, San Francisco
14. Arvo J (1994) Graphics gems II. Morgan Kaufmann, San Francisco
15. Heckbert PS (1994) Graphics gems IV. Morgan Kaufmann, San Francisco
16. Bowman DA (2005) 3D user interfaces: theory and practice. Addison-Wesley, Boston, p 478

17. Sezgin T, Davis R (2005) HMM-based efficient sketch recognition. In: Proceedings of the 10th international conference on intelligent user interfaces (IUI '05)
18. Hong P, Huang T (2000) Constructing finite state machines for fast gesture recognition. In: 15th international conference on pattern recognition (ICPR'00), vol 3, p 3695
19. Hong P, Huang T, Turk M (2000) Gesture modeling and recognition using finite state machines. In: IEEE conference on face and gesture recognition, Mar 2000
20. Pittman J (1991) Recognizing handwritten text. In: Human factors in computing systems: reaching through technology (CHI '91), New York, pp 271–275
21. Rubine D (1991) Specifying gestures by example. *ACM SIGGRAPH Comput Graph* 25(4):329–337
22. MacLean S, Labahn G (2010) Elastic matching in linear time and constant space. In: International workshop on document analysis systems 2010 (DAS '10)
23. Kara L, Stahovich T (2005) An image-based, trainable symbol recognizer for hand-drawn sketches. *Comput Graph* 29(4):501–517
24. Notowidigdo M, Miller R (2004) Off-line sketch interpretation. In: AAAI fall symposium, pp 120–126
25. Wobbrock J, Wilson A (2007) Gestures without libraries, toolkits or training: a \$1 recognizer for user interface prototypes. In: Proceedings of the 20th annual ACM symposium on user interface software and technology (UIST '07)
26. Anthony L, Wobbrock J (2010) A lightweight multistroke recognizer for user interface prototypes. In: Proceedings of graphics interface 2010 (GI'10), Toronto
27. Li Y (2010) Protractor: a fast and accurate gesture recognizer. In: Proceedings of the 28th international conference on human factors in computing systems (CHI '10), New York
28. Johnson G, Gross M, Hong J (2009) Computational support for sketching in design: a review. *Found Trends Human-Comput Inter* 2: 1–93
29. Press WH, Flannery BP, Teukolsky SA, Vetterling WT (2007) Numerical recipes, 3rd edn. The art of scientific computing. Cambridge University Press, Hong Kong
30. Kruger R, Carpendale S, Scott S, Tang A (2005) Fluid integration of rotation and translation. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 601–610
31. Reisman J, Davidson P, Han J (2009) A screen-space formulation for 2D and 3D direct manipulation. In: Proceedings of the 22nd annual ACM symposium on User interface software and technology, pp 69–78
32. Wright RS, Haemel N, Sellers G, Lipchak B (2010) OpenGL superbible. Comprehensive tutorial and reference. Addison-Wesley, Reading
33. Han J, Kim J (2011) 3D graphics for game programming. Chapman and Hall, London
34. Kiriathy Y, Moroney L, Goldshtein S, Fliess A (2009) Introducing windows 7 for developers. Microsoft Press, Redmond
35. Laufs U, Ruff C, Zibuschka J (2010) MT4j-a cross-platform multi-touch development framework. In: ACM EICS 2010, workshop: engineering patterns for multi-touch interfaces, pp 52–57
36. O'Brien T, Keefe D, Laidlaw D (2008) A case study in using gestures and bimanual interaction to extend a high-DOF input device. In: Proceedings of the 2008 symposium on interactive 3D graphics and games (I3D '08), New York
37. Lazar J, Jinjuan Heidi Feng D, Harry Hochheiser D (2010) Research methods in human-computer interaction. Wiley, New York

Chapter 34

TCP with Extended Window Scaling

Michal Olšovský and Margaréta Kotočová

Abstract With the expanding amount of data transferred over communication links it is necessary to improve the links and appropriate network devices to match the traffic requests. The most common way of increasing network throughput and improving performance in general is the replacement of the network devices and links. This way is reliable but usually expensive. Different way of network performance is the change of protocol stack. This paper introduces an effective way of increasing network performance by means of improving the most common transport protocol TCP. The improvement, called extended window scaling, increases the amount of unacknowledged data in the network from 1 GB introduced in RFC1323 to future 64 TB.

34.1 Introduction

The first version of the most common transport protocol TCP was introduced in RFC793 in 1981 [1, 2]. To match the increasing traffic requests (bandwidth, delay, etc.), it was necessary not only to improve hardware part of the communications, but software part as well. Improvements of the TCP, usually called TCP variants or

M. Olšovský (✉)
IEEE Conference Publishing, Ilkovičova 3, 84216,
Bratislava 4, Slovakia
e-mail: olsovsky@fiit.stuba.sk

M. Kotočová
Institute of Computer Systems and Networks, Faculty of Informatics and Information
Technologies Slovak University of Technology in Bratislava, Ilkovičova 3,
84216, Bratislava 4, Slovakia
e-mail: kotocova@fiit.stuba.sk

extensions, are mainly focused on the most equitable use of available communication lines [3, 4, 5].

The first improvement of the TCP for higher performance was published in RFC1323 [6]. The most important part of this document introduces window scaling extension. Since 1992 [6], there have been many new TCP variants, which can be divided into 2 main groups based on the end network type—wired and wireless networks [7, 8]. These groups consist of smaller groups whose key element is the way how the congestion is detected. Based on this hierarchy, we can recognize following variants: CUBIC, CompoundTCP, Sync-TCP for wired networks and JTCP, TCP CERL and CompoundTCP + for wireless networks [9–13]. All these variants have one thing in common—after detected congestion, they want to decrease and later increase congestion window while keeping in mind intra/inter-protocol fairness.

The window scaling extension, introduced in RFC1323, solved the problem with small congestion window. Based on the reserved field for window size in TCP header, the maximal size of congestion window can be 64 kB [1, 6]. It means that the amount of unacknowledged data in network cannot exceed the limit of 64 kB. Using this extension, it is possible to extend the unacknowledged amount of the data in the network up to 1 GB [6] which seems to be sufficient for the most transmission links. On the other hand it can be insufficient in some specific situations—reliable long-fat networks (LFN) with proxy generating aggregated traffic can found this size limiting [14].

Our new approach increases today's 1 GB–64 TB which can be found senseless these days but the same opinions were shared when 64 kB and 1 GB congestion window was introduced. As we later explain, the goal was to improve performance but to keep the overhead and additional processing delay as minimal as possible. Therefore the extension is used only in situations, when it is really necessary. This extension can be used in combination with any existing TCP variant and extension.

34.2 Theoretical Background

The original size 1 GB has its justification in the way, how TCP handles and verifies data in the network. As we know, TCP is numbering every single data segment which will be sent out and is comparing the numbers in received segment together with expected numbers. Based on the sequence and acknowledgment 32-bit numbers TCP is able to reorder received segments and decide whether received segment isn't stray and out of actual window. Therefore TCP cannot use whole 4 GB (2^{32}) as addressing space and as congestion window as well. To decide whether received segment isn't old, TCP checks if sequence number of the received segment belongs to interval of 2^{31} bytes from the left border of the congestion window. The same rule is applied to the right border. Based on this, maximal congestion window size can be calculated using (34.1) which can be simplified to (34.2).

$$2^{\text{congestion_window_size}} < 2^{31} \quad (34.1)$$

$$\text{congestion_window_size} < 2^{30} \quad (34.2)$$

Maximal congestion window is defined as a product of $2^{\text{scaling_factor}}$ and $2^{16}-1$ (maximal unscaled congestion window). It means that scaling factor must be less than or equal to 14.

Based on this equations we can said, that any further congestion window enlargement cannot be done without the change of sequence and acknowledgement number's size together with support of larger than 32 bits variables.

34.3 The Principles of Main Features

While creating new TCP extension, it was necessary to keep in mind some important factors—features, which had to be observed. Some of the main features are:

- Backward compatibility with TCP protocol header and extension field format.
- The usage of the highest version of window scaling type supported on both sides.
- Automatic usage with minimal overhead.

34.3.1 Backward Compatibility

To keep the backward compatibility with original TCP header, the only way how to create a new TCP extension was to use the field for extensions (Options) despite the fact that the easiest way will be to modify the reserved fields for sequence and acknowledgement numbers in the TCP header. When we want to keep backward compatibility with extension field format, we had to use format shown in Table 34.1. The minimum length of the extension is 3 with changeable Value field. Basic window scaling uses 3 B length formats. The appropriate fields are filled as shown in Table 34.2.

It was necessary to use similar extension format as the window scaling does to achieve the requested compatibility with basic window scaling. It means that fields Type and Length have the same values. The only difference is the Value field, which values can range from 43 up to 186. The new format of advanced window scaling option is shown in Table 34.3. Details why it was necessary to use these values are explained in the following chapters.

Table 34.1 General extension format

| Type | Length | Value |
|------|--------|-------|
|------|--------|-------|

Table 34.2 Window scaling extension format

| | | |
|----------|------------|-----------------|
| Type = 3 | Length = 3 | Value = <0; 14> |
|----------|------------|-----------------|

Table 34.3 Extended window scaling extension format

| | | |
|----------|------------|-------------------|
| Type = 3 | Length = 3 | Value = <43; 186> |
|----------|------------|-------------------|

34.3.2 Window Scaling Selection

Before the end node can use any extension, it has to be sure, that the other end node will support the certain extension and chooses it as well. As it was introduced in [6], when the end node wants to use TCP with some extension, it has to announce this extension in appropriate TCP header part called Options during initial three-way handshake at the beginning of the communication. This process was clear with the basic window scaling. The new approach with backward compatibility made it a bit more complicated so it had to be improved and extended with some kind of coding and decoding process of the window scaling extension’s announced type.

Coding and decoding processes are performed only when announcing the extended window scaling extension. Standard window scaling extension is announcing only scaling factor while the extended window scaling has to advertise two parameters—scaling factor and number of bits which will be used for enlargement of the sequence and acknowledgement number’s fields. All these details need to be store in 8 bits.

34.3.3 Message Coding

As we stated extended window scaling isn’t announcing only scaling factor but the number of bits which will be used for extension of the sequence and acknowl- edgement number’s fields as well. Therefore the main goal of the coding process is to calculate the number, which will be announced in the Value field from these two values. The length of the extension format can be increased at least by 1 byte with minimal length of 3 bytes (1 byte for data). It means that if the end node wants to transfer additional bits of sequence and acknowledgement number during com- munication with minimal overhead, it will use 1 byte for data which works 4 bits for sequence number and 4 bits for acknowledgement number. Based on this, we can said that number of bits which will be used for enlargement starts at 4, can be increased with steps of 4 up to 32. Usage of any additional bits means to support

Table 34.4 Ω factor dependence on the number of bits

| | | | | | | |
|----------|----|----|----|----|----|----|
| Bits | 12 | 16 | 20 | 24 | 28 | 32 |
| Ω | 4 | 12 | 24 | 40 | 60 | 84 |

operations with numbers larger than 32 bits as well. To make sure that this number will be a multiple of number 4 we use (34.3).

$$bits_rounded = 4 * (bits / 4) \quad (34.3)$$

$$scaling_factor = 14 + 4 * (bits / 4) \quad (34.4)$$

$$Value = (43 + 4 * ((scaling_factor / 4) - 1) + bit - 15 + \Omega) \quad (34.5)$$

Maximal scaling factor which can be used for scaling process depends on the number of bits for enlargement. It starts at 15, ends at 46 and can be calculated using (34.4). The number which will be put in the Value field can be calculated using (34.5) where Ω can be selected from the Table 34.4 and can vary from 43 up to 186.

This unique number includes the information about the type of the highest scaling extension the end node supports, about scaling factor which will be used and number of bits which will be used for extension.

34.3.4 Message Decoding

The goal of decoding is to find out, what the second end node supports. At first, the field Value needs to be checked. If the number belongs to the interval $<0; 14>$ it means that the number is not coded and the node supports only basic scaling. Field Value represents the used scaling factor. On the other hand, if the number belongs to the interval $<43; 186>$ it means that the node support basic and extended scaling as well. To get the offered scaling factor together with the number of bits for the extension we cannot use any simple formula due to more factors used in coding process. The best way is to use Table 34.5. Line Range represents the received number in the Value field. Scaling factor for specific range can be calculated using the formula in line Factor. This same rule can be applied to the number of bits for extension which is placed in line Bits.

Both nodes have to keep their offered scaling factor and number of bits but cannot use them without comparison with received offer from second node.

Nodes will compare the number of bits and in case these two values don't match, both nodes will use smaller number. Node which has to decrease the number of bits has to decrease the scaling factor as well—node will use maximal scaling factor allowed for used number of bits. Second node will be aware of this as it has the same algorithm implemented and will expect congestion window scaled with the new decreased scaling factor.

To this point, we wrote about nodes that support extended scaling. But in case the node doesn't support it, it cannot decode the message. As we stated in the

Table 34.5 Extended window scaling decoding rules

| | | | | |
|--------|--------|---------|---------|---------|
| Range | 43–46 | 47–54 | 55–66 | 67–82 |
| Factor | v-28 | v-32 | v-40 | v-52 |
| Bits | 4 | 8 | 12 | 16 |
| Range | 83–102 | 103–126 | 127–154 | 155–186 |
| Factor | v-68 | v-88 | v-12 | v-140 |
| Bits | 20 | 24 | 28 | 32 |

beginning, one of the main goals was to keep the backward compatibility. It means that if the node doesn’t support extended scaling, it won’t ignore this offer but will simply use the basic scaling and the second node will use the same option. This is possible because of these two things:

1. Default rule in RFC1323
2. Extension format.

Default rule in RFC1223 says, that if the node has received offer for window scaling (Type = 3) and the field Value is greater than 14, it will accept the offer (in case it supports it), set the local copy of the second end node’s scaling factor to 14 and use the basic scaling instead of ignoring the offer.

This brings us to the second thing—extension format. As we have written extended window scaling uses nearly the same extension format—the same type and length. It means that the node which doesn’t support the extended scaling will look at the offer (Type = 3), check Value, set it to 14 and simply use the basic scaling.

The end nodes can use:

- extended scaling if both nodes support it.
- no scaling if at least one of them doesn’t support scaling.
- standard scaling in other cases.

34.3.5 Usage

Extended window scaling can be used anytime during communication. At the beginning both end nodes will increase the maximal values of congestion window together with the limits of sequence and acknowledgement numbers. Standard scaling process is extended with two new operations which enable the usage of acknowledgement and sequence numbers larger than 32 bits while keeping the backward compatibility with standard TCP header.

First operation is responsible for breaking the sequence and acknowledgement numbers into 32 bit part and the rest. This operation is done after the packet is created but before the TCP header is finalized (Fig. 34.1).

As soon as the sequence (SEQ) and acknowledgement (ACK) numbers are increased, they are checked if they have exceeded standard limit 2^{32} . If not, the

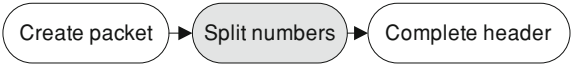


Fig. 34.1 Split operation

Table 34.6 Extended window scaling usage format

| | | | |
|-----------|------------------|-----------------------------|-----------------------------|
| Type = 30 | Length = <3; 10> | R_A = <0; 2 ³² > | R_S = <0; 2 ³² > |
|-----------|------------------|-----------------------------|-----------------------------|

whole 32-bits SEQ and ACK are inserted into TCP header and the header is released for further processing. If yes, the ability of extended scaling is checked. If not present, numbers will overflow and the whole 32-bits SEQ and ACK are inserted into TCP header and the header is released for further processing. If yes and numbers are larger than $2^{(32+bits)}$ the numbers will overflow. Otherwise TCP header is checked if extension with type 30 is created (Table 34.6). If not, extension is created. Lower 32 bits of ACK and SEQ are inserted into appropriate field in the TCP header. The rest bits (up to 32) are inserted into created extension—the rest of ACK into R_A and the rest of the SEQ into R_S. Based on the inserted bits, length of the extension is calculated. Both fields R_A and R_S have to have the same length due to reading the values on the other node. Therefore any empty bits are filled with ‘0’. As these bits represent upper bits, it doesn’t change the value of the ACK and SEQ. The whole operation is shown in Fig. 34.2.

The goal of the second operation is to create SEQ and ACK of full-length. It merges 32 bit number from header together with the rest from the extension field (if present). This operation is done at the beginning of the L4 packet’s processing but before the SEQ and ACK are checked (Fig. 34.3).

Once the received packet enters L4 processing phase, the header is checked if it contains extension with type 30. If not, received SEQ and ACK from header are used. If yes, the support of this extension in local node is checked. If local node doesn’t support this extension, received SEQ and ACK from header are used. If this extension is supported and present in header as well, new SEQ and ACK numbers are created. Lower 32 bits of this number are taken from the standard fields in the header, upper from the extension—from R_A for ACK and R_S for SEQ. Once new numbers are created, header is released for further processing. The whole operation is shown in Fig. 34.4.

Introduced extended window scaling allows the end nodes to increase the congestion window beyond the 1 GB limit. Once the usage of this extension is confirmed during initial handshake, it will dynamically adapt to the network situation and send additional bytes (overhead) in the header only when it is necessary. This extension can be used in combination with any existing TCP variant and extension.

Fig. 34.2 Split operation workflow

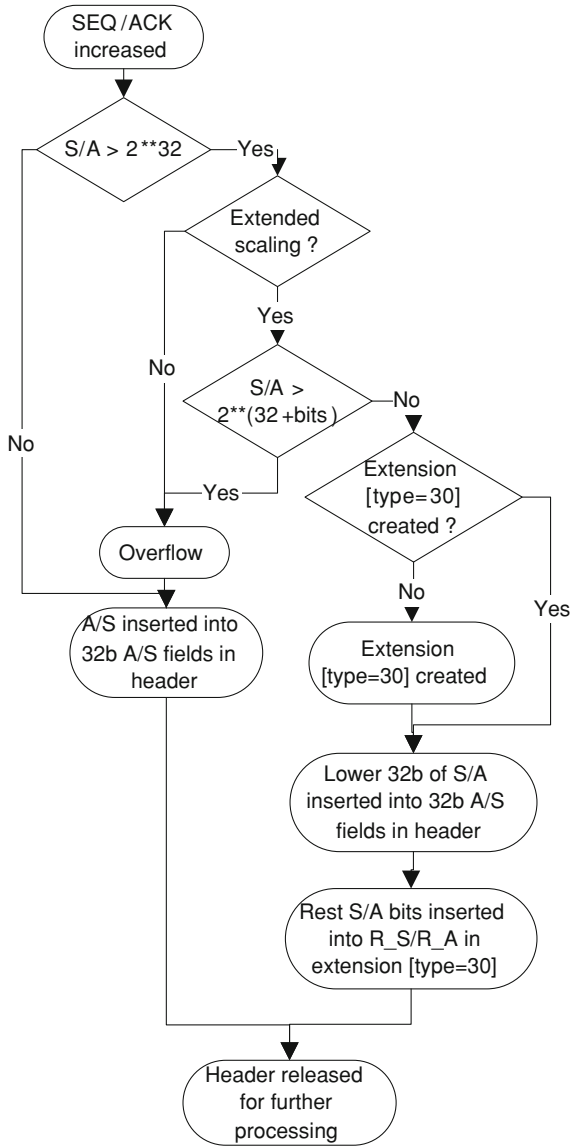
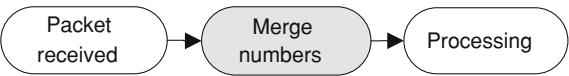


Fig. 34.3 Merge operation



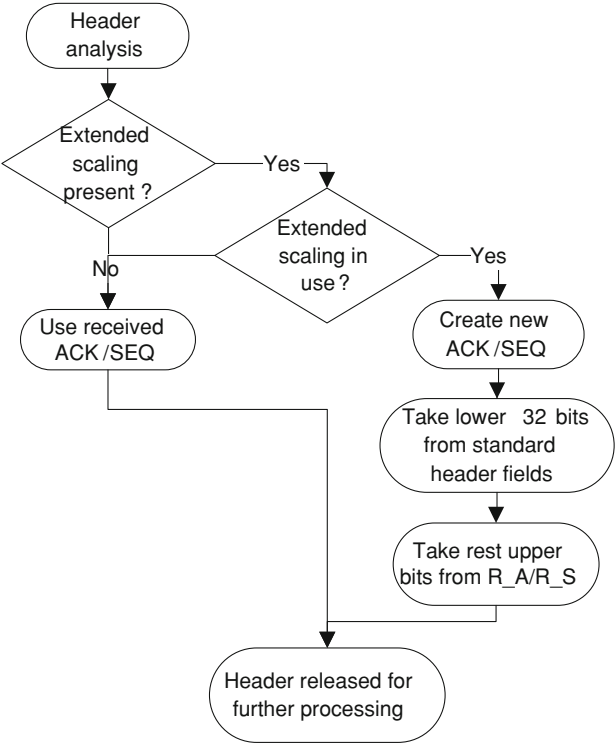


Fig. 34.4 Merge operation workflow

Table 34.7 New limits of the congestion window

| # | Additional bits | All bits | Maximal scaling factor | Maximal congestion window |
|---|-----------------|----------|------------------------|---------------------------|
| 1 | 4 | 36 | 18 | 16 GB |
| 2 | 16 | 48 | 30 | 64 TB |
| 3 | 32 | 64 | 46 | 4096 PB |

34.4 Testing Model

New approach, introduced in this paper was tested theoretically and in ns-2 simulator, where it was implemented [15].

Before we move to ns-2 part let’s point out some specific situation:

- 1. Minimal overhead
- 2. Semi overhead
- 3. Maximal overhead

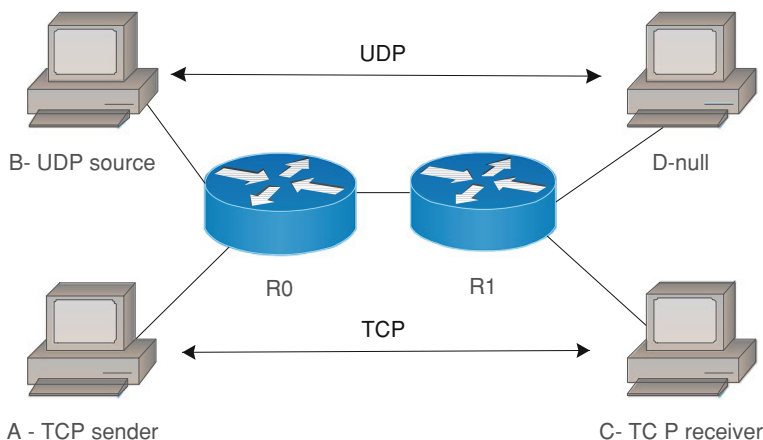


Fig. 34.5 Simulation scheme

In all situations we want to show how can be congestion window increased by using specific overhead. These results are shown in Table 34.7. As we can see, using the minimal overhead of 3 bytes we are able to increase congestion window up to 16 GB.

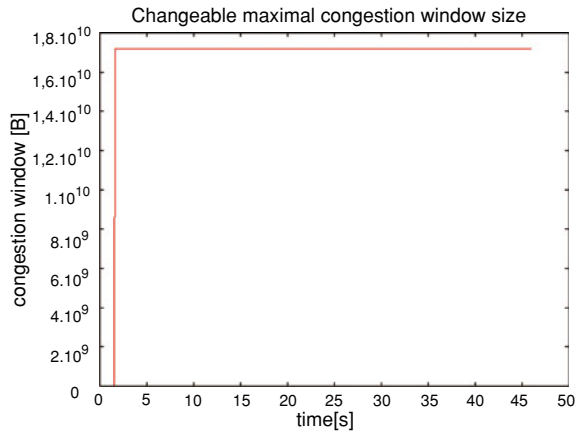
We have done many ns-2 tests, which were focused mainly on improving the TCP performance by means of increasing the limit of congestion window. First sets of tests were focused on extended three-way handshake functionality and the validation of provided parameters. The rest of the tests were used to validate our assumptions about congestion window limit.

Our testing scheme (Fig. 34.5) consists of two routers which were connected with bottleneck link. Two end nodes were connected to each router—nodes A and B to Router R0 and nodes C and D to Router R1. TCP communication was established between the nodes A and C, while the aggressive UDP traffic was flowing between the nodes B and D. UDP traffic wasn't generated all the time, only in specific intervals to achieve the state of congestion in the bottleneck link. Every link had 250 ms delay and 1,000 Mbps bandwidth. Generated UDP traffic had 1,000 Mbps rate. The queue size was 20,000 packets. During the performance test when we wanted to reach maximal throughput we have suppressed any aggressive UDP traffic.

Implementation of this extension has been associated with many obstacles—e.g. absence of basing scaling, usage of 64 bit variables together with modification of many core files. To eliminate processing delay many operations are implemented as bit operations. There have been many outputs and results, which confirmed the basic functionality of the extended window scaling. Basic functionality was confirmed in tests, where we checked the three-way handshake.

As we can see in Fig. 34.6 end nodes at the beginning of the communication approved extended window scaling—they increased the maximum size of the congestion window up to 16 GB (simulation with minimal overhead). Later, when

Fig. 34.6 Changeable maximal congestion window size



the congestion window crossed the limit of 350 MB, simulation faced new problem—simulator became unstable and the simulation crashed. We have done further investigation and find out that the problem is in core functions of the simulator. To confirm these assumptions, we worked only with 32 bit variables but the problem remains. When the crash occurred, actual throughput was about 233,3 MB/s. We found this throughput as a limitation for ns-2, but the question if ns-2 is suitable for high throughput simulations is another topic.

Apart from this issue, using the 32 bit (maximal) extension in this simulation topology (RTT = 1.5 s) we would be able to achieve almost 21 EBps (21,845 PBps) throughput. In fact this number will be smaller, because the delay, which occurs while processing a large number of packets in the routers, cannot be neglected anymore and this would lead to increased RTT. The usage of the basic scaling will allow throughput only up to 5 Gbps. It means that while using the extended scaling we are able to achieve throughput greater nearly $4 \cdot 10^9$ times.

34.5 Conclusion

In this paper we have proposed the new approach of increasing the TCP performance. Firstly, this approach allows the usage of congestion window larger than 1 GB. Secondly, it shows how to improve TCP performance while keeping the full backward compatibility.

It seems that new available congestion windows are too large and can be overestimated for most transmission links. On the other hand it can be insufficient in some specific situations. Based on our assumptions it will be common in the next few years if we consider that using the minimal overhead of 3 bytes we are able to extend congestion window from today's 1 GB to 16 GB. This introduced approach can be taken as a sophisticated solution without the usage of replacement methods.

We believe that this paper can contribute to the important research field in the communication network performance.

Acknowledgments The support by Slovak Science Grant Agency (VEGA 1/0649/09 “Security and reliability in distributed computer systems and mobile computer networks”) is gratefully acknowledged.

References

1. RFC793—Transmission Control Protocol (1981)
2. Yee-Ting L, Leith D, Shorten RN (2007) Experimental Evaluation of TCP Protocols for High-Speed Networks. *IEEE/ACM Trans Netw* 15(5):1109–1122
3. Ha S, Rhee I, Xu L (2008) CUBIC: a new TCP-friendly high-speed TCP variant. *ACM SIGOPS Oper Sys Rev* 42:64–74
4. El-Ocla H, (2010) TCP CERL: congestion control enhancement over wireless networks. *J Wirel Netw* 16:183–198
5. Brosh E, Baset SA, Misra V, Rubenstein D, Schulzrinne H (2010) The delay-friendliness of TCP for real-time traffic networking. *IEEE/ACM Trans Netw* 18(5):1478–1491
6. RFC1323—TCP Extensions for High Performance (1992)
7. Abdeljaouad I, Rachidi H, Fernandes S, Karmouch A (2010) Performance analysis of modern TCP variants: a comparison of cubic, compound and NewReno. In: *Proceedings of 25th biennial symposium on communications*, pp 80–83
8. Henna S (2009) A throughput analysis of TCP variants in mobile wireless networks. In: *Proceedings of the third international conference on next generation mobile applications, services and technologies*, pp 279–284
9. Miao X, Feng Q, Ping D, Yajuan Q, Sidong Z, Hongke Z (2009) Fairness evaluation of the default highspeed TCP in common operating systems. In: *Proceedings of IC-BNMT 2009*, pp 100–105
10. Waghmare S, Parab A, Nikose P, Bhosale SJ (2011) Comparative analysis of different TCP variants in a wireless environment. In: *2011 3rd international conference on electronics computer technology (ICECT)*, 8–10 April 2011, pp 158–162
11. Charoenwatana L, Rattanabung S (2011) Coexistence of SCTP and TCP variants under self-similar network. *2011 eighth international joint conference on computer science and software engineering (JCSSE)*, 11–13 May 2011, pp 17–22
12. Grieco LA, Mascolo S (2004) Performance evaluation and comparison of Westwood+, New Reno and Vegas TCP congestion control. *ACM Comp Commun Rev* 34:25–38
13. Dukkupati N, Refice T, Cheng Y, Chu J (2010) An argument for increasing TCP’s initial congestion window. *SIGCOMM Comput Commun Rev* 40:26–33
14. Sundararajan JK, Shah D, Médard M, Jakubczak S, Mitzenmacher M, Barros J (2011) Network coding meets TCP: theory and implementation. *Proc IEEE* 99(3):490–512
15. The Network Simulator—ns-2. <http://www.isi.edu/nsnam/ns/>

Chapter 35

Offering SaaS as SOA Services

Ali Bou Nassif and Miriam A. M. Capretz

Abstract Software as a Service (SaaS) is an approach in which software applications are delivered on demand to users as services. Service Oriented Architecture (SOA) is a promising paradigm that allows companies that run their software on different platforms to interoperate among each other. The popularity of SaaS has been soaring since customers do not have to carry the burden of paying money upfront to purchase software licensing. The interest in using SOA to run different applications has been proliferating, especially since web services have started to implement SOA. Research is being conducted to observe how SOA can benefit SaaS. This paper presents an overview of SaaS, SOA and web services. Moreover, a new model is proposed to show how SaaS can be offered as SOA services. Furthermore, a real-life example is given to demonstrate the benefits of using the proposed model.

35.1 Introduction

Web evolution has been visible right from its initiation in order to support several important services. These include social networking, entertainment and business applications. The term *service* has been widely used in the web to support people's demands in many activities. Amazon Web Services (AWS) and Google Apps offer a

A. B. Nassif (✉) · M. A. M. Capretz
Department of Electrical and Computer Engineering,
Faculty of Engineering, The University of Western Ontario,
London, ON, Canada
e-mail: abounass@uwo.ca

M. A. M. Capretz
e-mail: mcapretz@uwo.ca

variety of services. Microsoft found that offering services to customers is inevitable and has introduced Microsoft Azure to compete with AWS and Google Apps.

There has been a huge tendency and intrigue to turn conventional software into services by both vendors and customers. Software as a Service (SaaS) is a technology that offers software applications as services. SaaS is an online delivery of software to customers [1]. Rather than purchasing a software license (On-Premise software) for an application such as Customer Relationship Management (CRM) or Enterprise Resource Planning (ERP) which might cost thousands of dollars, customers subscribe to applications from a SaaS provider and pay fees based on the usage of the application. In a SaaS environment, software becomes a collection of services. The service is deployed from a centralized datacentre and accessed through the Internet on a recurring fee basis [1, 2]. Demand for SaaS providers is increasing and the SaaS architecture should be improved to accommodate the substantial increase of new customers [3]. The advantages of SaaS are obvious as a SaaS provider is the one who manages the application and supporting infrastructure and not the customer. For the customer, the use of SaaS reduces costs of maintaining an infrastructure and the updating of Software.

Many SaaS providers such as Salesforce.com (the top SaaS vendor in CRM), offer SaaS solutions in the form of editions. Each edition has a set of features. Reference [4] shows an example of Salesforce.com's editions and their features.

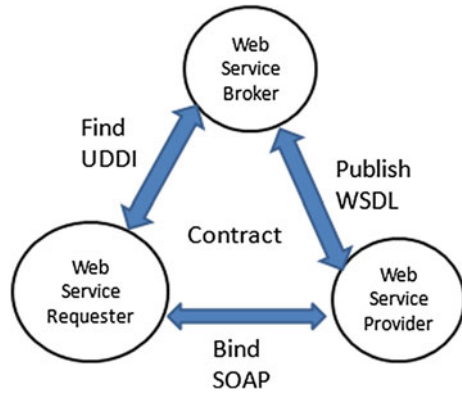
The main drawback of most SaaS vendors such as Salesforce.com, is that customers cannot choose the features of the application that they are interested in. Instead, they have to subscribe to a specific edition and pay for that edition, even if only a few features of this edition are being used.

Service Oriented Architecture (SOA) can be defined as a flexible set of design architectures that embodies a collection of services that communicate amongst each other [5]. SOA allows software functionalities to be available in a form of services to anyone who requests it, provided they are authorized to use these services. SOA is not new. People in the past used SOA for Object Request Brokers (ORBs) which were based on the CORBA mechanism.

SaaS and SOA are completely different. The two terms can be confusing because they both contain the word "service". SaaS can be summarized as the way that software is delivered and used, while SOA is concerned about how software is structured. SaaS applications might or might not rely on SOA. However, if the functionalities of SaaS applications can be converted into SOA services and deployed through SOA, the benefits of leveraging these services in business will be enormous. In this case, any enterprise can satisfy its own business needs by selecting the services they wish from different SaaS providers through SOA.

The rest of the paper is structured as follows. [Sections 35.2](#) and [35.3](#) present the principles and the architecture of SaaS and SOA respectively. [Section 35.4](#) introduces the relationship between SOA and web services. [Section 35.5](#) illustrates the related work. [Section 35.6](#) demonstrates the proposed model to provide SaaS as SOA services. [Section 35.5](#) presents a real-life example to describe the model. Finally, [Section 35.6](#) concludes the paper.

Fig. 35.1 Mapping between web services and SOA [7]



35.2 SOA and Web Services

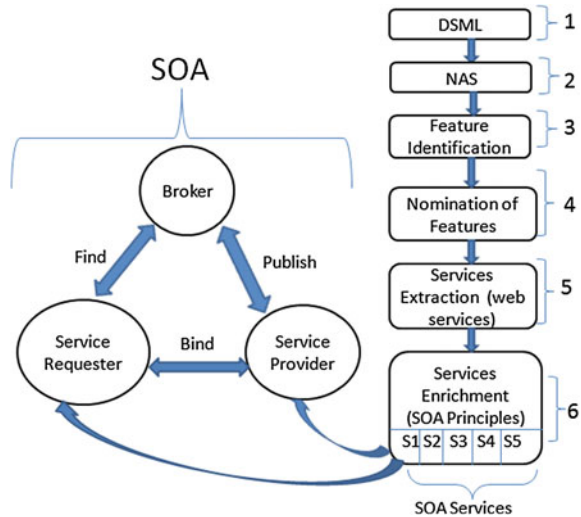
The terms SOA and web services could be confusing. SOA and web services are not the same. SOA is an architecture, while web services can be seen as a technology. Web services can implement SOA but SOA can also be implemented by other technologies such as REST, CORBA, Distributed Computing Environment (DCE) and Java Enterprise Edition (Java EE) [6].

Although SOA can be implemented with service-based technologies other than web services, these technologies have been criticized for not being loosely coupled. Web services became highly acceptable as a standard for SOA implementation because they are pervasive, loosely coupled and are supported by most software vendors. In web services, messages are used for communication as opposed to using operations in older technologies.

Web services can explicitly map the elements of SOA to form a reliable and unobtrusive solution of registering, discovering and executing services. Figure 35.1 depicts the mapping between web services and SOA technologies [7]. Figure 35.2, shows that the web service provider publishes its services with the web service broker using Web Services Description Language (WSDL). The web service requester searches for services registered with the service broker using Universal Description, Discovery, and Integration (UDDI). The role of the service broker is important as customers (service requesters) might search for multiple services and each might have a different service provider. Registering the services with a central broker provides the flexibility required for customers to send queries to find services. Finally, when a customer finds a service, the web service requester binds the web service provider through SOAP protocol.

According to Sneed, there are four main sources for web services [8]. The first source of web services can be purchased from a web service supplier. Secondly, web services can be accessed from the open source community. The third source of web services can be developed from scratch and finally, the fourth source of web services can be extracted from existing applications.

Fig. 35.2 Transition from SaaS to SOA



Sneed argued that if reliability, maintainability, cost, time and testability are taken into consideration, the fourth source which is extracting services from existing applications can be the smartest choice especially for companies who have been in the business for a long time, and thus they have a repository of software functionalities that can be converted to SOA services.

35.3 Related Work

The related work in this area focuses on two main parts. The first part is how to offer On-Premise applications as SaaS. The second one is how to migrate legacy systems to SOA. The two approaches are different, since it is not necessary for SOA to be the infrastructure of SaaS.

35.3.1 Offering On-Premise Applications as SaaS

There are several approaches to offer On-Premise applications as SaaS. These include:

- 1 Hosting the application with a third party server: If an application is written in a third generation language such as Java or C#, the application can be hosted with a third party provider [2, 9]. With little modification of the application (especially the interface), customers can access the software online. In this case, the application is offered to customers as a service. This is similar to the case of the Application Service Provider. This method has been criticized because the

offered service is tightly coupled. For instance, customers do not have a lot of freedom to tailor the application to meet their needs. Furthermore, upgrading the application might be inconvenient to some customers.

- 2 On-Premise applications offered as Web services: In this case, the application can be transformed into a web service [9]. Web services can then be offered to customers through SaaS. Several technologies exist such as SOAP and REST to convert applications to web services.
- 3 On-Premise applications can be migrated to SaaS through Commercial tools: The application can be moved to a SaaS provider through commercial tools [10]. Some of these tools act as a mediator between the customer and the SaaS vendor. In this case, an enterprise might choose to migrate part of its business to SaaS and keep the other part on their servers.

35.3.2 Migration from Legacy Systems to SOA

Although SOA is not new, the demands for SOA have been rapidly increasing after the frills that have been added by web services to SOA environment. Many technologies exist to build SOA applications from scratch, but there is a substantial amount of legacy systems that have an enormous value to their business and consequently these legacy systems cannot be ignored. This stimulated the search for techniques to migrate legacy systems to SOA.

Zhang and Yang [11] defined three main categories to transform legacy system to services. The first category is called *black-box* approach. In this approach, legacy code is integrated into services through adapters. Adapters are software layers which surround legacy code. The second category is called *white-box* approach. In this approach, a reengineering process is applied through investigating, analysing and modifying the legacy code to discover the business logic and then to obtain the code components that can be offered as web services. The third category is the *grey-box* technique. This technique includes both the black-box and the white-box techniques to integrate legacy systems to services.

In this section, three approaches to migrate legacy systems to SOA are demonstrated. These approaches include:

- 1 Wrapping Legacy Software for Reuse in a SOA: This work was proposed by Sneed [8]. In this approach, legacy software is wrapped into an XML shell such that software functionalities can be proposed as web services. This can be achieved through three stages. These stages include salvaging the legacy code, wrapping the legacy code and linking the web services to the business processes. In the first stage, legacy code which is valuable to reuse is identified and business operations and functions are discovered. The second stage is to wrap the legacy code with a WSDL interface. This is achieved by transforming each entry into a method and each parameter into an XML data element. The legacy code is then packaged with SOAP framework. The final stage is to link the web services to the business processes and SOA. This is achieved through a proxy.

- 2 **Architecture Transformation, From Legacy to Three-Tier and Services:** This work was proposed by Heckel et al. to transform legacy systems to three-tier architecture to support SOA [12]. The methodology used in this approach is mainly composed of four steps. The first step is to annotate the legacy source code by code categories. This is to associate each code category with the elements of the target architecture (3-tier), e.g., as user interface, business logic or data. The second step is called Reverse Engineering. In this step, a source graph model is constructed from the annotated source code. This graph is also an instance of a metamodel graph. The third step is redesigning the source graph model to the target graph model to conform to the three-tier and SOA. The redesign state is done using graph transformation rules. The relation between the annotated code and the target model is also saved to support transforming target graph model into the target code. The last step in this methodology is called Forward Engineering. In this step, the target code is generated using the target graph model with the annotated source code.
- 3 **A Wrapping Approach for Migrating Legacy System Interactive Functionalities to Service Oriented Architectures:** This work was proposed by Canfora et al. [13]. The authors presented a black-box approach to integrate interactive functionalities of legacy software into services. Interactive systems are session-based systems where a user exchanges messages with a computer. This can be achieved through a browser or input forms when the user submits a query to the system. The system then processes the query and executes it by presenting an output to the user. The black-box approach is represented by a wrapper which encompasses the user interface and interacts autonomously during the execution of each query with the legacy system on behalf of the user.

35.4 Transition from SaaS to SOA

In this section, a novel model is proposed as presented in Fig. 35.2 to illustrate how SaaS applications can be offered as SOA services. This model is divided into six main stages. These stages include:

- A **Determination of SaaS Maturity Level (DSML):** In this stage, the SaaS application is thoroughly investigated and analyzed to determine for example, the size, maturity level of the application, and programming language used to build the application. The goal of this stage is to determine the coupling level (loose or tight or in between) among the functionalities of the application.
- B **Nomination of Applications for SOA (NAS):** This is the second step of the model. The goal of this step is to select a SaaS application, or mainly, the parts of an application that will be transformed to SOA services. In many cases, it is not required to transform the whole SaaS application to SOA. It is possible that parts of the application will be converted to SOA and the other parts will be

kept as SaaS. For example, a SaaS vendor might choose to select a feature, which they believe is very competitive in the market, and offer it as SOA service.

- C Feature Identification: Each SaaS software has a predefined set of features [14, 15]. An example about the features of Salesforce.com CRM application is shown in Ref. [4]. For instance, Salesforce.com offers its CRM application as five editions: Contact Manager, Group, Professional, Enterprise, and Unlimited. Each edition has a collection of features. In this stage, we are concerned to define and list all the features of the SaaS application part which was nominated to be transitioned to SOA.
- D Nomination of Features: In the previous stage, all features of an application part were listed. In this stage, some features will be selected to be candidates for SOA services. The selection of these features is based on the contribution which they may provide to business.
- E Services Extraction: In the previous stage (Nomination of Features), some features may not be candidates to become SOA services because they might be tightly coupled or depend on other features. In this stage, selected features will be offered as web services through SaaS. Not all web services are SOA services because some web services are not coarse grained and loosely coupled. This depends on the technology being used with web services, e.g. REST or SOAP. In this stage, services that are written with legacy code will be wrapped with XML shell.
- F Services Enrichment: This is the final stage of the model and after this stage, services are ready to be engaged with SOA. In this stage, web services will be checked if they conform to SOA principles (reuse, loosely coupled, coarse grained, etc.). For instance, REST services are not loosely coupled as SOAP services. Here, some services might be redesigned to match SOA principles or standards. Another important point in this stage, is with regards to Salesforce.com; instead of offering five different editions, each containing a group of features. Salesforce.com can now offer more than fifty services, each with a different price.

After the SaaS application is divided into services that comply with SOA principles, the services are now ready to participate with SOA. As shown in Fig. 35.2, the SaaS vendor can act as a service provider who publishes its services with a broker. The SaaS vendor can also be a service requester. In this case, the SaaS provider can acquire additional services from other providers and add them to its application.

The main advantage of the proposed model is that any SaaS provider can offer services as opposed to offering features that are packaged within editions. Moreover, any customer can select the services that they are interested in from a service provider and also select different services from another provider to form a new business model or application. The new business application will be a collection of services supported by different providers that communicate with each other to

provide business values to the company. [Section 35.5](#) provides a real-life example to demystify the premise of the proposed model.

35.5 Real-Life Example

In this section, an example is presented to demonstrate the proposed model. The proposed model was evaluated using a SaaS company called TopCRM. TopCRM is a relatively small company that provides CRM to their customers. Our proposed model was used to help TopCRM convert part of its application into SOA services. TopCRM underwent the five stages of the proposed model during the transition process. This included:

- A DSML: During this stage, an investigation was conducted thoroughly to study the characteristics of the application. This included the requirements, architecture, design, programming language as well as the network and the hardware infrastructure. The results showed that the application was designed using UML models, implemented using C#. There were three main servers that were used to run the business and to store the database. The maturity level of the TopCRM was considered as medium. The database architecture used could be expressed as assigning a single instance and separate database for each tenant (customer). The reason behind using separate databases is that separate databases for each customer are easy to be managed.
- B Nomination of Applications for SOA (NAS): In this stage, a preliminary study was performed to see whether or not the whole system or just a fraction of it would be converted to SOA services. TopCRM proposes CRM solutions through five main sections. These sections include *Contact Manager*, *Sales*, *Marketing*, *Data Warehousing* and *Communications*. At this stage, TopCRM recommended the Data Warehousing section to be offered as SOA services because TopCRM believed that this section could be competitive in the market.
- C Feature Identification: During the Feature Identification stage, all features of the Data Warehousing section were identified. These include *data_analysis*, *report_generator* and *data_storage*.
- D Nomination of Features: In this stage, the *data_storage* feature was selected to be a candidate to become an SOA service. This was because TopCRM believed that this feature can compete with similar features from other providers in the market. When this feature is offered as a service, users can seamlessly store their data securely at TopCRM's servers.
- E Services Extraction: In this step, the *data_storage* feature was converted to a web service using SOAP, WSDL and UDDI. At this stage, the web service *data_storage* was not guaranteed to be a valid as an SOA service.
- F Services Enrichment: In this stage, the *data_storage* service was tested using SOA principles. The service was found to be loosely coupled because of the characteristics of the SOAP protocol. However, there was an issue regarding the

reusability of this service. SOA services must be reusable by multiple applications. To support reusability of the service, new servers were added to handle the large number of users that will use the service. Virtualization technology was involved in the new system to support maintenance and security. Furthermore, the old database structure (single database per tenant) was not the best choice to support the new system because the old database would be very costly if it would be used for a larger number of customers. To resolve this problem, the old database was modified to become a shared database to support more users.

After passing through the six steps of our proposed model, services could be extracted from SaaS applications and would be ready to be published with SOA brokers to be used by customers. Finally, TopCRM published its `data_storage` service with a broker and many customers are now enjoying the feature.

35.6 Conclusion

Software as a Service has been rapidly adopted because customers save the money needed to purchase software licensing and hardware infrastructure to run the software upfront. On the other hand, web services and SOA have been offering many benefits by allowing heterogeneous systems to communicate with each other. The benefits of offering software and hardware as services are enormous. Because legacy systems exist in almost any organization, several works have been conducted to transition legacy systems to SOA. On the flip side, SaaS applications exist and SaaS vendors are looking for techniques to merge SaaS with SOA.

In this paper, a novel model was proposed to discuss how SaaS can be offered as SOA services. This model is mainly composed of six stages. First, the maturity level of the SaaS application will be determined. Second, applications or parts of these applications will be nominated for SOA services. Third, the features that belong to an application or to an application section will be listed. Fourth, features that are nominated to participate with SOA will be listed. Fifth, the nominated features will be converted to web services. In the last stage, web services will be modified to conform to SOA principles and finally these services become ready to be published through SOA. A real-life example was also presented to demonstrate the proposed model.

References

1. Hoch F, Kerr M (2001) Software as a service: strategic background. Software & Information Industry Association, Washington. <http://www.sii.net/estore/pubs/SSB-01.pdf>
2. Chong F, Carraro G (2006) Architecture strategies for catching the long tail. Microsoft Corporation. Technical report MSDN Library

3. Turner M, Budgen D, Brereton P (2003) Turning software into a service. *Computer* 36:38–44
4. Selecting the right sales edition (2011) Salesforce.com. <https://www.salesforce.com/ap/assets/pdf/cloudforce/PricingEditions-SelectingTheRightSalesforceCRMEdition.pdf>
5. Newcomer E, Lomow G (2005) Introduction to SOA with web services, in *Understanding SOA with Web Services*. Addison-Wesley, Upper Saddle River, pp 1–50
6. Vinoski S (2007) REST Eye for the SOA Guy. *Internet Comput IEEE* 11:82–84
7. Amirian P, Mansurian A (2006) Potential of using web services technologies in distributed GIS applications. *GIS Development, Middle East*
8. Sneed HM (2005) Wrapping legacy software for reuse in a SOA. AneCon GmbH, Wien
9. Sun W, Zhang K (2007) Software as a service: an integration perspective. In: Hutchison D *Service-oriented architecture*. Springer Berlin, Heidelberg, pp 558–569
10. Cloud intelligent (2010) <http://www.cloudint.com>
11. Zhang Z, Yang H (2004) Incubating services in legacy systems for architectural migration. In: *Software engineering conference, 2004. 11th Asia-Pacific*, pp 196–203
12. Heckel R, Correia R, Matos C, El-Ramly M, Koutsoukos G, Andrade L (2008) Architectural transformations: from legacy to three-tier and services. In: Mens T, Demeyer S (eds) *Software evolution*. Springer, Berlin, pp 139–170
13. Canfora G, Fasolino AR, Frattolillo G, Tramontana P (2008) A wrapping approach for migrating legacy system interactive functionalities to Service oriented architectures. *J Syst Softw* 81:463–480
14. Nassif AB, Lutfiyya H (2011) Measuring the usage of SaaS applications based on utilized features. In: *The first international conference on cloud computing and services science*, pp 452–459
15. Nassif AB (2011) Measuring SaaS applications based on utilized features. LAP Lambert Academic Publishing, Köln, Germany

Chapter 36

Using Conceptual Mini Games for Learning: The Case of “The Numbers’ Race” (TNR) Application

C. T. Panagiotakopoulos and M. E. Sarris

Abstract This study reports the basic characteristics of an experimental conceptual mini-game called “The Numbers’ Race” (TNR), developed with Microsoft Visual Studio. The TNR mini-game concerns a standalone training application that aims at raising achievement levels in the simple mathematical task of addition and giving insights into the strategies used by children when performing simple computational tasks. In this paper the design and the development issues for TNR application are also discussed. Finally, the results of summative evaluation are presented concerning TNR effects on a student sample. The results are very encouraging regarding the improvement of the sample’s computational skills.

36.1 Introduction

Probably the greatest problem that teachers encounter in contemporary education contexts is the lack of motivation that students often exhibit. As described by researchers [1], this problem is aggravated in the case of mathematics, and debilitated motivation may have major repercussions for the learning process, such as the disinclination to do exercises or consult textbooks. In an effort to tackle these issues, educators have begun to reconsider the traditional forms of learning

C. T. Panagiotakopoulos (✉)

Department of Primary Education, University of Patras, Greece
e-mail: cpanag@upatras.gr

M. E. Sarris

Department of Primary Education, University of Patras, Greece
e-mail: m.sarris@upatras.gr

instruction. Recent reforms, in mathematics education particularly, explore and implement alternative approaches [2].

Since the 2000s, a major shift towards the use of computer mini-games for teaching and training has been observed, replacing the former obsolete rote learning methods. What is particularly appealing about mini-games to the field of primary education is that they incorporate two important features. They can target both part-task training and smaller learning objectives [3]. In that sense, current teaching approaches can support the notion of student-centeredness in education by reinforcing a small set of goal-oriented learning objectives suitably adjusted for each individual student.

The emphasis that current teaching approaches put on students' engagement in constructing knowledge on their own is relevant to the constructivist learning theory [4, 5]. According to constructivism, learning is an internal process where knowledge is accumulated (built up) when individuals assimilate and incorporate new experience into an existing framework [6, 7]. One way of providing children with interesting experiences is via games and simulations. During the last decades, when ICT was embedded in mainstream teaching practice, educational computer games tended to exploit children's natural tendency to play. Thus, possible positive effects of educational computer games may be attributed to the fact that children are willing to spend time and effort on mastering the games.

For this to happen, children must be actively engaged in this process. As observed by Prensky, a combination of 12 elements makes computer games engaging [8]. Among others, fun, interaction, goal problem-solving, outcomes and feedback are probably the most significant features. In an interesting review by another researcher [9], it was found that an equally important factor is the educational context in which the games are used. Educational computer games are most likely to present positive results in mathematics, physics and language. The beneficial effects of gaming are to be found when specific content is targeted and objectives precisely defined [10].

One subcategory of educational games is mini-games, which can help children to obtain basic skills in mathematics or reading, providing both entertainment and instruction to the target audience [11, 12]. Mini-games are short games with simple and immutable rules. The mini-games that focus on a concrete concept are often called conceptual mini-games [13].

In this paper we describe the operation of a mathematical mini-game named "The Numbers' Race" (TNR). The TNR application was designed to address simple computation skills in addition and was developed as an alternative educational tool to enhance mental calculation skills. We also analyze the operation of the application, describe its structure, illustrate the parameters affecting its operation and present the initial results of its final evaluation. The sample consisted of students from primary education and the results are very encouraging as regards the improvement of their computational skills.

36.2 TNR Application: Identification

36.2.1 Description

The TNR application was developed during 2011 in the Computers and Educational Technology Laboratory of the Department of Primary Education of the University of Patras (www.cetl.elemedu.upatras.gr).

The target group of the developed application is primary education students. The TNR application operates as follows: the educator defines two integer numbers, creating a subset of integers between them. The application, using random variables [13, 14], creates a predefined number of additions in this subset. For each addition the application presents the result (SUM) and the first addend (ADDEND1), whereas the second addend is requested (ADDEND2). The second addend appears consecutively, among nine possible addends (ADDEND2_A[i], $i = 1$ to 9), only one of which is displayed each time. The requested possible addend moves clockwise each time and stops in the middle of the application's window, between the pictures of ADDEND1 and SUM. The user has a specified time for response, answering on the correctness or non-correctness of the proposed combination of addends—sum.

The user is able to:

- (a) Ignore the proposed combination of addends—sum by not responding.
- (b) Identify that the combination is wrong.
- (c) Identify that the proposed combination is correct.

Figure 36.1 displays how the application operates, with the appearance of a trinity of numbers (sum—addends). The number SUM is the possible sum of the two numbers: the ADDEND1 and the ADDEND2_A(i), when $i = 1$ to 9. The number ADDEND1 is the first addend. The ADDEND2_A(i) represents each time one of the nine possible addends, from that only one if added to the ADDEND1 gives sum the number SUM.

Note that each time only one of the proposed number from the ADDEND2_A(i) array is visible, which is moved between ADDEND1 and SUM.

If the user correctly decides that the displayed combination of addends—sum is incorrect (X), the score of the “No match” counter rises. If the answer is wrong, the “Error” counter, that indicates the total number of wrong answers, increases.

If the user decides that the proposed combination of addends—sum is correct (✓) then again, if they are right, the score in the counter “Score” rises; otherwise, the “Error” counter increases.

In the second case, when the proposed combination of addends—sum is correct and the user's selection is correct, the application begins a new race, proposing a new set of numbers.

In the next figure (Fig. 36.2) an actual screenshot of the application is presented.

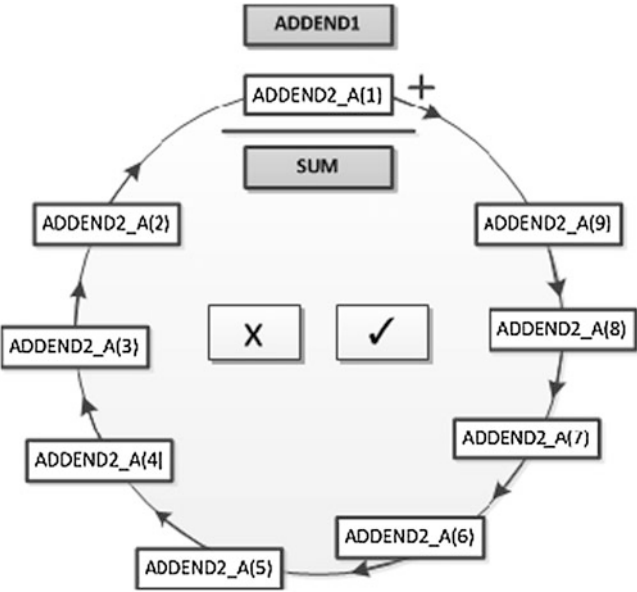


Fig. 36.1 Graphical representation of the application’s operation. The numbers are moving circularly, stopping for a predefined time for users to decide whether the action is right or wrong

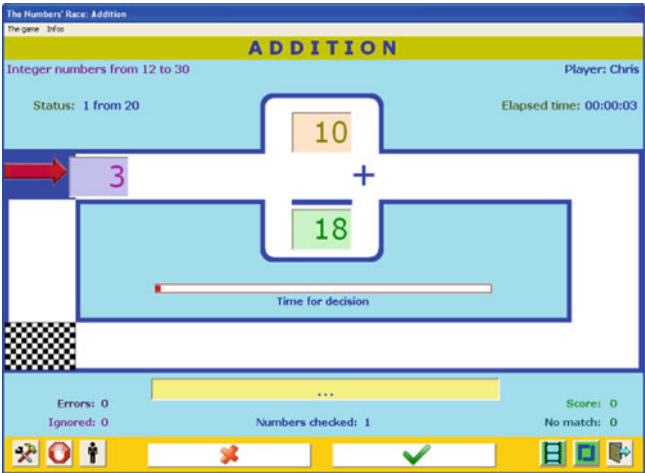


Fig. 36.2 An actual screenshot of the TNR application

At the bottom of the application’s window there are the controls, with “X” for the identification of the wrong combination and “✓” for the identification of the correct combination. In the middle, beneath the numbers, there is a progress bar with a red mark, indicating the time remaining for the user’s decision. The clock

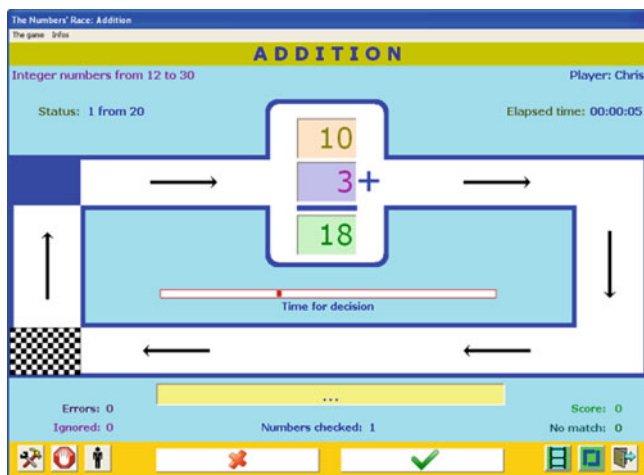


Fig. 36.3 An actual screenshot of TNR while the system is waiting for the user's decision. Arrows demonstrate the numbers' movement

starts counting when the moving `ADDEND2_A[i]`, stops under the `ADDEND1` and its duration is determined by the educator/administrator (Fig. 36.3).

36.2.2 Application's Operational Features

The application runs under Microsoft Windows 7 or earlier versions of the Windows operating system, using a typical microcomputer. Access to the application, from the user's perspective, is personalized. All of the application's operational features are configured through an appropriate window with scroll bars and check boxes (Fig. 36.4).

The educator/administrator using this window is able to modify:

- (1) The moving speed of the numbers. To this end the values of two variables can be changed. The first one changes the movement's step (in pixels) and the other changes the time (in ms) between two successive steps of the moving number.
- (2) The minimum and maximum number of sums. These two numbers are the limits for the calculation of the sequences of sums.
- (3) The number of the requested additions, i.e. the number of repetitions of the game.
- (4) The time allowed for the user to make a decision. For this purpose the movement step (in pixels) and the time (in ms) must be set between two successive steps of the moving mark (progressive bar), which is located under the three numbers, in the middle of the application's window.

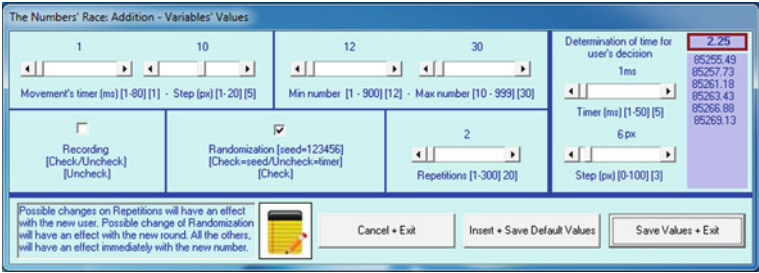


Fig. 36.4 Configuration window of the application’s operational features. The highlighted number shows the time for the user’s decision in seconds

- (5) Either the full set of user’s actions and decisions or only the maximum score of the user’s performance to be recorded. If all of the user’s actions are recorded the administrator has the ability to trace the interaction between user and application in detail (with response times and actions), either reading these from the screen or passing on the data in a Microsoft Excel file for further statistical analysis.
- (6) The sequence of random numbers’ sets. If this checkbox is ticked the sequence will remain the same while the application is running on multiple different computers. This is important for research, when we want all students to deliver the same sequence of number sets and in the same order.

36.2.3 Development Issues

Several timers were used for the development of the application. Bearing in mind compatibility, ease of installation and usage with regard to different computers, we were careful not to use any custom control files (Object Linking and Embedding—OLE custom controls) or ActiveX files.

Special sounds point out the crucial events, e.g. when the numbers start moving, when the user is rewarded or when an error has to be noted. These sounds were recorded and embedded in the application, so as to operate without any prompts from the system or from other peripherals.

The application starts running when the appropriate button is clicked (the first of the group of three buttons, bottom-right of the window in Fig. 36.3). After that, the application checks if a user name is imported and, if so, reads the parameters of the operation. Then, the two numbers DN and UP that determine the limits of the SUM are calculated with random variables; the first SUM between A1 and A2, the ADDEND1 between one and SUM, and the ADDEND2 as the difference between SUM and ADDEND1 (stage 1).

The specific process is described in the following flowchart (Fig. 36.5).

Fig. 36.5 The flowchart of the sum's production

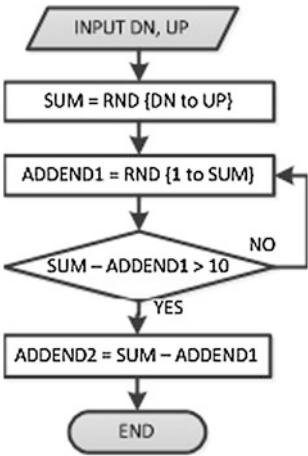
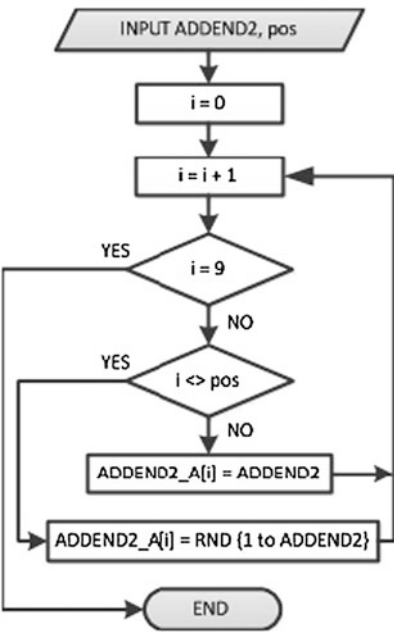


Fig. 36.6 The data flow in which all remaining items of possible second addends in the array ADDEND2_A[i] are placed

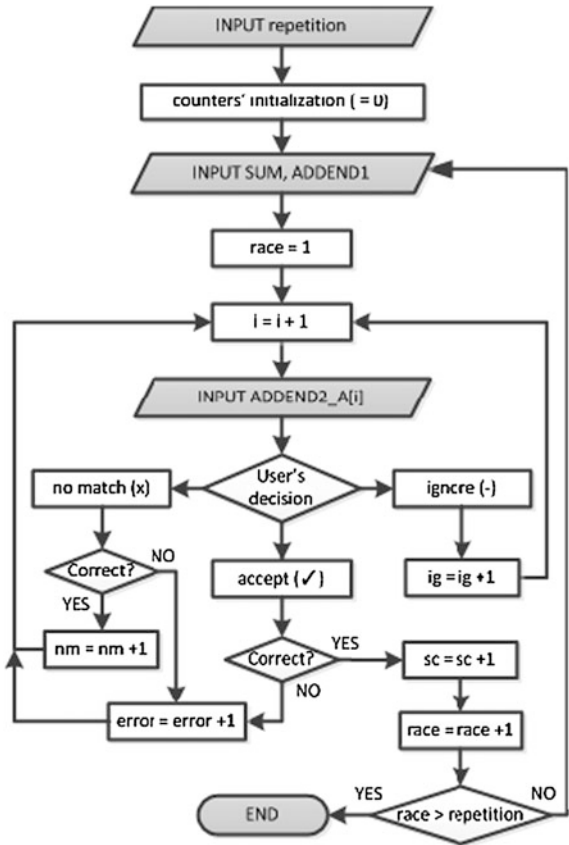


After stage 1, a one-dimensional array is created, i.e. the ADDEND2_A(i) with nine items, which will be filled with the nine alternative ADDEND2 values. First of all, with a random variable the position (pos) is selected, and the unique correct ADDEND2 is placed in position in the array (stage 2).

In the next stage (stage 3) the remaining items of the array ADDEND2_A[i] and random variables are set so that all are different.

The next flowchart presents the iterative process of the control and data placement in the array ADDEND2_A[i] (Fig. 36.6).

Fig. 36.7 The generic flowchart of the application



All three stages mentioned above are repeated as many times as the variable that controls the iterations of the additions (scroll bar “Repetitions” in the window of the application’s configuration of the operational parameters) determines. For this purpose an appropriate array (Repetition_Array) is created, with rows equal to the number of repetitions and items in every row: the SUM, the ADDEND1 and the nine items of the array ADDEND2_A[i].

The application then runs as presented in the flowchart of Fig. 36.7, acquiring every time all the required information from the Repetition_Array. Figure 36.7 shows the generic flowchart of the application for the user interface.

The application’s development was based on the linear model with multiple prototypes [15] and the programming language used is Microsoft Visual Studio.

36.2.4 Design and Esthetic Issues

In previous software development [12–14] it was noted that although the construction of numbers' images is a lengthy process it can be avoided. We noticed, however, that unlike the motion of a picture-object, the movement of a label-object is accompanied by flickering. In TNR application we decided on the automatic creation of numbers with label-objects, which are placed into picture-objects, for smooth movement and flicker-free motion. This may reduce the esthetic and the potential of custom pictures but on the other hand it also decreases the difficulty of implementation and increases the autonomy of the application and the ability to adapt to new circumstances and possible extensions.

We also made an attempt to insure that the integrated sounds enhanced the functionality without discouraging users but highlighted certain crucial situations.

Our effort was focused on labeling at least two points: when every race starts and when the ADENND2_A[i] stops between ADDEND1 and SUM, when the clock for user response starts ticking.

To avoid misleading visual messages the colors used were pale, except for the areas where we wanted to attract the attention of the user.

36.2.5 Application's Evaluation

The development of the application was performed through continuous tests and formative evaluation by the members of the research group of the Computers and Educational Technology Laboratory (CETL). A small sample of three primary education students in the sixth grade was also used. This sample was excluded from the final evaluation of the application.

Gradually, until the application was finalized, a few prototypes (of the application) were produced and examined by the evaluators, who implemented parts of the predefined requirements [15–17].

Subsequently, the application was pilot-tested with a summative evaluation, whose results will be analyzed in detail in the next section.

36.3 Evaluation Results

36.3.1 Methodology

Overall, 38 sixth-grade primary school children were selected for this study. The participants were tested individually in the computer laboratory. Data collection was co-ordinated by a single person in order to avoid any possible experimenter effects during both the treatment and the data collection phases. The main

Table 36.1 Mean Reaction Time Scores (In Seconds)

| Response | 2digit addends | 3digit addends |
|-------------|------------------------------|-----------------|
| No-Error | 0.4754 (0.2588) ^a | 0.5205 (0.2226) |
| No-Success | 0.4617 (0.2324) | 0.5045 (0.2117) |
| Yes-Error | 0.4614 (0.2440) | 0.6205 (0.2693) |
| Yes-Success | 0.4697 (0.2404) | 0.4731 (0.2006) |

^a Standard deviation is given in parentheses

experimental procedure was preceded by a pilot test. Results led to minor revisions of the calibration settings. Typical desktop personal computers with Microsoft Windows 7 were used for collecting the data.

A pre-test/post-test design was administrated for assessing children’s simple mathematical calculation skills. Two sets of stimuli were used in the pre-test/post-test session. In the first set participants were asked to calculate an array of 15 double-digit addends addition tasks, whilst the second set constituted 15 three-digit addends. Both sets of tasks were visually presented on a whiteboard. A fixation point (1 s) preceded the stimuli and participants were asked to determine whether the presented sums were correct or wrong and put their answers on a separate answer sheet. Stimuli disappeared after 60 s.

In the main experimental procedure participants were requested to complete a 10-stage session (approx. 60 min) for both the double-digit and the three-digit addends. Individual scores were calculated for each participant. The post-test followed the main experimental procedure. A two-week interval between pre- and post-test evaluation was allowed to insure the minimization of any warm-up effects.

36.3.2 Results

Table 36.1 presents the mean reaction time scores (in seconds) for both double- and three-digit addends. Children’s responses were classified in four categories. Thus, the **No-Error** category contains the negative answers in the case where the proposed sum was correct. When the proposed sum was incorrect and participants correctly selected the “X” button, answers were allocated to the **No-Success** category. The **Yes-Error** category denotes answers where the proposed sum was incorrect, but participants failed to identify it and pressed the “✓” button, whereas the **Yes-Success category** denotes answers where participants correctly identified the proposed sum as correct.

The data were subjected to a univariate ANOVA with *Reaction Time* as the dependent variable. *Response* (No-Error, No-Success, Yes-Error, Yes-Success) and *Addends* (two-digit, three-digit) were used as the fixed factors. Any significant differences between *Responses* were analyzed with Sheffé post hoc tests.

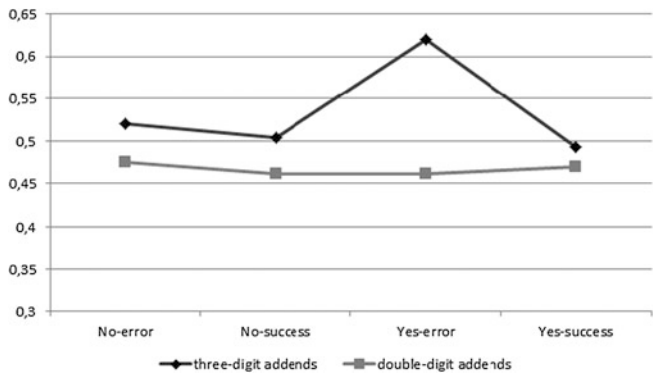


Fig. 36.8 Mean reaction time scores for each response category (in seconds)

Simple main effects analysis revealed significant effects of *Responses* [$F_{(7,3834)} = 4.23, p < 0.01$] and *Addends* [$F_{(1,3834)} = 27.57, p < 0.001$] on *Reaction Time* scores. There was also a significant interaction between the effects of *Responses* and *Addends* on *Reaction Time* scores [$F_{(3,3834)} = 4.71, p < 0.005$]. The post hoc multiple comparisons were conducted by applying a Sheffé correction ($p < 0.05$). The analysis unveiled significant differences between the **No-Success** and **No-Error** responses in the double-digit set ($p < 0.05$), as well as between **Yes-Success** and **Yes-Error** in the three-digit set ($p < 0.001$).

Figure 36.8 illustrates the mean reaction time scores for each set of addends. To evaluate the TRN application a pre-test/post-test design was selected. Our main focus was on assessing both usability issues and the application’s potential, if any, for enhancing children’s mathematical skills. Table 36.2 displays the mean accuracy scores in the pre- and post-test assessments.

As Fig. 36.9 depicts, children achieved higher scores in the post-test evaluation session for both double- and three-digit addends. Mean differences were calculated by using separate paired-samples *t*-tests. For both the addition of double-digit addends and three-digit addends tasks the analysis revealed significant differences [$t(569) = -6.46; p < 0.001$] and [$t(569) = -6.15; p < 0.001$], respectively.

After the main experimental procedure participants were asked to complete a brief structured non-disguised questionnaire. The questions mainly concerned the application’s ease of use, esthetics and effectiveness.

With regard to the application’s ease of use, participants were of the opinion that there were no difficulties regarding its use, as illustrated in Table 36.3.

Even if the students, according to them at least, did not find the use of the application difficult, we performed an additional analysis so that the distinctive features of TNR could be investigated at greater length. The analysis revealed statistically significant differences between the factors that caused the most difficulties while the application was in use (Friedman’s $\chi^2 = 45.03; df = 3; p < 0.001$). The mean rankings are presented in Table 36.4. The speed with which the numbers move seems to be the most problematic factor.

Table 36.2 Mean accuracy scores in the pre- and post-tests (%)

| Mathematical operation | Pre-test | Post-test |
|---------------------------------|----------------------|---------------|
| Addition (double-digit addends) | 59 (49) ^a | 68.42 (46.52) |
| Addition (three-digit addends) | 56 (50) | 69.12 (46.24) |

^a Standard deviation is given in parentheses

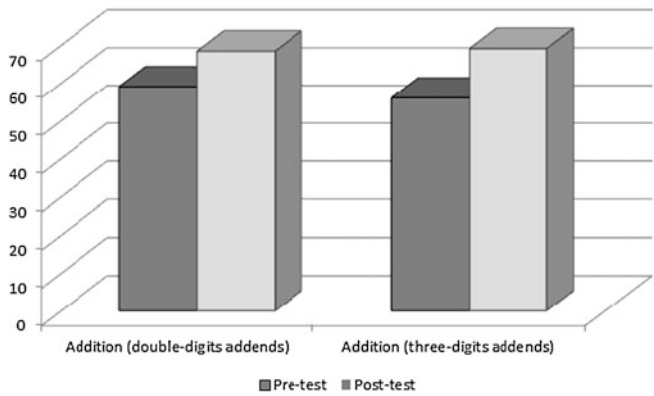


Fig. 36.9 Mean performance scores for pre- and post-test assessments (%)

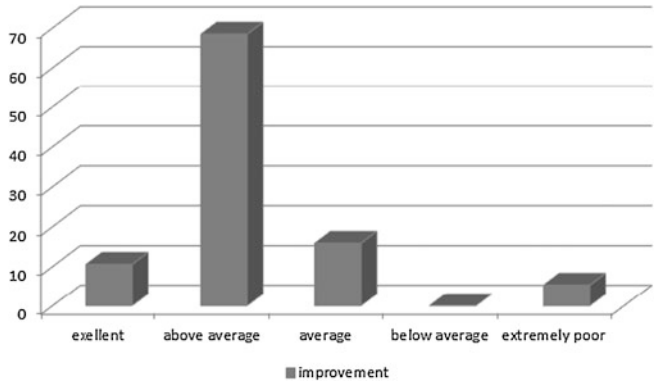


Fig. 36.10 TNR's effectiveness (%)

TNR's attractive features, as reported by children, were analyzed with a Friedman's test. The analysis revealed statistically significant differences ($\chi^2 = 48.57$; $df = 3$; $p < 0.001$). Table 36.5 shows the mean ranking for every factor. The most attractive factor for the participants was the ease of use.

A final analysis concerned whether and to what extent participants believed that TNR improved their performance while executing arithmetical calculations (Fig. 36.10). A Chi-square test of goodness-of-fit was performed to determine the degree to which TNR improved participants' mental addition skills. The effectiveness was not equally distributed in the population ($\chi^2 = 19.53$; $df = 3$; $p < 0.001$).

Table 36.3 Frequency and Percentage of Answers about the Ease of Use

| Answer | Frequency | Percentage |
|----------------|-----------|------------|
| Very easy | 16 | 42.1 |
| Easy | 14 | 36.8 |
| Neutral | 8 | 21.1 |
| Difficult | 0 | 0.0 |
| Very difficult | 0 | 0.0 |
| Total | 38 | 100.0 |

Table 36.4 Rankings for Ease of Use

| Factor | Mean ranking |
|----------------------|--------------|
| Mathematical task | 2.26 |
| Mouse use | 3.68 |
| Speed | 1.84 |
| Attention detachment | 2.21 |

Table 36.5 Rankings for Attractiveness

| Factor | Mean ranking |
|-------------|--------------|
| Score | 3.0 |
| Movement | 2.37 |
| Colors | 3.26 |
| Ease of use | 1.37 |

36.4 Discussion and Conclusions

This paper discusses development and evaluation issues concerning a mathematical mini-game entitled “The Numbers’ Race—TNR.” The TNR application was developed primarily as an alternative educational tool intended to support the current mathematical teaching approaches. Like most educational mini-games, TNR’s objective is to exploit children’s tendency to play in order to get them engaged in different mental mathematical calculations.

The application described so far is comprised of two separate modules: addition and subtraction. In this study we focused on the results of the addition. It is worth noting that the design of the application allows the integration of addition and subtraction of decimal numbers.

During the formative evaluation several programming and operational features improved so much as to meet the theoretical background for developing educational software, such as the user interface design and appearance, usability, motivational aspects and ease of use [16].

To evaluate the TNR application a multi-stage process was selected. The pre-test/post-test assessment was followed by a questionnaire designed to gather information on both the overall performance of the application and its specific components regarding ease of use, esthetics, etc.

Data analysis revealed differences between double- and three-digit addends in terms of both accuracy percentages and reaction time scores. Although the focus of this paper is not the cognitive factors that lie behind mental computation tasks, it is important to note that the analysis revealed significant variation between the different response categories. In the double-digit addends addition tasks differences were observed between the No-Error and No-Success responses, and the same pattern of results emerged between the Yes-Error and Yes-Success categories.

The pre-post assessments clearly showed substantial improvement in mental computation skills both for double- and three-digit addends between the initial and the final tryout. It is important to note that the effectiveness of the TNR application was also pointed to by the participants, as the questionnaire analysis revealed.

Finally, participants found the TNR application to be user-friendly and easy. It should be noted at this point that when children were asked to select the most attractive factor the ease of use received the highest ranking.

Our results suggest that the implementation of ICT applications in mainstream teaching practice may offer significant positive effects and seem to be in line with previous research on the field [12–14]. Future work includes the development of the TNR application with Java for use with mobile appliances.

References

1. Eliens A, Ruttkay Z (2009) Record, replay & reflect: a framework for serious gameplay. In: Proceedings of *EUROMEDIA 2009*, Brugge (Belgium), 2009
2. Lowery VN (2003) Assessment insights from the classroom. *Math Educator* 13(1):15–21
3. Smith PA, Sanchez A (2010) Mini-Games with Major Impacts. In Cannon-Bowers J, Bowers C (eds), *Serious game design and development: technologies for training and learning*. ICI Global, USA
4. Piaget J (1970) *Genetic epistemology*. W. W. Norton and Company, New York
5. Papert S (1993) *The children's machine: rethinking school in the age of the computer*. Basic Books, New York
6. Jong MSY, Shang J, Lee F, Lee JHM (2010) Constructivist Learning through Computer Gaming, doi:[10.4018/978-1-60566-934-2.ch014](https://doi.org/10.4018/978-1-60566-934-2.ch014)
7. Bruner J (1960) *The process of education*. Harvard University Press, Cambridge
8. Prensky M (2001) *Digital game-based learning*, McGraw-Hill, New York
9. Randel JM, Morris BA, Wetzel CD, Whitehill BV (1992) The effectiveness of games for educational purposes: a review of recent research. *Simul Gaming* 23(3):261–276
10. Mitchell A, Savill-Smith C (2004) The use of computer and video games for learning: A review of the literature. Accessed 12/8/2011 from Learning and Skills Development Agency www.LSDA.org.uk
11. Kickmeier-Rust MD (2009) Talking digital educational games. In: Kickmeier-Rust MD (ed.), *Proceedings of the 1st international open workshop on intelligent personalization and adaptation in digital educational games*, Graz, Austria pp 55–66 14 Oct 2009
12. Panagiotakopoulos C (2011) Applying a conceptual mini game for supporting simple mathematical calculation skills: students' perceptions and considerations. *World J Education* 1(1):3–14
13. Illanas AI, Gallego DF, Satorre CR, Llorens LF (2011) Conceptual mini-games for learning, IATED international technology, education and development conference, Valencia

- (Spain), 2008, Accessed 12 January 2011, from <http://rua.ua.es/dspace/bitstream/10045/8495/1/illanas08conceptual.pdf>
14. Panagiotakopoulos C, Sarris M, Koleza E (2010) Playing with numbers: development issues and evaluation results of a computer game for primary school students. In: Proceedings of international joint conferences on computer, information, and systems sciences, and engineering, 3–12 Dec 2010
 15. Pfleeger SL (2010) Software engineering: theory and practice. Prentice-Hall, New Jersey
 16. Panagiotakopoulos C, Pierrakeas C, Pintelas P Educational software design, Hellenic Open University Publications [in Greek], Patras, Greece
 17. Sommerville I (2004) Software Engineering, 7th Edition, Addison-Wesley, Reading

Chapter 37

Visual Cryptography Based on Optical Image Projection

Rita Palivonaite, Algiment Aleksa and Minvydas Ragulskis

Abstract A visual cryptography scheme based on optical image projection is proposed in this paper. Initially the secret image is split into two shares. Then, such digital images are constructed in share's planes that their projections in the projection screen would correspond to each of the appropriate shares. Geometrical parameters describing the location of shares' planes and focus points of projectors are additional security parameters of the encoded image. Direct overlapping of the reconstructed shares does not leak any information on the encrypted image. The original image can be interpreted by a naked eye when appropriate projectors are placed at predefined locations of the geometrical setup.

37.1 Introduction

Visual cryptography is a cryptographic technique which allows visual information (pictures, text, etc.) to be encrypted in such a way that the decryption can be performed by the human visual system, without the aid of computers. Visual cryptography was pioneered by Naor and Shamir in 1994 [1]. They demonstrated a

R. Palivonaite (✉) · A. Aleksa · M. Ragulskis
Research Group for Mathematical and Numerical Analysis of Dynamical Systems,
Department of Mathematical Research in Systems, Kaunas University of Technology,
Studentu 50-222, LT-51638 Kaunas, Lithuania
e-mail: rita.palivonaite@ktu.lt

A. Aleksa
e-mail: algiment.aleksa@ktu.lt

M. Ragulskis
e-mail: minvydas.ragulskis@ktu.lt

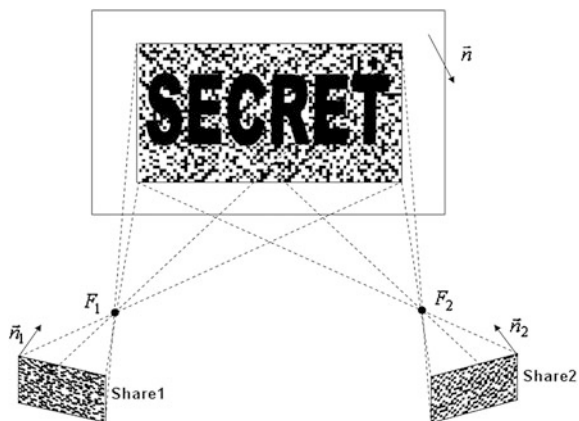
visual secret sharing scheme, where an image was broken up into n shares so that only someone with all n shares could decrypt the image, while any $n-1$ shares revealed no information about the original image. Each share was printed on a separate transparency, and decryption was performed by overlaying the shares. When all n shares were overlaid, the original image would appear.

Since 1994, many advances in visual cryptography have been done. Extended visual cryptography is presented in [2]. Four—share visual cryptography scheme for color images is proposed in [3]. A Visual Cryptography (VC)-based system for sharing multiple secret images is developed in [4]. Visual cryptography schemes are defined and analyzed for grey level images whose pixels have g grey levels ranging from 0 to 1 are presented in [5]. Colored visual cryptography without color darkening is proposed in [6]. Cheating prevention in visual cryptography is developed in [7]. Visual cryptography schemes with optimal pixel expansion are introduced in [8]. Image encryption by random grids is presented in [9]. A new method of producing multicolored share images based on visual cryptography is proposed in [10]. A novel technique named halftone visual cryptography is proposed to achieve visual cryptography via halftoning [11]. Colored visual cryptography scheme based on additive color mixing is presented in [12]. The best pixel expansion of various models of visual cryptography schemes is investigated in [13]. A new definition of the contrast of the visual cryptography is proposed in [14]. General construction of extended visual cryptography schemes is proposed in [15]. A high contrast and capacity efficient visual cryptography scheme for the encryption of multiple secret images is introduced in [16]. Image encryption by multiple random grids is presented [17].

We propose a modification of the classical visual cryptography scheme when each share has to be projected on the screen. We exploit the principle of the projection moiré technique [18] when an image projected on to the projection plane undergoes non-affine transformations (if only the projection angle is not perpendicular to the projection plane). The proposed method requires n projectors (if the original image is broken into n shares). Each of the projectors must project its share on the screen at the predefined geometrical location (projection angles for each projector can be different). Without loss of generality we will describe the method when the original image is broken into 2 shares only.

Each share is constructed in such a way that its projection (at strictly predefined geometrical parameters of the projector) would result into a projected image of a share which can be used to visualize the original secret image. Someone who has all shares can decrypt the secret image only if he knows how to project all these shares. Direct overlaying of n transparent shares (or projection of at least one of the shares at a wrong angle) would not leak any information about the secret image. Such an encryption technique can be considered as a visual cryptography scheme with additional security protection.

Fig. 37.1 A schematic diagram illustrating the principle of visual cryptography based on the projection technique



37.2 Description of the Projection Technique

As mentioned previously, we will use 2 shares to illustrate the proposed visual cryptography method based on projection techniques. The basic principle of the method is illustrated in Fig. 37.1. Two shares are projected on the projection plane; every share is located at a different position in the 3D space. Coordinates of focal points and the geometrical location of each share determine unique locations of each of projectors. The secret image appears on the projection plane when both shares are projected appropriately. It should be noted that Fig. 37.1 is only a schematic diagram. We do not show geometrical deformations of the projected rectangular images; exact geometrical locations of the projection plane and two shares are determined not only by their normal vectors. In fact, one has to solve an inverse problem of an image construction. One has to construct a share given a structure of the projected share and geometrical parameters of the projector.

Initially we assume that the equation of the projection plane is $z = 0$; $n = (0; 0; 1)$. As mentioned previously, a projected image on the projection plane must form a matrix of dots (it is assumed that a pixel is smaller object than a dot). Since we consider a classical visual cryptography scheme, every element of the projected matrix can be described as

$$M_1(i, j) \in \{0; 1\}; \quad i = 1, 2, \dots, r_1; \quad j = 1, 2, \dots, r_2 \quad (37.1)$$

where M_1 is the projection of the Share 1; r_1 and r_2 define the resolution of the visual cryptography scheme. The numerical value 0 corresponds to the black color; 1 corresponds to the white color (all intermediate values would correspond to appropriate grayscale colors). Similarly, both shares are also represented as matrixes of dots.

Lets assume that the origin of the 3D frame is a point $O(0; 0; 0)$; coordinates of the focus point are $F_1(f_x; f_y; f_z)$; $n_1 = (n_x; n_y; n_z)$; the equation of the first share's

plane is $n_x x + n_y y + n_z z = p$ (p is such that the focus point is between the share and the projection plane).

The first step is the selection of a local 2D frame in the share's plane. This is necessary because a share needs not only to be placed in a correct plane, but also rotated around its normal vector up to a correct angle. Initially, the origin of the 2D local frame O_1 is set as an intersection point between the line $F_1 O$ and the share's plane: $O_1(-t_O \cdot f_x + f_x; -t_O \cdot f_y + f_y; -t_O \cdot f_z + f_z)$, where $t_O = \frac{p - n_x f_x - n_y f_y - n_z f_z}{-n_x f_x - n_y f_y - n_z f_z}$. Next, images of points $A(1; 0; 0)$ and $B(0; 1; 0)$ are computed in the share's plane and denoted as A_1 and B_1 . Then, base vectors of the local 2D frame in the share's plane are denoted as:

$$\mathbf{i}_1 = \frac{\mathbf{O}_1 \mathbf{A}_1}{|\mathbf{O}_1 \mathbf{A}_1|}; \mathbf{j}_1 = \frac{\mathbf{O}_1 \mathbf{B}_1}{|\mathbf{O}_1 \mathbf{B}_1|} \quad (37.2)$$

Elementary transformations yield:

$$\begin{aligned} i_1 &= \frac{(t_A(1 - f_x) + t_O f_x; -t_A f_y + t_O f_y; -t_A f_z + t_O f_z)}{\sqrt{(t_A(1 - f_x) + t_O f_x)^2 + (-t_A f_y + t_O f_y)^2 + (-t_A f_z + t_O f_z)^2}}; \\ j_1 &= \frac{(-t_B f_x + t_O f_x; t_B(1 - f_y) + t_O f_y; -t_B f_z + t_O f_z)}{\sqrt{(-t_B f_x + t_O f_x)^2 + (t_B(1 - f_y) + t_O f_y)^2 + (-t_B f_z + t_O f_z)^2}}; \end{aligned} \quad (37.3)$$

where,

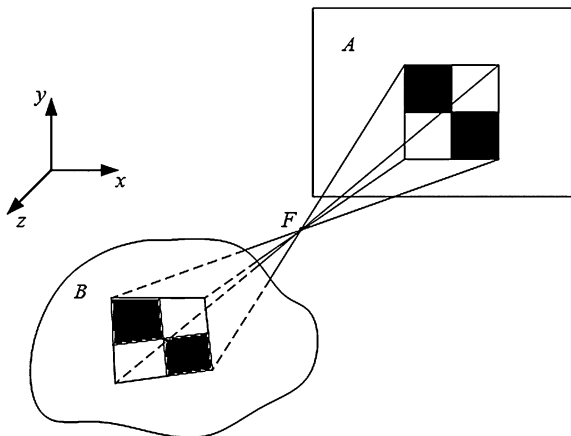
$$\begin{aligned} t_A &= \frac{p - n_x f_x - n_y f_y - n_z f_z}{n_x(1 - f_x) - n_y f_y - n_z f_z}; \\ t_B &= \frac{p - n_x f_x - n_y f_y - n_z f_z}{-n_x f_x + n_y(1 - f_y) - n_z f_z} \end{aligned}$$

It can be noted that vectors \mathbf{i}_1 and \mathbf{j}_1 are not necessarily orthogonal. The next step is the construction of an image of a dot from the projection plane in the share's plane. In general, computation of coordinates of an image point in a 2D local frame in the share's plane involves solution of a linear algebraic system of equations. Let coordinates of a point in the projection plane are $C(x; y; 0)$. Then, coordinates of its image point in the share's plane are $C_1(t_C(x - f_x) + f_x; t_C(y - f_y) + f_y; -t_C f_z + f_z)$ where $t_C = \frac{p - n_x f_x - n_y f_y - n_z f_z}{n_x(x - f_x) + n_y(y - f_y) - n_z f_z}$. Now, a vector $\mathbf{O}_1 \mathbf{C}_1$ has to be expressed in a linear combination of base vectors \mathbf{i}_1 and \mathbf{j}_1 :

$$\mathbf{O}_1 \mathbf{C}_1 = c_x \cdot \mathbf{i}_1 + c_y \cdot \mathbf{j}_1 \quad (37.4)$$

where c_x and c_y are 2D coordinates of the point C_1 in the share's plane. It can be noted that Eq. (37.4) produces 3 linear algebraic equations. One of these equations can be omitted (or used as a criterion for checking if the point C_1 is exactly mapped on the share's plane). Other two equations can be used for determination of c_x and c_y (we omit details for brevity):

Fig. 37.2 A schematic diagram representing of the construction of the images in the share plane; A stands for the projection plane; F is the focus point; B is the share plane. Note that only one share plane is shown



$$\begin{cases} i_x c_x + j_x c_y = t_C(x - f_x) + t_O f_x; \\ i_y c_x + j_y c_y = t_C(y - f_y) + t_O f_y. \end{cases} \quad (37.5)$$

Reconstruction of a point's local coordinates in the second share's plane is analogous to the procedure described above. As mentioned previously, a dot can be comprised from many pixels (this is determined by the resolution of the projected image). A schematic diagram of the image computation process is presented in Fig. 37.2. A set of four black and white dots in the projection plane A is shown at the top of Fig. 37.2; the corresponding image is projected through the focus point F into the share plane B . Note that only one share plane is illustrated in Fig. 37.2; the splitting rule is a standard random scheme used in classical visual cryptography [1].

37.3 The Construction of Images in the Share Plane

The construction of digital image in the share plane is not a straightforward task simply due to fact that the image in the share plane B is skewed in respect of the orthogonal matrix of pixels in the share image (Fig. 37.2). It is clear that the proposed system of visual cryptography will not work if the size of a pixel in the share plane is comparable to the size of a projected dot in the projection plane.

The algorithm for the computation of grayscale levels of pixels in the share plane is illustrated in Fig. 37.3. The shape of the inclined grid of dots in the share plane is illustrated in Fig. 37.3a. Note that the projected grid in the projection plane is rectangular and corresponds to the position of pre-defined dots (Fig. 37.2). The size of pixels in the share plane is illustrated by the dashed grid in Fig. 37.3a. The computational procedure for the reconstruction of the grayscale level at the pixel in the i th row and the j th column in the share plane is explained in Fig. 37.3b.

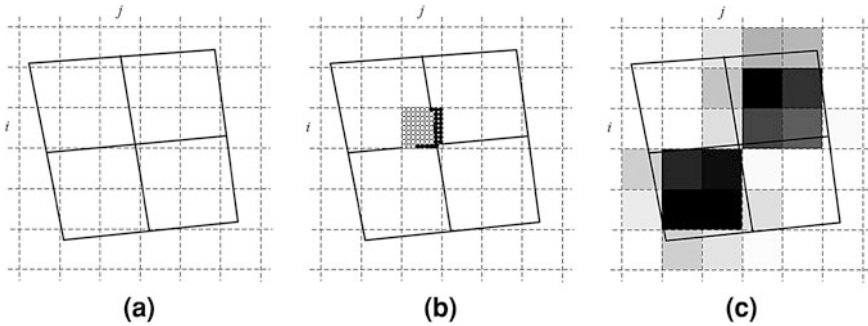


Fig. 37.3 A schematic diagram representing of the reconstruction of the grayscale levels of pixels in the share plane; the size of pixels in the share plane is denoted by dashed grid. The inclined grid corresponding to the projected image in the projected plane is shown in (a). The algorithm for the computation of the grayscale level at the i - j th pixel is illustrated in (b). The reconstructed image in the share plane is shown in (c). Note that only one share plane is shown

We cover the surface of the i - j th pixel in the share plane by a rectangular matrix of points. Every point of this matrix is projected to the projection plane.

A point is marked by an empty circle in Fig. 37.3b if the coordinates of the projected point in the projection plane correspond to a white dot. Analogously, a point is marked by a black circle if the projected point is located inside a black dot in the projection plane.

Now, the grayscale level g_{ij} at the i - j th pixel is computed according to the equation:

$$g_{ij} = \text{round}\left(255 \frac{w_{ij}}{n}\right) \quad (37.6)$$

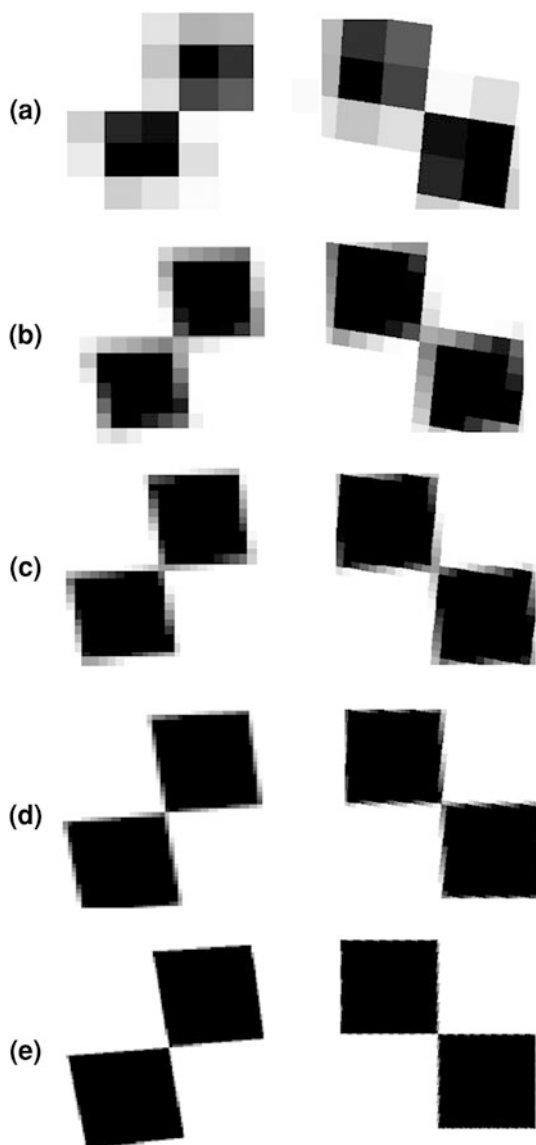
where w_{ij} is the number of white points and n is the total number of points in the area occupied by the i - j th pixel in the share plane.

The reconstructed grayscale digital image is shown in Fig. 37.3c. It is clear that the quality of the projected image is directly related to the size of a pixel in the share plane.

The effect of the pixel size in the share plane to the quality of the image in both share planes is illustrated in Fig. 37.4. We show the digital images in the share plane (left columns in Fig. 37.4) and the projected images to the projection plane (right columns in Fig. 37.4). The quality of digital images in the share plane depends on the ratio between the size of the dot in the projection plane and the size of the pixel in the share plane; A stands for the ratio 2:1; B stands for the ratio 5:1; C —10:1; D —20:1 and E —50:1.

It is clear that it is pointless to discuss a visual cryptography scheme based on projected images if one cannot reproduce a realistic image in the projection plane. Our computations show that at least 50×50 pixels in the share plane should correspond to one dot in the projection plane. Moreover, the quality of the

Fig. 37.4 The quality of digital image in the share plane depends on the ratio between the size of the dot in the projection plane and the size of the pixel in the share plane. The left column shows the image in the share plane; the right column—the projected image in the projection plane. The angle of projection $s = 0.01$



projected image depends of the geometrical set-up. The more inclined is the angle of the projection the higher must be the ratio between the size of the dot in the projection plane and the size of the pixel in the share plane.

It is possible to assess the quality of the projected image by comparing the original image (Fig. 37.2a) and the projected image (images in the right column in Fig. 37.4). We use root mean square error estimate (RMSE) and plot these errors as a function from the angle of projection s (Fig. 37.5).

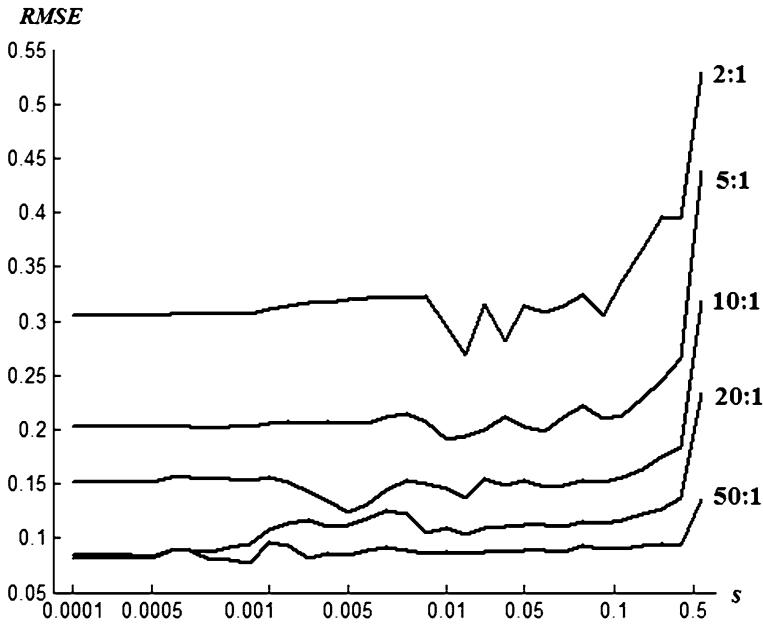


Fig. 37.5 The relationship among the ratio between the size of the dot in the projection plane and the size of the pixel in the share plane, the angle of projection s and the RMSE of the projected image

It can be seen that RMSE errors depend both on the ratio between the sizes of dots but also on the projection angle s . The general recommendation is to select larger dots in the projection plane if the projection angle is rather inclined.

37.4 Computational Experiments

Initially we use a classical cryptography scheme to split a digital image (three letters KTU) into two shares. Both shares are shown in Fig. 37.6a and Fig. 37.6b. Direct geometrical superposition yields an image which can be interpreted by a naked eye (Fig. 37.6c). It can be noted that inaccurate superposition of two shares prevents visual interpretation of the encoded image.

The next step is the construction of such digital images in shares' planes that their projected images would coincide with original shares shown in Fig. 37.6a and Fig. 37.6b. This is an inverse problem of image construction described in the previous section.

Following parameters of the geometrical setup were selected for computational experiments: $F_1 = (0; 0; 10)$; $F_2 = (0; 4; 10)$; $n_1 = (-s; s; -1)$; $n_2 = (s; -s; -1)$; distance between the plane of the share 1 and F_1 is equal to 10; distance between the plane of the share 2 and F_2 is also equal to 10; the width of the secret image is

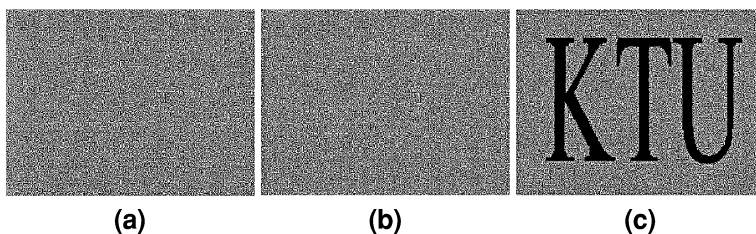


Fig. 37.6 Illustration of a classical visual cryptography: **a**, **b** show two shares in the projection plane; **c** exact geometrical superposition of both shares in the projection plane produces the secret image

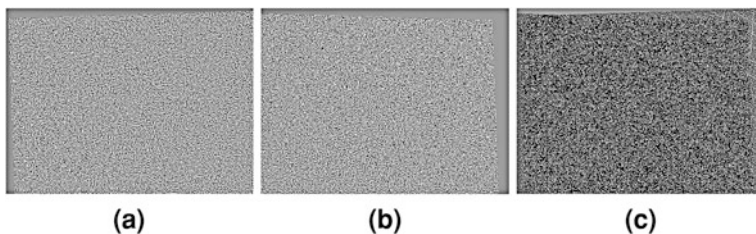


Fig. 37.7 Reconstructed images of share 1 (**a**) and share 2 (**b**); direct geometrical superposition of share 1 and share 2 does not reveal the secret image (**c**)

4 units; the height is 3 units. Reconstructed images of share 1 and share 2 at $s = 0.1$ are shown in Fig. 37.7a, b.

As mentioned previously, projections of share 1 and share 2 in the projection plane should correspond to digital images in Fig. 37.6a, b. But nonlinear transformations occurring during the process of projection and described in the previous section cause the appearance of parasitic moiré patterns in the superposed image in the projection plane (Fig. 37.8a, b).

Reconstructed images of shares in their planes can be considered as a next security level in a visual cryptography scheme. Direct superposition of these shares prevents visual interpretation of the encrypted image, which becomes observable only when shares' planes become almost parallel to the projection plane (Fig. 37.9c).

37.5 Concluding Remarks

A classical visual cryptography scheme is extended by introducing projection effects which distort original shares when they are projected on the surface of a projection plane. Such extensions can be considered as a next step in the security level of the image encryption. Both shares are constructed in such a way that their

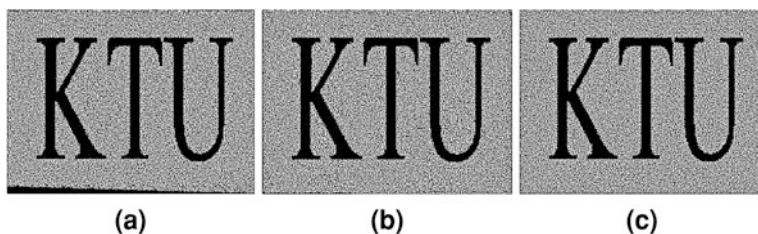


Fig. 37.8 Projected and superposed images of share 1 and share 2 in the projection plane; normal vectors of the shares' planes are $n_1 = (-s; s; -1)$; $n_2 = (s; -s; -1)$; **a** at $s = 0.1$; **b** at $s = 0.01$; **c** at $s = 0.0001$

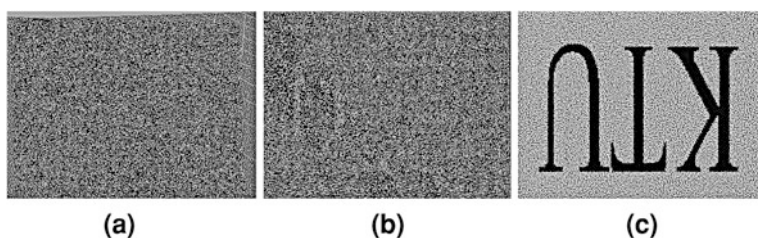


Fig. 37.9 Direct geometrical superposition of share 1 and share 2; **a** at $s = 0.1$; **b** at $s = 0.01$; **c** at $s = 0.0001$

projections would coincide with the original shares of the encoded image (at strictly defined geometrical parameters of the projection scheme). Distortions occurring during the construction of both shares damage the allocation of appropriate pixels and thus direct overlapping of both shares cannot leak any information about the encoded image.

We have used a simplified geometrical projection scheme. In practice (if a CCD projector is used to project an image on the screen) one should take care of distortions caused by non-ideal lenses. Also one should consider the effect of the depth of the projection's focus on the screen, especially if the projection angle is high and the projector is far from the screen.

The proposed image sharing scheme is still a visual cryptography scheme. Computational algorithms are necessary to construct the shares, but visualization does not require a computer; this is a completely visual process. But one needs to place two projectors with high accuracy in reference to the projection screen, instead of simply overlapping two shares.

Finally, it can be noted that the proposed scheme can be extended to an n shares scheme, halftone or even color visual projection cryptography schemes.

References

1. Naor M, Shamir A (1995) Visual cryptography. LNCS 950:1–12
2. Nakajima M, Yamaguchi Y (2002) Extended visual cryptography. In: Conference Proceedings 10th international conference on computer graphics, visualization and computer vision, vols I and II, University of West, Bohemia pp 303–310
3. Leung BW, Ng FY, Wong DS (2009) On the security of a visual cryptography scheme for color images. *Pattern Recogn* 42(5):929–940
4. Chen SK (2007) A visual cryptography based system for sharing multiple secret images. In: Proceedings of the 7th WSEAS international conference on signal processing, computational geometry and artificial vision (ISCGAV'-07) pp 113–118
5. Blundo C, De Santis A, Naor M (2000) Visual cryptography for grey level images. *Inform Process Lett* 75(6):255–259
6. Cimato S, De Prisco R, De Santis A (2007) Colored visual cryptography without color darkening. *Theoret Comput Sci* 374(1–3):261–276
7. Hu SM, Tzeng WG (2007) Cheating prevention in visual cryptography. *IEEE Trans Image Process* 16(1):36–45
8. Blundo C, Cimato S, De Santis A (2006) Visual cryptography schemes with optimal pixel expansion. *Theoret Comput Sci* 369(1):169–182
9. Shyu SH (2007) Image encryption by random grids. *Pattern Recogn* 40(3):1014–1031
10. Wu HC, Wang HC, Tsai CS (2006) Multiple image sharing based on colour visual cryptography. *Imaging Sci J* 54(3):164–177
11. Shou Z, Arce GR, Di Crescenzo G (2006) Halftone visual cryptography. *IEEE Trans Image Process* 15(8):2441–2453
12. Yang CN, Chen TS (2008) Colored visual cryptography scheme based on additive color mixing. *Pattern Recogn* 41(10):3114–3129
13. Hajiabolfassan H, Cheraghi A (2010) Bounds for visual cryptography schemes. *Discret Appl Math* 158(6):659–665
14. Liu F, Wu CK, Lin XJ () A new definition of the contrast of visual cryptography scheme. *Inform Process Lett* 110(7):241–246
15. Wang D, Yi F, Li X (2009) On general construction for extended visual cryptography schemes. *Pattern Recogn* 42(11):3071–3082
16. Lee KH, Chiu PL (2011) A high contrast and capacity efficient visual cryptography scheme for the encryption of multiple secret images. *Optics Commun* 284(12):2730–2741
17. Shyu SJ (2009) Image encryption by multiple random grids. *Pattern Recogn* 42(7):1582–1596
18. Kobayashi AS (1993) Handbook on experimental mechanics, 2nd edn. Bethel, SEM

Chapter 38

A New Service Offered by Digital Radio for Vehicle Drivers

Cabani Adnane and Mouzna Joseph

Abstract In this article, we present the advantages of digital radio and specially the T-DMB standard services for road users like TPEG and BIFS. We propose a new T-DMB service using BIFS for the rescue operation during the French red plan.

38.1 Introduction

During these last years, various services are offered to the motorists via the radio and mainly via the FM band. These services are included under the Radio Data System (RDS). Limitations due to the bandwidth (1.2 kbit/s) were felt quickly. With the emergence of the digital radio, the opportunity to offer richer services became possible. The bandwidth of digital radio is hundred times superior than RDS.

In France, it is in December 2007, that the Ministry of Culture and Communication published an order which fixes the regulatory framework of the deployment of the digital radio. The Terrestrial- Digital Media Broadcasting (T-DMB) standard was chosen for the broadcasting of the digital radio.

The purpose of the deployment of digital radio is to provide listeners with better sound quality, access to data associated with programs that complement the sound like album cover, the guest picture, information about the weather, etc.

C. Adnane (✉) · M. Joseph
ESIGELEC/IRSEEM, Saint-Etienne du Rouvray, France
e-mail: cabani@esigelec.fr

M. Joseph
e-mail: mouzna@esigelec.fr

In this paper, we present the contribution of digital radio and the various services that may be offered to road users.

This work was carried out under the project Radio NUMérique TERrestre (RANUTER). It is a research program that aims to design and prototype, then test and evaluate new services made possible by digital radio.

38.2 Analogical Radio and RDS

The Radio Data System (RDS) [1] was developed within the European Broadcasting Union during the 1980s. The RDS is standardized by the European Committee under the name EN 62106. This system can offer a range of useful services to motorists. We can cite the following ones:

- Traffic Program (TP): It indicates if the radio channel is likely to disseminate traffic information or not.
- Traffic Announcement (TA): It indicates if the radio channel can inform about the traffic.
- Enhanced Other Networks (EON): It performs automatic handover to TA. Once it is completed, the receiver returns to its previous sound channel.
- Traffic Message Channel (TMC): It informs with a text or a panel on the screen about the traffic (accidents, delays, etc.).

38.3 Digital Radio

38.3.1 *The Broadcasting Standards*

With the development of digital radio, richer services can be offered to motorists thanks to a larger bandwidth. The T-DMB permits a bit rate up to 1.5 Mbit/s. It offers a better sound quality with the Advanced Audio Coding (AAC), interactive services like slideshows and Binary Format for Scenes (BIFS). T-DMB uses Transport Protocol Expert Group (TPEG) standard for traffic alerts. Its advantage is to provide a richer alert content which is independent of the language of the country. This was not the case in RDS.

Digital radio is born with the Digital audio Broadcasting (DAB) standard [2]. The standard has evolved to DAB+ and then to T-DMB.

DAB uses the MPEG-2 Layer II compression. With this type of compression, it is possible to aggregate 6–9 programs using the same multiplex (on a bandwidth of 1.5 MHz). The number varies with the selected compression ratio.

The DAB+ uses the algorithm of MPEG-4 HE-AACv2 audio. It is possible to broadcast 12–18 channels via a multiplex.

The standard T-DMB has been specially designed for video broadcasting for handheld devices like smart phones. The encoding used is MPEG-4 HE-AACv4 for channel audio and MPEG-4 H.264 for the video. Only 2–3 television programs can be broadcast on the same multiplex. For the audio, it is possible to group 6–9 programs on the same multiplex. This standard also allows the streaming of BIFS.

Note that on the same DAB multiplex, it is possible to broadcast DAB, DAB+ and T-DMB. The choice of the broadcast standard does not impact the network design which remains the same for all audio broadcasting standards.

To listen to the digital radio, there are three different receiver profiles [3]:

- Profile 1—Standard Radio Receiver: It does not have a screen or a basic alphanumeric display.
- Profile 2—Medium Radio Receiver: It has a small color screen. They can display text and static images.
- Profile 3—Multimedia Receiver: It has large screens that can view videos.

For these three profiles, it is recommended that receivers can decode analog radio on AM and FM bands jointly with the digital radio.

We will focus on T-DMB services as it's the most updated and richest standard of digital radio.

38.3.2 Services on T-DMB

38.3.2.1 TPEG

TPEG is a standard included in T-DMB and used for traffic alerts [4].

The TPEG message composition is shown in Fig. 38.1. Every message (Application Event) is associated with a location (TPEG-location). The TPEG-location container indicates the GPS coordinates and additional location description (e.g. in a junction, middle of the street, near a fuel station, etc.). Every description is referenced in a table as a number so that the location message does not depend on the language. The language used is indicated in the default language code field.

The application Event container is composed of one of the next information messages:

- Road Traffic Message (TPEG-RTM): The purpose of this service is to send messages about road traffic: traffic information, accidents, congestion and travel time. The application is specified in ISO TS 18234-4.
- Public Transport Information (TPEG-PTI): Provides information about public transportation: maritime, urban (bus, metro, rail) and aviation. The application is specified in ISO TS 18234-5.

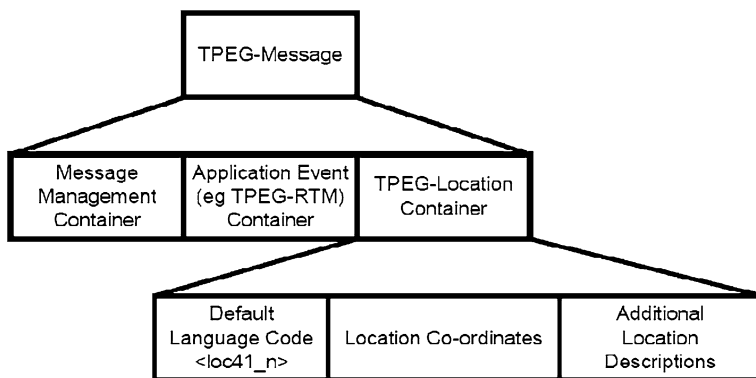


Fig. 38.1 TPEG message composition

- Parking Information (TPEG-PKI): This service provides information about parking: The number of vacancies, occupancy, etc. The application is specified in ISO TS 24530-5.
- Traffic Event Compact (TPEG-TEC): It's the equivalent of RDS-TMC. The main added value is the dynamic route guidance for GPS navigation system. The dynamic route guidance.
- Congestion and Travel-Time (TPEG-CTT): It gives information to drivers on congestion levels and journey times.
- Weather information for travelers (TPEG-WEA): It provides information on weather that may affect traffic conditions of roads.
- Traffic Flow and Prediction [5]: This application provides information on current traffic and future road network. For example: delay, speed, and travel times on sections of roads.
- The Wi-Fi hotspots (hotspot): It shows hotspots around the receiving point.
- Environment: It gives information on the air quality and index of pollution.

38.3.2.2 BIFS

BIFS is a binary format for two- or three-dimensional audiovisual content. It is based on VRML and part 11 of the MPEG-4 standard [6]. BIFS is an MPEG-4 scene description protocol composed of 3D geometric forms, text and videos. The standard describes the interaction between these objects and how to animate them. BIFS is a useful standard which permits mixing audio, video, animations and text in a single interactive MPEG4 file. Coding of animation and text uses vector representation permitting a better quality than text and animation included inside an MP4 video. Once coded, a BIFS application can be included in a video channel over the T-DMB multiplex.

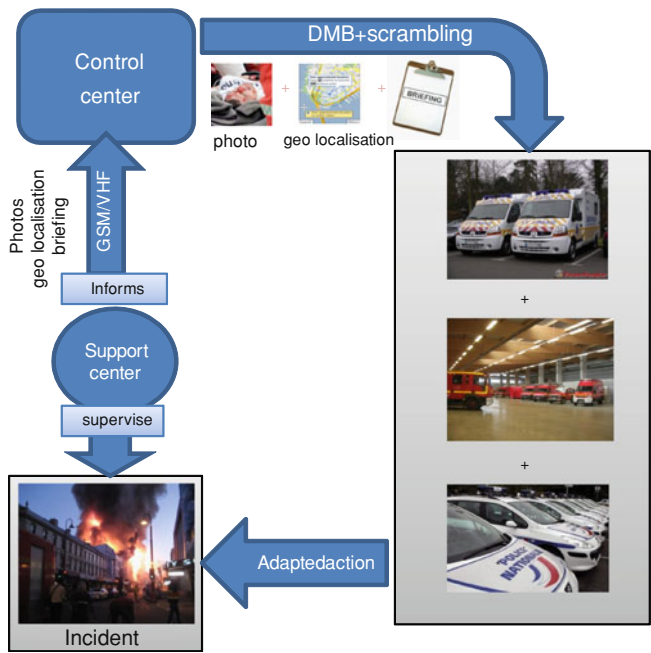


Fig. 38.2 The proposed service

In the next paragraph, we propose a new service on T-DMB using a BIFS application. This service is dedicated to organize rescue operations in the so-called French red plan.

38.3.3 The Proposed Service

Figure 38.2 shows the service that we have proposed to the rescue teams. An organization called Support Center (SC) establishes its headquarter near the incident which can be a building collapse, a fire or an explosion when a red plan is triggered. The support center guides the rescue operation but also provides information for the crews and for vehicles involved in the action. A Control Center (CC) filters the information (photos, geo localization, and briefing) that arrived from the SC. The CC generates a DMB signal containing the same information. It scrambles the data with a secret key known only by the vehicles concerned. This key is already stored in the receiver’s memory.

We implemented a BIFS application capable of being transported on the T-DMB Mean Service Channel (MSC), and an intuitive Graphical User Interface (GUI) that permits the CC to select the contents and generation for the BIFS application.

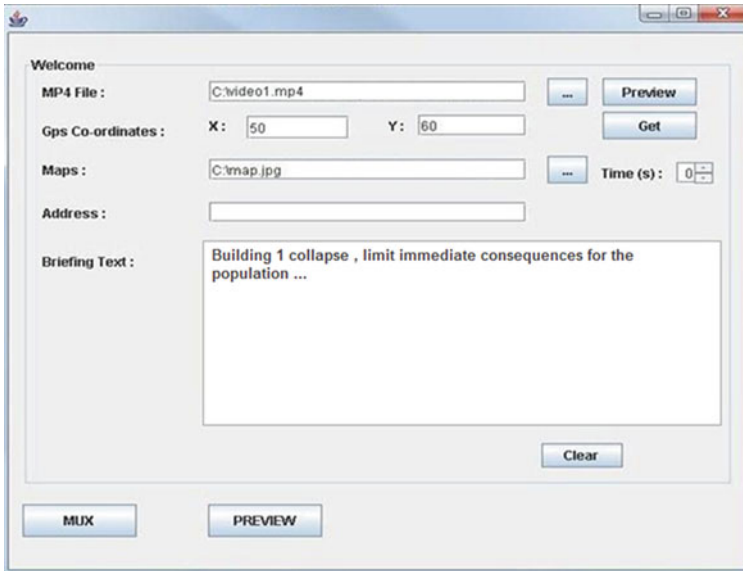


Fig. 38.3 Control Center interface

38.3.3.1 The Graphical User Interface

The GUI shown in Fig. 38.3 is developed using JAVA language. The user can choose a photo or a video as MP4 file, set GPS coordinates and choose a map like Google maps for the incident localization. He can also precise the address directly and type a briefing text, the button preview shows the BIFS application prior to its sending, and the button MUX mixes the inputs to a single output MP4 file. This JAVA based application, first, generates an XMT-A file containing the information elements (Video, audio, text and pictures), then, the XMT-A is transformed to an MP4 file which is sent to MSC channel.

The result of the generation is shown on Fig. 38.4. The output is a BIFS application read by the OSMO player [7].

38.3.3.2 Simulation of the T-DMB Scrambling

In order to provide a scrambled content, we used MATLAB © to simulate the scrambling process on

T-DMB. We conceived a T-DMB transmitter with the SIMULINK blocks. Figure 38.5 shows the composition of the conditional access scrambler which uses a real time triggered sub-system. This sub-system provides the Control Word (CW). An XOR operation is applied between the input and the CW to provide the scrambled data. The CW is built by a pseudo-random generator according to the standard Ref. [8].

Fig. 38.4 The BIFS application

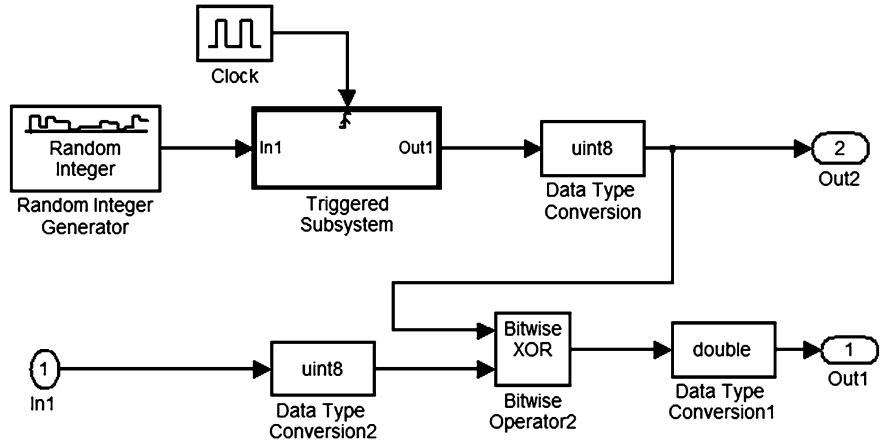
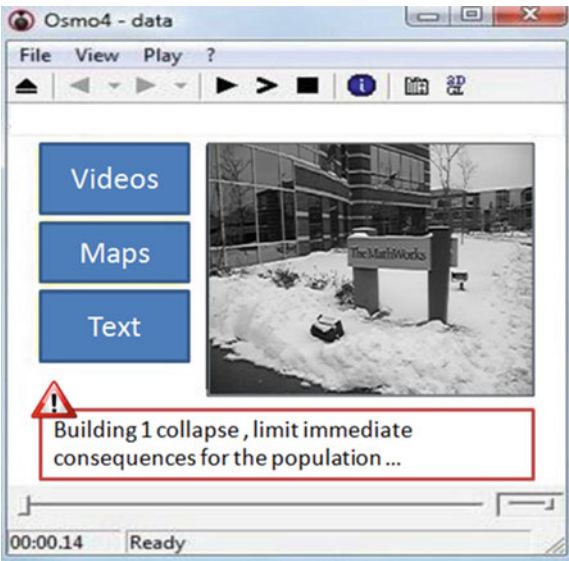
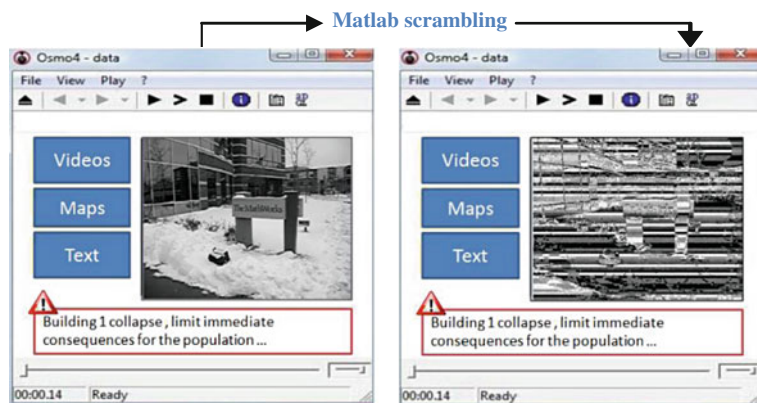


Fig. 38.5 The scrambling simulation process

The “IN1” source connects the input file to the scrambler. The file is the MP4 BIFS. Matlab extracts the picture or the video into an 8 bit element Matrix and processes the scrambling on the media.



38.4 Conclusion

In this article, we listed the added value of T-DMB as a standard chosen by France for the digital radio. We saw that TPEG is richer than RDS and that its language is independent. We then explained BIFS and proposed a service for the French red plan rescue. In this service, we developed the JAVA application for the Control Center and shown the result as a BIFS application. We simulated the T-DMB control access scrambling (CAS) to provide the content to the privileged rescue vehicles (firemen, police, medical staff).

References

1. Specification of the RDS standard: IEC 62106:1999
2. ETSI 300 401. Radio broadcasting systems; digital audio broadcasting (DAB) to mobile, portable and fixed receivers, European standard 2001–2005
3. EBU-R126. Digital radio broadcasting: common European digital radio profiles, Jan 2009
4. <http://www.tisa.org>, official website of the TMC and TPEG forum successor the traveller information services association (TISA)
5. <http://www.mobile-info.org/>
6. ISO/IEC 14496-11 (2005) Information technology—coding of audio-visual objects—part 11: scene description and application engine
7. http://gpac.sourceforge.net/tutorial/bifs_intro.htm
8. ETSI TS 102 367 v 1.2.1 (2006-01). Digital audio broadcasting (DAB); conditional access, European standard

Chapter 39

Separation of Concerns in Extensible Control Systems

Martin Rytter and Bo Nørregaard Jørgensen

Abstract The extensibility of non-trivial control systems is often constrained by unsatisfactory separation of concerns. Unfortunately, concerns frequently encountered in the control system domain are difficult to separate using domain independent approaches—e.g. aspects and other advise-based techniques. Thus, improved extensibility can only be achieved by inventing domain-specific software architectures for control systems that improve separation of concerns. In this paper, we analyze concerns emerging in a control system for industrial plant cultivation in greenhouses, and we present a software architecture that improves the separation of those concerns. The experience shared in the paper is the result of cooperation between software engineers, plant physiologists, and a control system vendor.

39.1 Introduction

An *extensible* system is a system that promotes the introduction of new functionality. Ideally, we would like *independent extensibility*—i.e. a situation where extensions can be combined without requiring a global integrity check [1]. Independent extensibility is particularly useful when a system is built from components that are independently developed.

M. Rytter (✉) · B. N. Jørgensen
The Maersk Mc-Kinney Møller Institute, University of Southern Denmark,
Campusvej 55, 5230, Odense M, Denmark
e-mail: mlrj@mmmi.sdu.dk

B. N. Jørgensen
e-mail: bnj@mmmi.sdu.dk

Extensibility is one of several good properties that tend to emerge when striving for *separation of concerns* [2]—in our opinion, perhaps the most important one. Other desirable properties emerging from separation of concerns include modular reasoning and smaller codebases.

Many works strive for improved separation of concerns by suggesting general approaches that are independent of the problem domain—e.g. novel contributions such as hyperspaces [3], aspects [4], open modules [5], and many more. While the pursuit for general solutions have contributed many novel approaches, and while such approaches certainly can appear to be very elegant, they also have built-in limitations that constrain extensibility.

Specifically, domain-independent approaches to separation of concerns tend to compromise independent extensibility. The fundamental source of the problem is that the combination of independently developed system components requires a global integrity check.

Examples of domain-independent techniques that compromise independent extensibility include AspectJ and Hyper/J [6].

In the case of Hyper/J [3], it is possible to integrate incompatible hyperslices by using a new hypermodule to define new versions of existing types. The problem with this approach is the need for *client migration*—whenever a new version of a type is defined, there may be clients that need to be migrated to use the new version. Ensuring that all relevant clients are migrated requires a global view of the system. In other words, Hyper/J does not satisfy the requirement for independent extensibility.

While Hyper/J promotes a copy-and-migrate approach to integrating new functionality, AspectJ [4] relies on *in-place modification* of existing types. The fundamental idea is to modify existing types so that new functionality is woven around the main program's structure. When weaving a single aspect with a base program, intimate knowledge of the base program is often required, to guarantee that no invariants are violated. A more serious problem emerges when multiple aspects are woven with the same base program. In this case, an aspect may not merely break invariants in the base code—it may also break any other aspect in the system with which it interacts. Therefore, the combination of independently developed aspects requires a global integrity check, and thus the requirement for independent extensibility is not met.

It might be tempting to suggest that aspects do not produce conflicts unless they modify shared state in the base program. On the contrary, the order in which aspects are woven into the code may greatly influence program behavior [7]. Some languages—e.g. AspectJ—provide means to control the order in which aspectual advices are executed. This might be used to solve some aspect interaction problems, but not all. And in any case, using this approach requires a global view of the system.

Many works have proposed ideas that strive to partly overcome the problems mentioned above. We see primarily three different kinds of approaches: first, it is possible to use tools such as IDE support [8] and aspect-oriented test practices [9] to help programmers manage aspect interaction—needless to say, this does not

remove the need for a global integrity check. Second, it is possible to create specialized composition mechanisms [10] or coordination aspects [11] that allow a composer to resolve certain conflicts—using such approaches also requires the composer to have global knowledge. Third, it is possible to introduce language features that constrain aspects so that the room for interaction is smaller [5, 12]—at best, such approaches help manage aspect interactions, but still, independent extensibility cannot be guaranteed.

While we acknowledge the novelty of many aspect-oriented approaches, we have no hope that aspects can facilitate extensibility and independent development [13].

Failing to see how domain-independent approaches to separation of concerns can improve extensibility, we suggest that more efforts are focused on less general solutions. Specifically, we believe that the best way to achieve improved extensibility is to discover domain-specific software architectures that support separation of concerns that are common in a particular domain. In other words, there is no general solution—thus, the problem of separating concerns must be solved once for each domain using component frameworks [14].

In this paper, we will investigate a domain where we find separation of concerns to be a particularly challenging task: non-trivial control systems. Here we consider a *control system* to be a system that controls other systems—e.g. an autopilot system in a car controls the car’s throttle system. We consider a control system to be *non-trivial* when more than one concern affects the control of a controlled subsystem—this will be illustrated in more detail in Sect. 39.2.

The main contribution of this paper is the presentation of a software architecture for non-trivial control systems. The architecture promotes separation of control-related concerns and thus facilitates extensibility. We have used the architecture to construct a control system for industrial plant cultivation in greenhouses. We use this case study to motivate and evaluate our work.

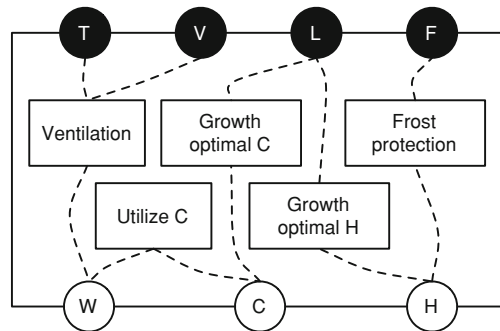
The rest of the paper is organized as follows. Section 39.2 introduces the reader to plant cultivation in greenhouses—in particular, we focus on control concerns that frequently occur in this domain, and why they are difficult to separate. In Sect. 39.3, we present a software architecture that we have used to achieve separation of control concerns. Section 39.4 presents experience with the use of our architecture to design a real system. Finally, Sect. 39.5 concludes the paper.

39.2 Greenhouse Climate Control

In this section, we will introduce greenhouse climate control—a non-trivial control system domain.

Industrial plant cultivation in greenhouses is a domain with vast amounts of variability and many interacting concerns. Greenhouses may be equipped with different sets of sensors, different sets of actuators, the plants being grown may

Fig. 39.1 Conceptual overview of a greenhouse control system



have different plant physiological properties, and various control concerns may be selected and prioritized differently by different growers.

Traditionally, plant cultivation in greenhouses has been carefully controlled by domain experts—i.e. growers—while automated control systems only have had responsibility for simple tasks such as maintaining setpoints, raising alarms, etc. However, during the past decades plant physiologists and control system vendors have pushed for increased automation. Increased automation has the potential to utilize resources more effectively—e.g. exploit fluctuating energy prices when controlling supplementary light, refrain from using heating energy that does not result in increased plant growth, and so on.

While plant physiologists have invented many control strategies that require increased automation [15–17], unfortunately, only a few attempts have been made to combine multiple strategies in a common platform [18]. There are many reasons for this. However, we consider many of the most important reasons to be related to the difficulty of separating control-related concerns. When control-related concerns are not separated,

- it is difficult to develop new control concerns without compromising existing ones,
- it is difficult to customize a control system for a particular production environment, and
- it may be difficult for the grower to understand the system’s behavior.

To illustrate the problem of separating control-related concerns, we will now introduce a small control-system example. The example is a small part of a system we are working on. An overview of the system is shown in Fig. 39.1. The full circles indicate *inputs* to the system, e.g. sensor values, user-specified configuration parameters etc. The empty circles indicate *outputs*, e.g. setpoints. Boxes indicate *control concerns*. Finally, the dashed lines indicate how control concerns are related to inputs and outputs. To improve readability, the example shows only a few selected inputs, outputs, and concerns. Statistics on the size of the full system is given in Sect. 39.4.

We will now briefly discuss the goals of each concern in Fig. 39.1:

- *Ventilation* must ensure that the temperature inside the greenhouse, T , never exceeds a specified maximum temperature, V . Ventilation is achieved by opening windows, W .
- *Growth optimal C* is concerned with dosing an optimal amount of supplementary CO_2 into the greenhouse. The concern uses the current light level, L , as input to a model that produces an optimal CO_2 setpoint, C .
- *Utilize C* must ensure that all supplementary CO_2 is put to good use. This means that opening windows (see W) and dosing of supplementary CO_2 (see C) must not happen simultaneously.
- *Growth optimal H* must ensure that the temperature in the greenhouse promotes growth. The concern uses the current light level, L , as input to a model that produces an optimal heating setpoint, H .
- *Frost protection* must ensure that the temperature in the greenhouse never gets below a specified minimum temperature, F . This means that H must always be greater than F .

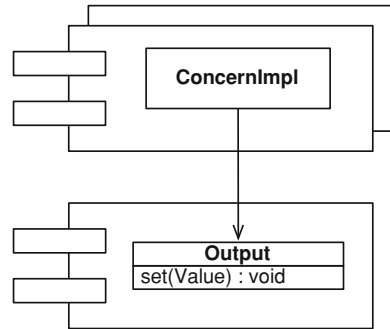
In Sect. 39.1, we referred to a control system as being non-trivial, when multiple concerns in the system affect the same controlled subsystem. In our system, outputs are the interface to controlled subsystems—e.g. the value of the heating setpoint, H , is used to control a heating system. Figure 39.1 illustrates that our system is non-trivial, since outputs—e.g. H —are affected by multiple concerns.

In our experience, it is the presence of shared outputs that makes separation of concerns in non-trivial control systems difficult to achieve. The problem is that an output is a shared resource, and that multiple concerns attempting to control this resource may produce a resource conflict [19]. The kinds of resource conflicts that emerge cannot be resolved using domain-independent techniques for separation of concerns—the conflicts are inherent to the domain. E.g. multiple concerns with a desire to set the heating setpoint, H , may produce a conflict, as only one value can be chosen. If multiple concerns are programmed to control a shared resource in ways that contradict each other, then no domain-independent resolution of the conflict can be found.

Note that the presence of shared inputs does not produce similar problems, since the interface of inputs cannot be used in ways that produce conflicts—inputs are not “controlled”, they are merely “observed” or “read”.

The only approach by which conflicts over shared outputs can be solved, while maintaining separation of concerns, is to design the resource interface in such a way that it supports a *protocol* for conflict resolution. When this approach is used on a larger scale, it is often referred to as a component framework [14]. An operating system is an example of a component framework. For example, access to files—i.e. a kind of resource—is usually coordinated by a protocol that allows multiple programs—i.e. concerns—to simultaneously read a file, while write access requires exclusive access. Coordinated access to files is possible, only because the system call API is designed with the required protocol in mind.

Fig. 39.2 Resolving conflicts over a shared output using a chain of responsibility requires global knowledge



The choice of conflict-resolution protocol depends on the problem domain. In Sect. 39.3, we will elaborate on a number of protocols, and we will suggest one that is fit for the greenhouse control system domain.

39.3 Separation of Control Concerns

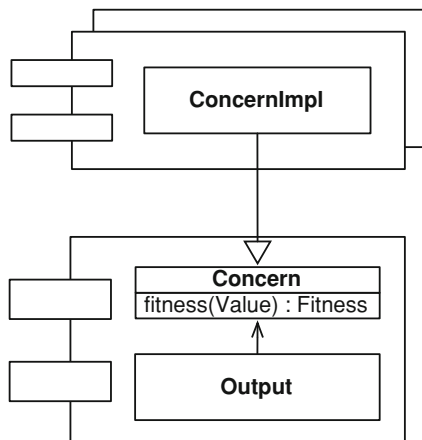
In Sect. 39.2, we have seen that conflicts over shared outputs may hinder separation of control concerns, and that this prevents us from creating extensible systems. In this section we will evaluate different protocols for their ability to separate control concerns that depend on shared outputs. The goal of this separation is extensibility—i.e. the ability to add new control concerns to the system—or remove existing ones—without modifying existing control concerns.

A naive protocol for negotiating the value of a shared output is *chain of responsibility*—see Fig. 39.2. When using this protocol, the interface of a shared output, e.g. *H*, may be as simple as a traditional set-method. The concerns are organized in an ordered chain. Each concern in the chain gets to invoke the set-method in turn. Thus, concerns late in the chain may override decisions made by concerns earlier in the chain. The problem may also emerge with more complex shared resources—e.g. in a blackboard system [20].

Protocols that rely on a chain of responsibility suffer from the following related problems:

- Creating a chain of responsibility that works require knowledge of the relative importance of all concerns that share access to outputs. When such knowledge is needed, we find it problematic to claim that concerns are separated. In any case, the protocol does not satisfy the requirements for independent extensibility [1].
- Even if we are prepared to require global knowledge of the relative importance of concerns, it cannot be guaranteed that a suitable chain can be constructed. This problem can be seen with many composition mechanisms: it is not always possible to create a suitable mixin ordering [21]. Similarly, global knowledge does not imply that a suitable aspect ordering exists.

Fig. 39.3 Resolving conflicts over a shared output using a utilitarian protocol may produce situations, where individual concerns are suppressed in ways that are unacceptable



- In our experience, the need for global knowledge is not only required when constructing the chain. In practice, this knowledge tends to become assumptions inside the implementation of individual concerns. In particular, concerns late in the chain tend to rely on implementation details of previous concerns.

In summary, protocols that rely on a chain of responsibility do not facilitate independent extension with respect to control concerns that share outputs.

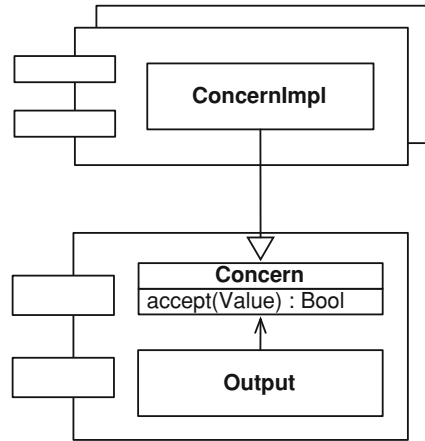
Another attempt to achieve the required separation of concerns is *utilitarian protocols*—e.g. see Fig. 39.3. In this approach, each concern provides a fitness function. Given a proposed value, the fitness function returns a *utility*—a measure of the extent to which the concern desires the proposed value. Given a set of fitness functions, the protocol finds the output value by maximizing the sum of all utilities. The rationale behind this protocol is that the total utility offers the best compromise between concerns with partially conflicting goals.

Protocols that rely on the summation of utilities suffer from the following problems:

- It is difficult to establish a common utility scale among independently developed concerns. When maximizing the sum of “different kinds” of utility, we are not really searching for “an optimal solution”, as there is no common notion of optimality. Instead, we are performing a form of satisficing [22].
- Even if a common utility scale can be established, it is possible that a concern is suppressed in ways that are clearly unacceptable. E.g. even if the frost-protection concern “thinks” that lowering the heating setpoint to minus 20 celsius would be disastrous, then this decision might be executed, if other concerns prefer this result.

In our experience, the lack of a common notion of optimality is often acceptable when dealing with control systems in the greenhouse domain. A more important problem is the cases where a control concern is suppressed by other concerns in ways that are clearly unacceptable. When the protocol may lead to this result, it is

Fig. 39.4 Resolving conflicts over a shared output using a constraint-based protocol may produce results that are unnecessarily suboptimal



impossible to develop concerns independently, because critical assumptions cannot be guaranteed when independently developed concerns are combined. If it is difficult—or impossible—to satisfy all concerns, then “no solution” is preferable to an “unacceptable solution”. If no solution can be found, it may be possible to alert a user, or it is possible to control the climate using a fallback strategy.

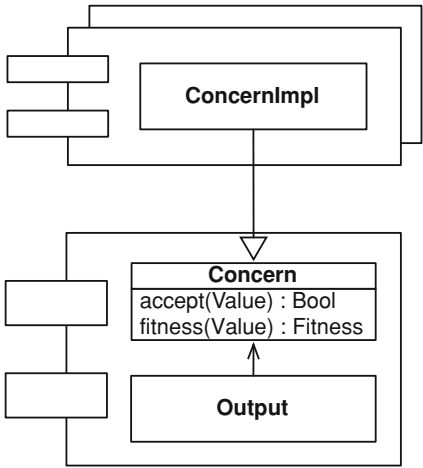
The suppression problem experienced with the utilitarian approach may be solved by *constraint-based protocols*—see Fig. 39.4. A constraint-based protocol requires each concern to implement an accept function that determines if a given value is acceptable or not. From the protocol’s point of view, each concern is a constraint, and the role of the protocol is to find a solution that satisfies all constraints.

Constraint-based protocols avoid the suppression problem because each concern can clearly specify what part of the solution space is acceptable, and what is not. However, a purely constraint-based approach suffers from other problems:

- Purely constraint-based approaches treat all solutions as equally good. This is not a problem when implementing “hard concerns” such as “make sure the heating setpoint is above five celsius”. However, when implementing “soft concerns” such as “prefer the heating setpoint to be close to the growth-optimal temperature” constraint-based approaches are insufficient.
- It is our experience that due to the limitation outlined above, developers of “soft concerns” are easily tempted to over-constrain the solution space—e.g. to write accept functions that reject suboptimal solutions even though this is not strictly necessary. With respect to extensibility, over-constraining the solution space is particularly problematic, because it is too late for any developer to relax problematic constraints at the time independently developed concerns are combined.

In summary, constraint-based approaches are insufficient—in particular when the domain contains “soft” control concerns.

Fig. 39.5 Resolving conflicts over a shared output using a hybrid protocol facilitates extensibility and separation of control concerns



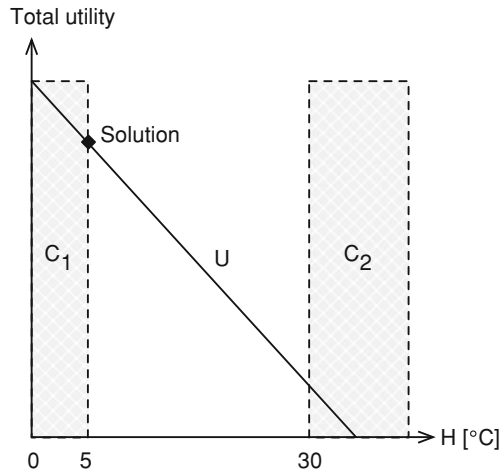
While utilitarian and constraint-based approaches all have limitations, we have developed a hybrid protocol that separates control concerns to an extent that makes independent extension attainable—see Fig. 39.5. The hybrid protocol allows the developer of a control concern to implement both an accept function and a fitness function. This hybrid approach solves the problems that emerge when utilitarian and constraint-based approaches are used separately:

- The protocol uses the set of accept functions to constrain the solution space to only contain acceptable solutions.
- The protocol uses the set of fitness functions to find the best solution within the space of what is acceptable to everyone.

The solution space in Fig. 39.6 illustrates how the best acceptable solution is found. First, accept functions constrain the solution space—e.g. C_1 ensure frost protection and C_2 avoid high temperatures that would be harmful to plants. Second, the solution is found by searching for the maximum total utility—e.g. the highest point on the U curve that is acceptable. In Fig. 39.6, the solution space contains only a single output. To illustrate a solution space for the control system depicted in Fig. 39.1, four dimensions are needed—three output dimensions and one for total utility. Note that when multiple output dimensions are present, a concern’s “opinion” about the value of one output may depend on the value of another output. Therefore, it is insufficient to search for appropriate output values one output at a time—all outputs must be evaluated together.

Until now we have discussed the choice of protocol primarily in terms of the interface that control concerns depend on. Interfaces are sufficient in order to explain how concerns are separated—in fact, the ability to separate concerns is the very purpose of interfaces. However, to see how “it really works” we will now give an outline of the protocol implementation.

Fig. 39.6 The best solution is found by maximizing total utility within the space of what is acceptable to everyone



Essentially, the purpose of our protocol is to solve a constrained optimization problem [23, 24]: the goal is to find a set of output values that all concerns can accept and that concerns collectively consider as “fit” as possible. Many different search algorithms could be applied to solve this kind of problem. However, a problem that is seen with many algorithms is that they perform badly given a complex fitness landscape—e.g. they get stuck at local optima. In our architecture, the search algorithm implementation cannot rely on assumptions about the shape of individual concerns’ fitness functions. Therefore, the composite fitness function might very well be complex. It is therefore important to choose a search algorithm that can deal with complex fitness landscapes. We have chosen to use a genetic algorithm, as it is easy to implement and known for its ability to handle complex fitness landscapes [25]. Needless to say, a genetic algorithm for our problem may be configured in various ways. Our typical configuration of the algorithm is summarized below:

- A chromosome is a set of output values—one for each output in the system.
- The algorithm’s starting population is initialized with a set of random chromosomes.
- The algorithm’s selection strategy is designed so that a chromosome is always rated the lowest possible value if just a single concern in the system cannot accept it. Acceptable chromosomes are rated according to the concerns’ normalized average perception of the chromosome.
- For each generation the worst half of the population is thrown away and replaced by new chromosomes. New chromosomes are randomly created using either mutation (making a change to the value of a random output) or crossover (combining the output values of two chromosomes). Only chromosomes that are in the better half are used for mutation and crossover.

- The algorithm terminates after a fixed set of generations. If the best chromosome is acceptable, then it is used by the application. If not, then the application must act accordingly—e.g. inform the user, use a fallback strategy etc.

Given this configuration, we have been able to satisfy control concerns encountered while developing a control system for plant cultivation in greenhouses. Experience from this effort is discussed in [Sect. 39.4](#).

39.4 Experience

Using our architecture, we have created an extensible control system for industrial plant cultivation in greenhouses. In this section, we will present a few facts about the system and we will share experiences gained from architecting and maintaining the system.

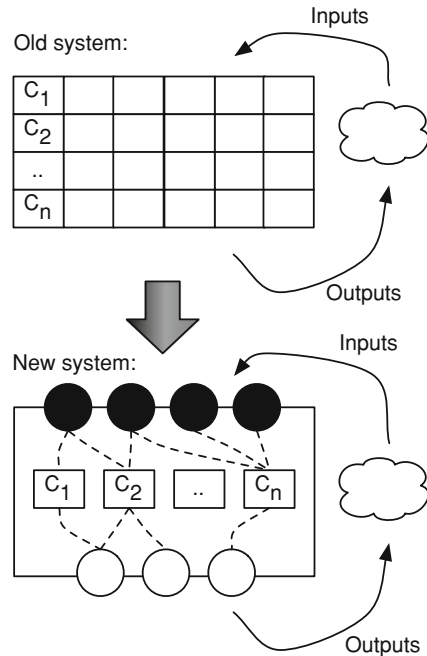
The current system is the result of an iterative process spanning three years. At the beginning of this process, the system's architecture was very different. The old system was created by porting a number of essential components from the IntelliGrow system [18]. In IntelliGrow, control concerns were organized in a chain of responsibility—this legacy of IntelliGrow's architecture greatly dominated the old system.

The new system started to emerge about one year ago. The new architecture relies on the hybrid protocol, as it is described in [Sect. 39.3](#). Another important trait of the new system is that it is built on top of the NetBeans Rich Client Platform [26]—a service-oriented component platform. Thus, all functionality—including the core of the hybrid protocol—is distributed as components—modules in NetBeans lingo. At the time of this writing, the system is composed from 23 components. The total size of the system is 15,163 lines of Java code (excluding blank lines and comments). We use the system in various configurations. However, in total we support 29 input resources, 5 output resources, and 20 control concerns.

Obviously, a non-trivial control system has many features that are not directly related to control—this is the case for the old as well as the new system. Both systems have components for user-customizable graphs, integration with databases, integration with sensors, integration with actuators, and much more. However, the heart of both systems is the ability to control the production environment. For the reasons previously discussed, this is also the part of the system that is most difficult to make extensible.

Figure 39.7 illustrates how the separation of control logic was changed fundamentally in the new system. In the old system (top part of the figure), control concerns were organized in a chain of responsibility: first, input values were read from the environment. Second, all outputs were manipulated by one concern at a time. Finally, outputs were written back to the environment. In the new system (bottom part of the figure), control concerns are organized using our hybrid protocol: first, inputs are read from the environment. Second, our hybrid protocol is

Fig. 39.7 While the old system organized control concerns in a chain of responsibility, the new system uses our hybrid protocol to resolve conflicts over shared outputs



used to satisfy all concerns in the system. Finally, if an acceptable solution is found, the output values are written to the environment.

After having performed an informal, yet thorough, comparison of the two systems, we consider the following differences to be particularly important:

- It has become easier to add control concerns without performing a global integrity check. In the old system, the introduction of a new control concern often required detailed knowledge or even modification of existing concerns. In the new system, the usual scenario is that new control concerns can be introduced without modifying existing ones.
- It has become possible to model control concerns with finer granularity. In the old system, a typical configuration is built from 5 complex concerns—a roughly functionally equivalent configuration of the new system is built from 17 simpler concerns, each of which is significantly smaller than concerns in the old system. The finer granularity indicates that individual concerns are more self-contained—this corresponds well with our general experience.
- The mindset required to write a concern is very different in the two systems: in the old system, writing a concern is all about focusing on writing an algorithm for how to come up with “the result”. In the new system, writing a concern is about evaluating proposed solutions. The new style of writing concerns may seem backwards at first. However, in our experience, with a little practice, it becomes possible to write much smaller control concerns that are much more mutually oblivious to each other.

- The architecture of the new system has improved our ability to explain the system's composite behavior to the user. In the old system, it was difficult to explain interaction among control concerns to the user. In the new system, we are able to dynamically analyze which control concerns are in conflict. This is an aspect of the system that we would like to improve even more in the future.

In summary, it is our experience that the proposed architecture improves extensibility with respect to control concerns in non-trivial control systems.

39.5 Conclusion

Non-trivial control systems constitute a domain where extensibility is particularly difficult to achieve. We have suggested that poor separation of control-related concerns is an important source of these difficulties. Unfortunately, we do not see how domain-independent techniques for separation of concerns can improve the situation. Thus, the best way to improve extensibility is the design of a domain-specific software architecture that improves separation of control concerns.

We have presented an architecture that promotes separation of control-related concerns in non-trivial control systems. The core of the architecture is a protocol capable of resolving conflicts among control concerns that share output resources. Our protocol is a hybrid that uses a genetic algorithm to combine a utilitarian approach—i.e. each concern implements a fitness function—with a constraint-based approach—i.e. each concern implements an accept function. The combination of both approaches is important in order to separate control-related concerns without encountering problems that emerge from using one approach without the other: first, with purely utilitarian approaches, a control concern may be suppressed by others in ways that are unacceptable. Second, with purely constraint-based approaches, there is a tendency to choose suboptimal solutions, as all acceptable solutions are treated as equally good. Our hybrid approach improves the situation with respect to both problems.

We have used our architecture in a non-trivial control system for industrial plant cultivation in greenhouses. It is our experience that the architecture has facilitated separation of control-related concerns and thus improved extensibility of our system.

We think there is a need for research that pursues separation of concerns without insisting on domain-independent solutions. In this paper, we have investigated the domain of non-trivial control systems. In the future, we would like to see our architecture applied to a larger set of systems. We would also like to investigate other domains where separation of concerns is difficult to achieve.

References

1. Szyperski C (1996) Independently extensible systems—software engineering potential and challenges. *Aust Comput Sci Commun* 18:203–212
2. Hürsch W, Lopes C (1995) Separation of concerns, Northeastern University
3. Ossher H, Tarr P (2002) Multi-dimensional separation of concerns and the hyperspace approach. In: *Proceedings of the software architectures and component technology*, pp 293–323
4. Kiczales G, Lamping J, Mendhekar A, Maeda C, Lopes C, Loingtier J, Irwin J (1997) Aspect-oriented programming. In: *ECOOP 1997—object-oriented programming*, pp 220–242
5. Aldrich J (2005) Open modules: modular reasoning about advice. In: *ECOOP 2005—object-oriented programming*, pp 144–168
6. Ostermann K, Kniesel G (2000) Independent extensibility—an open challenge for aspectj and hyperj. In: *ECOOP 2000—international workshop on aspects and dimensional computing*
7. Aksit M, Rensink A, Staijen T (2009) A graph-transformation-based simulation approach for analysing aspect interference on shared join points. In: *Proceedings of the 8th ACM international conference on aspect-oriented software development*. ACM, pp 39–50
8. Kiczales G, Hilsdale E, Hugunin J, Kersten M, Palm J, Griswold W (2001) An overview of aspectj. In: *ECOOP 2001—object-oriented programming*, pp 327–354
9. Restivo A, Aguiar A (2008) Disciplined composition of aspects using tests. In: *Proceedings of the 2008 AOSD workshop on linking aspect technology and evolution*. ACM, pp 1–5
10. Douence R, Fradet P, Südholt M (2002) A framework for the detection and resolution of aspect interactions. In: *Generative Programming and Component Engineering*. Springer, pp 173–188
11. Zambrano A, Vera T, Gordillo S (2006) Solving aspectual semantic conflicts in resource-aware systems. In: *ECOOP 2006—workshop on reflection, AOP, and meta-data for software evolution*. Citeseer, p. 79
12. Herrmann S (2003) “Object teams: Improving modularity for crosscutting collaborations.” *Objects, Components, Architectures, Services, and Applications for a Networked World*. Springer, Berlin, pp 248–264
13. Steimann F (2006) The paradoxical success of aspect-oriented programming. *ACM SIGPLAN Notices* 41(10):481–497
14. Weck W (1997) Independently extensible component frameworks. In: Mühlhäuser M (ed) *Special issues in object-oriented programming*, dpunkt Verlag, pp 177–183
15. Tantau H, Lange D (2003) Greenhouse climate control: an approach for integrated pest management. *Comput Electron Agric* 40(1–3):141–152
16. Van Pee M, Berckmans D (1999) Quality of modelling plant responses for environment control purposes. *Comput Electron Agric* 22(2–3):209–219
17. Van Straten G, Challa H, Buwalda F (2000) Towards user accepted optimal control of greenhouse climate. *Comput Electron Agric* 26(3):221–238
18. Aaslyng J, Lund J, Ehler N, Rosenqvist E (2003) Intelligrow: a greenhouse component-based climate control system. *Environ Model Softw* 18(7):657–666
19. Bisbal J, Cheng B (2004) Resource-based approach to feature interaction in adaptive software. In: *Proceedings of the 1st ACM SIGSOFT workshop on self-managed systems*. ACM, pp 23–27
20. Corkill D (1991) Blackboard systems. *AI Expert* 6(9):40–47
21. Ducasse S, Nierstrasz O, Schärli N, Wuyts R, Black A (2006) Traits: a mechanism for fine-grained reuse. *ACM Trans Program Lang Syst (TOPLAS)* 28(2):331–388
22. Simon H (1956) Rational choice and the structure of the environment. *Psychol Rev* 63(2):129
23. Joines J, Houck C (1994) On the use of non-stationary penalty functions to solve nonlinear constrained optimization problems with ga’s. In: *Evolutionary computation, 1994*. IEEE

- world congress on computational intelligence. Proceedings of the First IEEE Conference on evolutionary computation. IEEE, pp 579–584
24. Homaifar A, Qi C, Lai S (1994) Constrained optimization via genetic algorithms. *Simulation* 62(4):242
 25. Mitchell M (1998) An introduction to genetic algorithms. The MIT press, Cambridge
 26. Boudreau T, Tulach J, Wielenga G, (2007) Rich client programming: plugging into the NetBeans platform, vol 1. Prentice-Hall PTR, Englewood Cliffs, pp 79–84

Chapter 40

Illicit Image Detection: An MRF Model Based Stochastic Approach

Mofakharul Islam, Paul Watters, John Yearwood, Mazher Hussain
and Lubaba A. Swarna

Abstract The steady growth of the Internet, sophisticated digital image processing technology, the cheap availability of storage devices and surfer's ever-increasing interest on images have been contributing to make the Internet an unprecedented large image library. As a result, The Internet quickly became the principal medium for the distribution of pornographic content favouring pornography to become a drug of the millennium. With the arrival of GPRS mobile telephone technology, and with the large scale arrival of the 3G networks, along with the cheap availability of latest mobile sets and a variety of forms of wireless connections, the internet has already gone to mobile, driving us toward a new degree of complexity. In this paper, we propose a stochastic model based novel approach to investigate and implement a pornography detection technique towards a framework for automated detection of pornography based on contextual constraints that are representatives of actual pornographic activity. Compared to the results published in recent works, our proposed approach yields the highest accuracy in detection.

M. Islam (✉) · P. Watters · J. Yearwood
Internet Commerce and Security Laboratory, School of Science, Information Technology
and Engineering, University of Ballarat, Ballarat, VIC, Australia
e-mail: mlrj@mmmi.sdu.dk

M. Hussain
Melbourne Institute of Technology, Melbourne, VIC, Australia

L. A. Swarna
IBM Australia, Ballarat, VIC, Australia

Table 40.1 Pornography statistics: top ten reviews

| |
|---|
| Every second— \$3,075.64 is being spent on pornography |
| Every second—28,258 internet users are viewing pornography |
| Every second—372 internet users are typing adult search terms into search engines |

40.1 Introduction

Accessing the Internet with the computers and a range of other communicating devices are increasingly a modern day must have for children and young people around the globe. Not only are they establishing new cultural norms, they are also becoming mainstream within education. A great majority of minors globally spend bulk amount of their time online to locate their educational stuff and as a source of perfectly innocent fun and games. But every child and young people who uses the Internet will almost certainly be exposed at some stage to material that will shock and possibly harm them, or they will come into contact with organization or individuals who mean them injury. To follow potentially catastrophic consequences throughout a children's lifetime, he/she needs only one such encounter to go wrong. It is therefore, no surprise that the Internet as well as the social networking sites pose a serious threat to our children who are more or less vulnerable, either some or all of the time while they are on the Internet. The risks posed by criminals who attempt to share, exchange, consume and produce child exploitation material, however, are fairly clear, and law enforcement is faced with the difficult task of trying to deal with the sheer volume of material (and offences) in a streamlined and systematic way.

Research [1] shows some of the statistics on pornography as shown in the pie-chart below (Tables 40.1, 40.2);

Apart from the exposure of age inappropriate content to minors, dealing with pornography in the workplace is a serious challenge for many large organizations but employing a block-all-images Internet browsing and email policy no longer provides a viable solution. In a paper written by McGuire et al. found that the viewing of pornography can serve as a source of a paraphilic "vivid sexual fantasy" which, when contemplated during masturbation, may condition men into perversion [2]. Further, research also established a direct link between pornography consumption and crimes like sexual abuse, rape, violence, and child molestation [3–5].

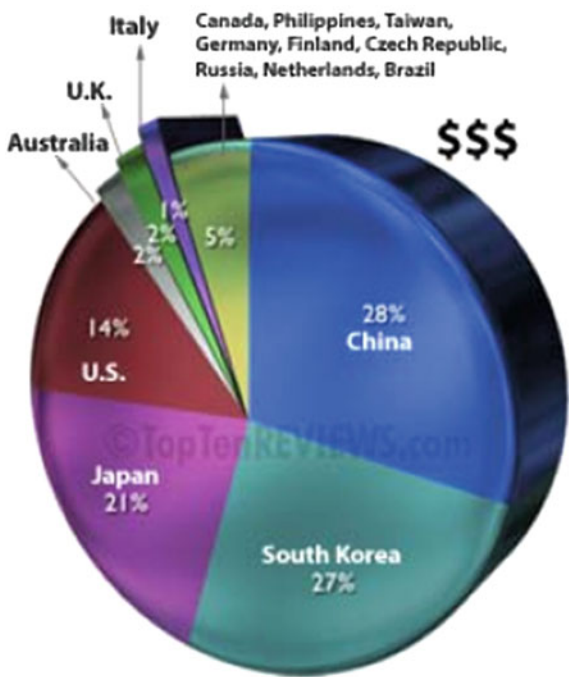
Research showed that most of the audience belongs to the lower levels that gradually progress towards higher levels. If we are able to put a stopping mechanism some where in this low level area, it will protect people from entering the most exiting world of arousal and being pornography addicted.

Contemporary research established a link between child pornography and adult pornography. From that point of view, it will not only prevent people from viewing the most erotic and bestial images but also put an end for the potential paedophiles who are currently belongs to less severe groups to progress further towards most

Table 40.2 Comparison of performance between the proposed and Belem and Cavalcanti [21] approach for lower body parts

| Method | TP | FP | FN | TN | PRE | REC | Acc |
|---------------------------|-----|----|----|-----|------|------|------|
| <i>Upper body parts</i> | | | | | | | |
| Belem and Cavalcanti [21] | 161 | 61 | 39 | 139 | 0.73 | 0.81 | 0.75 |
| Proposed | 189 | 32 | 19 | 168 | 0.86 | 0.95 | 0.89 |
| <i>Lower body parts</i> | | | | | | | |
| Belem and Cavalcanti [21] | 161 | 61 | 39 | 139 | 0.73 | 0.81 | 0.75 |
| Proposed | 189 | 32 | 19 | 168 | 0.86 | 0.95 | 0.89 |

Fig. 40.1 Worldwide pornographic revenue (courtesy: top ten reviews)



severe groups that eventually make them professional paedophiles at some stage. From the studied literatures, we find a progression model for pornography consumer as shown in Fig. 40.1.

So, there is a fundamental demand in real time classification of images to block inappropriate contents at home and in a business or commercial environment. The problem therefore is the detection and prevention of certain types of images. Real-time detection of pornography can effectively prevent pornographic images from entering at home and into the workplace via email and Internet browsing.

The software industry suggests two types of solutions to combat with the problem of pornography—collecting the adult web site addresses and collecting key words of adult web pages. While the former allows the user to access the requested page upon verifying its IP address, the latter performs the content

Fig. 40.2 Progression model for pornography consumer

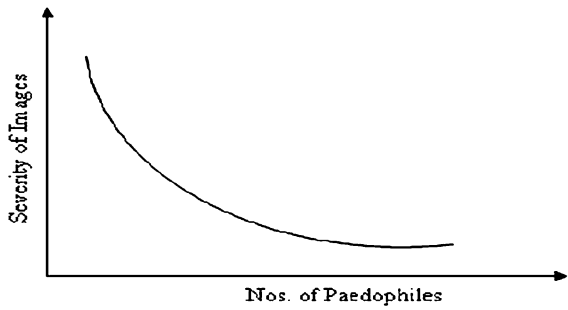
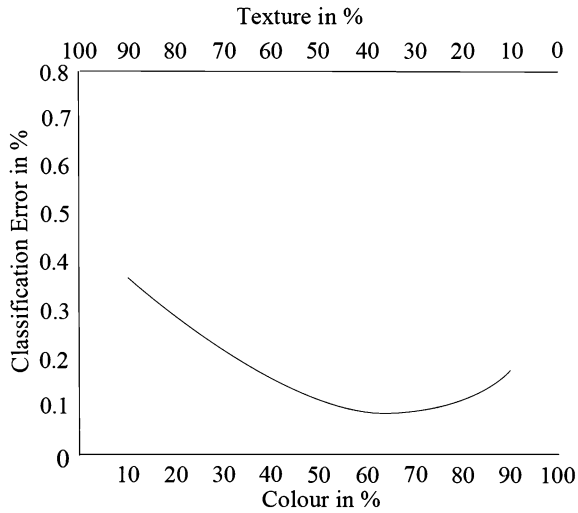


Fig. 40.3 Mixing weight of colour and texture



analysis for providing access to the user. Here, the filtering programs block the pages from viewing upon finding any related key word of the web page in user’s computer. The drawback of these systems is that it is an uphill task to collect all adult web site addresses because hundreds of new web sites are launching every day. The drawback of the latter system is that it tends to block the sex education pages due to the key words used on these pages. In addition, pornographic web-sites introducing new keywords to attract peoples and minors, that are being most often used by the users during browsing (Fig. 40.2).

We are interested in the computational foundations of vision that helps us designing machine vision systems with applications to pornographic image detection. Our proposed project aims to find out the solutions for some fundamental questions in computer vision. How can we recognize persons, guns, cars, boats and many other suspicious categories of objects in cluttered pictures? How can we be trained these categories in the first place? Can we provide machines with ability akin to this?

The remainder of this paper proceeds as follows: In [Sect. 40.2](#) we present previous work in this specific field. In [Sect. 40.3](#), we describe our proposed pornography detection model, which estimates body part appearance models for pictorial structures using latent relationships between the appearances of different body parts. Experimental results demonstrating the accuracy and efficiency of the proposed approach are discussed in [Sect. 40.4](#) and finally in [Sect. 40.5](#), we present our conclusion and future work ([Fig. 40.3](#)).

40.2 Related Work

While significant research have already been conducted on pornography or illicit image detection most often with miscellaneous low and high level features, and other visual cues in fact, no research has looked closely at the subject.

In fact, all the pornography detection approaches are based on skin detection module that defines a decision rule based on which skin and non-skin pixels will be discriminated. A decision rule based on the distance of the pixel color to skin tone using a metric plays the crucial role in the discrimination process as this metric is defined by the skin color modelling technique. The final goal of skin-color detection is to choose a classifier that will discriminate between a skin and a non-skin. From classification point of view, skin-color detection can be viewed as a two class problem: skin pixel vs. non-skin pixel.

Literature suggests different classifiers are employed by different authors for skin detection based on their respective contexts, where data distributions or data modelling played a crucial role. Our literature review suggests three specific types of skin modelling techniques—Explicitly defined skin region, Non-parametric distribution modelling, Parametric distribution modelling, and Adaptive modelling. In Explicitly Defined Skin Region techniques, threshold is a parameter that stipulates the values a pixel can be if it is to be considered as skin. In Nonparametric distribution modelling, Normalized Lookup Table (LUT), Naïve Bayes Classifier, and Self Organizing Map (SOM) are most often used as skin classifier. Single Gaussian, Mixture of Gaussians, Multiple Gaussian Clusters, and Elliptical Boundary Model are mostly used as Parametric Model for skin color detection. Adaptive skin model employs two or more skin-color models from the models as discussed above in combination or in sequence to make the technique robust against the varying conditions. In other word, it is called ‘Hybrid Model’.

Most of the adult image or pornographic detectors employed these skin color models based on different color space along with some other high level features like shape, color moment, color mean, region descriptors for their classification schemes.

A mathematical approach of grouping the images using the detected skin regions and extracted high level region features from these regions that might have link to the pornography is called pornographic classification, which is mainly focussed on classification of the skin region into benign or pornographic. We can

broadly categorize pornographic image classifier into three groups: Supervised Machine Learning, Geometric Classifier, and Boosting Classifier. Supervised Machine Learning can be further categorized hereunder into four groups. Supervised Machine Learning can be further categorized hereunder into four groups—Support Vector Machine [6–8], Neural Network, Decision Tree [9], and K-Nearest Neighbour [10]. Fleck et al. demonstrated an affine-invariant model that define human as a set of rules describing how to assemble possible girdles and spine-thigh groups, where both the individual geometry of the body parts and the relationships between parts are constrained by the geometry of the skeleton which eventually provides an appropriate and effective model for human body recognition [11]. Bootstrapping induces a classifier on a small set of labelled data and a large set of unlabeled data. Lee et al. reported an increase of sensitivity from 81.74–86.29 % with boosting algorithm [12].

Cusano et al. reported better performance of SVM than multiple decision trees [6]. Bosson et al. found that neural networks (83.9 specificity) gave slightly better results to that of k-NN and SVM [13]. Bosson et al. attained 94.7 % sensitivity and 95.1 % specificity using NN with MPEG-7 Descriptors. Zheng et al. [11] found that DT is able to provide better accuracy than NN and SVM [13]. Xu et al. demonstrated superior performance of k-NN over NN [10].

Almost all the existing techniques proposed region descriptors that are in fact, not directly aimed at capturing contextual constraints representatives of pornography rather based on some heuristic high level statistical and geometrical features on skin region like area, size, shape, location, and their relationships in terms of some ratios, resulting poor performance in finding the actual pornographic stuff on images and thus failing to utilize its full potential in areas like detection of pornography. Further, the extracted features as employed in the existing work are found to be vulnerable under the inherent imaging artifact like intensity inhomogeneity, scale, rotation, transformation, occlusion, and camera viewpoint. These are the void in the literature that this research is attempting to fill.

40.3 Our Proposed Approach

Although the identification and detection of pornographic activity in an image is easy to a human observer, automatic and accurate identification is difficult and complex. Obtaining satisfactory detection results depends mainly on the ability of the descriptors in characterizing pornographic contextual constraints.

Segmentation of skin colour often used as feature in different face and human tracking applications plays a crucial role in locating and identifying a naked human object in an image. Initially, skin regions of an image are segmented based on the colour to reduce the search space, resulting a faster approach appropriate for real time applications. Literature suggests a significant contribution in skin detection area over the last decade. Different authors suggested different approaches for skin detection based on different colour model out of which YC_b playing a dominant

role due to the fact that the luminance and chrominance components in this colour space are stored separately [14–20]. As a result, YC_rC_b is greatly suited to skin detection and some authors [14] reported found that the YC_bC_r gives the best skin detection results compared to seven other colour space transformations. In fact, we are not going to contribute anything in skin detection rather we use an existing skin detection technique [17] that has been found to be robust.

The resulting segmented skin region of a body is then divided into two parts—upper body part and lower body part that are independently searched with our proposed approach to detect pornography specific contextual constraint. While the upper body part is searched for detecting breast, the lower part is for pubic areas, especially the genitalia.

We employ Markov Random Fields (MRF) model to encode the contextual constraints representative of pornography and use these constraints as prior knowledge on pornographic activity in an image during classification. Markov Random Fields (MRF) model theory is found to be extremely robust in capturing contextual contexts in an image, especially under noisy condition.

40.3.1 Feature Extraction and Selection

Feature extraction and selection are the keys to robust and accurate image classification and in fact, play the most crucial role in pornography detection in this particular work. Here, we proposed a new hybrid feature derived from the generic low level features like colour and texture, which has been found to be robust and accurate in characterizing the appearance and internal structure of an image region.

- **Color Feature:** 50 skin patches are collected by just cropped off the segmented body region in each training image. A new low level color feature is derived utilizing the pure chromatic information ('u' and 'v') of the original CIE-Luv color model as;

$$Chroma_C_{(u,v)} = C_{(u,v)} - C_{(u,v)}_mean \quad (40.1)$$

where, $N = \text{Nos. of pixels in a single skin patch, and}$

$$C_{(u,v)}_mean = \frac{\sum_{n=1}^N C_{(u,v)}}{N}$$

Dropping the luminance component 'L' and considering 'u' and 'v' values, offering us a significant advantage for utilising pure color information in the detection process. Using pure color information, a skin model or detector gets an ability to be able to adapt to the changes in the lighting and the viewing environment. The key advantage of being a lighting invariant color spaces

made few spaces like *CIE-Luv*, *YC_bC_r*, and *HSI* the most popular choices to the research community working on skin detection.

In addition to the color, texture is another feature that can be employed as a cue for image analysis. Texture features are constructed in such a way that they characterize local variation in intensity or color within the neighbourhood surrounding the pixel. As a result, a value is assigned to the pixels with a given texture. Research suggests, color and texture based analysis are in combination capable of producing more accurate results than either the color or texture feature used independently [21].

Literature suggests the wavelet transform outperforms its other counterparts while applied in cluttered image having complex background. The wavelet transform has the capability to capture subtle texture information on objects appear on cluttered background more precisely than other texture descriptors [21]. Based on wavelet transform we derive a more powerful texture descriptor that is found to be robust in discriminating even subtle differences in texture.

- **Texture Feature:** In a similar manner like *chroma* feature above, we define another low level texture feature employing the diagonal, horizontal, and vertical coefficients as yielded by the DWT as;

$$dwt_{(d,h,v)} = coef_{(d,h,v)} - coef_{(d,h,v)}_mean \quad (40.2)$$

where, $N = \text{Nos. of pixels in a single skin patch, and}$

$$coef_{(d,h,v)}_mean = \frac{\sum_{n=1}^N coef_{(d,h,v)}}{N}$$

where $coef_{(d,h,v)}$ denotes diagonal, horizontal, and vertical coefficients.

However, mixing weight of colour and texture, which is pertinent to accuracy in capturing image appearance throws an open problem. Up to date research, computer vision and image processing in particular do not provide any authenticated relationship between these two while applied in combination, which incite us to run an experiment to find out an optimal weight ratios of these two. In our experiment, we have sourced 100 pornographic images freely available on the Internet and run an MLE (Maximum Likelihood Estimation) based segmentation scheme on these images to segment skin regions on them. Misclassification rate are plotted against both colour and texture component to find a cogent mixing weight.

Percentage of colour and texture components are taken at 5 % interval starting from 10 to 90 % and the graph is plotted on the mean classification error based on the segmentation results as obtained for the 100 images as chosen earlier. The plotted curve shows its minimum at 65 % colour and 35 % texture. So, we use this mixing weight of colour and texture in our work.

In addition to the colour and texture low level features, we add another powerful local image region descriptor called the ‘Local binary patterns’ (LBP). The LBP has been found to be a powerful feature allowing a classifier or detector to exploit fundamental properties of local image texture efficiently and effectively in a classification or matching. Applying LBP on an image yields occurrence histogram, which is a powerful texture descriptor that contains information about the distribution of the local micro-patterns like edges, spots, flat areas, over the image region. Label for every pixel of the image is assigned by the original LBP operator applying a thresholding technique on the centre pixel with his 3×3 neighbourhood.

- **Local Binary Feature:** Local Binary Patterns were originally introduced as a texture descriptor by Ojala [22], and have subsequently been employed as face and expression identification features [23, 24].

$$LBP_{P,R}(x_o) = \sum_{p=0}^{P-1} u(x_p - x_o) 2^p \quad u(y) = \begin{cases} 1. & y \geq 0 \\ 0. & y < 0 \end{cases} \quad (40.3)$$

where $p = \text{nos. of neighboring pixels}$, $R = \text{radius of the neighborhood}$.

In this particular work, local binary patterns are extracted from the individual chromatic channels to preserve channel-wise discriminative ability in terms of LBP.

Now, we combine these extracted low level features to get a high dimensional feature vector representative of each individual image and image object, which we call Luminance Invariant Region Descriptor (LIRD);

$$LIRD = \left[\sum_{n=1}^N \{ \{ chroma_{(u,v)} \} * 0.65 + \{ dwt_{(d,h,v)} \} * 0.35 + lb_{p(r,b)} \} \right] \quad (40.4)$$

40.3.2 Classification

For pixel classification, we employ the MAP-MRF labelling, where $P(f|d)$ is the posterior distribution of an MRF. In Bayes labelling, it’s an important step to derive this distribution. The problem here is to labelling the pixel into two class—malign (pornographic) or benign (normal). So, it’s a binary classification problem. Assuming the joint prior distribution of a pornographic image is

$$P(f) = \frac{1}{Z} e^{-U(f)} \quad (40.5)$$

where $U(f) = \sum_{ij} \sum_{i' \in \{i-1, i+1, j-1, j+1\}} (f_{ij} - f_{i'j'})^2$ is the *prior energy* for an image contains pornographic contents. Assuming that the observation is the true pornographic object on images plus independent Gaussian noise, $d_{ij} = f_{ij} - e_{ij}$, where $e_{ij} = N(\mu, \sigma^2)$, then the likelihood distribution will take the form

$$p(d|f) = \frac{1}{\prod_{i=1}^m \prod_{j=1}^n \sqrt{2\pi\sigma^2}} e^{-U(d|f)} \quad (40.6)$$

where $U(d|f) = \sum_{i=1}^m \sum_{j=1}^n (f_{ij} - d_{ij})^2 / 2\sigma^2$ is the *likelihood energy*. Now, we can compute the posterior probability as

$$P(f|d) \propto e^{-U(f|d)} \quad (40.7)$$

where

$$U(f|d) = U(d|f) + U(f) \quad (40.8)$$

$$= \sum_{i=1}^m \sum_{j=1}^n (f_{ij} - d_{ij})^2 / 2\sigma_{ij}^2 + \sum_{i=1}^m \sum_{j=1}^n (f_{ij} - f_{(i-1, j-1)})^2$$

is the *posterior energy*. The MAP estimate is equivalently found by minimizing the posterior energy function

$$f^* = \arg_f \min U(f|d) \quad (9)$$

The only parameter that we need to estimate is σ_{ij} , which can be done by employing the EM (Expectation and Maximization) algorithm. Now, $U(f|d)$ is fully specified and the MAP-MRF solution is completely defined.

40.4 Experimental Results and Discussion

We have sourced the pornographic images freely available on the Internet for training and testing purpose. Selected images are full body frontal and of upright position with a 10–15 % tilt in horizontal direction. All the images are divided into upper body part and lower body part. Upper body parts and lower body parts are now projected separately to a common coordinate frame, where they are roughly aligned in location and scale. We apply our novel descriptor LIRD on both the parts taken from the 200 images for feature extraction that eventually are employed for training our classifier to classify pornographic object and pornographic image in terms of *likelihood energy* and *prior energy* respectively. The *likelihood energy* and *prior energy* for non-pornographic object and benign (normal) image are considered as rest of the world i.e., $-U(d|f)$ and $-U(f)$ respectively.

For testing purpose, we have downloaded another 200 images—100 full body frontal pornographic images in upright position with a 10–15 % tilt in horizontal direction, and 100 normal human body images that do not contain any pornographic activity and have similar orientation. All the images are divided into two parts akin to the training images as done earlier. Skin segmentation is done on pornographic test images to detect the naked skin followed by division of the same into two parts—upper body parts and lower body parts cutting halfway through the belly button. We employ a person detector [25] to locate the person in the normal images to reduce the search space followed by the same procedure to get the upper body and lower body parts.

Pixel labelling is done on upper body parts and lower body parts separately for detection of breast and genitalia respectively using the MAP-MRF labelling as derived earlier in the previous section.

An experimental study reveals that pornography detection rate increases with increase of the number of pixel labelled to pornographic in an image up to a certain point. We have obtained such an optimal point i.e., 60 %, which yields maximum detection accuracy. Above 60 %, no significant improvement has been noticed in our experiments. Details of the experiments are not provided here due to space constraint.

A popular measure, called Precision and Recall (Eq. 40.10), based on True Positive (TP), True Negative, False Positive (FP), and False Negative (FN) is applied on both—our experimental results and the results obtained from another pornography image detector proposed by Belem and Cavalcanti [21] to evaluate the performance of our novel approach. We selected Belem and Cavalcanti [21] for performance comparison as its authors reported more than 90 % accuracy. Apart from the Precision and Recall, we also compared accuracy (Eq. 40.11) of our approach with Belem and Cavalcanti [8].

$$\text{Precision} = \frac{TP}{TP + FP} \quad (40.10)$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{P + N} \quad (40.11)$$

We experienced a bit lower performances with Belem and Cavalcanti [14] than what claimed by the authors. The low performance might be attributed to the use of unconstraint images in this particular work.

40.5 Future Work and Conclusion

We have proposed a novel pornography detection approach, where pornography specific contextual contexts based on few low level image features are encoded to detect pornography with maximum accuracy. The proposed approach is primarily

aimed at detection of pornography specific objects in an image using MRF model, which has been found extremely effective while dealing with unconstrained and cluttered images. The proposed approach is applicable to the frontal images with upright position but can be extended by modelling other pornographic orientations (non-frontal and non-upright) as well. To the best of our knowledge, this is the first of its kind which is able to recognize pornography using actual pornography specific contextual constraint and thus paves the way for research in this area to not only help pornography detection, but also to contribute significantly in security and surveillance, human body movement and pose identification, medical imaging and much more.

REFERENCES

1. Ropelato J (2009) Internet pornography statistics. Retrieved March 2009, from <http://internet-filter-review.toptenreviews.com/internet-pornography-statistics.html>
2. McGuire RJ, Carlisle JM, Young BG (1964) Sexual deviations as conditioned behaviour: a hypothesis. *Behav Res Ther* 2:185–190
3. Silbert MH, Pines AM (1984) Pornography and sexual abuse of women. *Sex Roles* 10:857–868
4. Carter DL, Prentky RA, Knight RA, Vanderveer PL, Boucher RJ (1987) Use of pornography in the criminal and developmental histories of sexual offenders. *J Interpers Violence* 2:196–211
5. Alexy EM, Ann WB, Robert AP (2009) Pornography use as a risk marker for an aggressive pattern of behavior among sexually reactive children and adolescents. *J Am Psychiatr Nurses Assoc* 14(6):442–453
6. Ruiz-del-Solar J, Castañeda V, Verschae R, Baeza-Yates R, Ortiz F (2005) Characterizing objectionable image content (pornography and nude images) of specific web segments: Chile as a case study. In: *Proceedings of the third Latin American web congress (LA-WEB'05, IEEE)*
7. Jeong C, Kim J, Hong K (2004) Appearance-based nude image detection. In: *Proceedings of the 17th international conference on pattern recognition (ICPR'04)*, IEEE computer society, Washington, DC, USA
8. Zheng QF, Zeng W, Wang WQA, Gao W (2006) Shape-based adult image detection. *Int J Image Graph (IJIG)* 6(1):115–124
9. Chai D, Bouzerdoun A (2000) A Bayesian approach to skin color classification in YCbCr color space. In: *IEEE TENCON'00*, vol 2, pp 421–424
10. Xu Y, Li B, Xue X, Lu H (2005) Region-based pornographic image detection. In *IEEE 7th workshop on multimedia signal processing*
11. Fleck M, Forsyth A, Bregler C (1996) Finding naked people. In: *Computer vision-ECCV'96*, 4th european conference on computer vision, Cambridge, UK, vol II, pp 593–602
12. Lee J-S, Kuo Y-M, Chung P-C, Chen E-L (2007) Naked image detection based on adaptive and extensible skin colour model. *Pattern Recognit* 40(8):2261–2270
13. Wang Y, Wang W, Gao W (2005) Research on the discrimination of pornographic and bikini images. In: *Proceedings of the 7th IEEE international symposium on multimedia (ISM'05)*
14. Storrington M, Anderson HJ, Granum E (1999) Skin colour detection under changing lighting conditions. In: *7th symposium on intelligent robotics system*, Coimbra, Portugal
15. Chai D, Ngan KN (1998) Locating facial region of a head-and- shoulders color image. In: *ICFGR'98*

16. Wong KW, Lam KM, Siu WC (2003) A robust scheme for live detection of human faces in color images. *Signal Process Image Commun* 18(2):103–114
17. Zheng Q, Zhang M, Wang W (2004) A hybrid approach to detect adult web images. In: *Advances in multimedia information processing-PCM 2004*, Proceedings of the 5th pacific rim conference on multimedia, Tokyo, Japan, Part II PCM (2), Nov 30–Dec 3, pp 609–616
18. Khan R, Stöttinger J, Kampel M (2008) An adaptive multiple model approach for fast content-based skin detection in on-line videos. In: *Proceeding of the 1st ACM workshop on analysis and retrieval of events/actions and workflows in video streams - AREA'08*, pp 89–96
19. Ye Q, Gao W, Zeng W, Zhang T, Wang W, Liu Y (2003) Objectionable image recognition system in compression domain. In: *Intelligent data engineering and automated learning*, 4th international conference-IDEAL 2003, Hong Kong, China, pp 1131–1135, Springer
20. Cao L, Li X, Yu N, Liu Z (2004) Naked people etrieval based on adaboost learning. In: *International conference on machine learning and cybernetics*, IEEE '2004, Beijing, vol 2, pp 1133–1138, ISBN: 0-7803-7508-4
21. Belem RJS, Cavalcanti JMB (2005) SNIF: a simple nude image finder. In: *IEEE Proceedings of the third Latin American web congress (LA-WEB'05)*
22. Ojala T, Pietikainen M, Harwood D (1996) A comparative study of texture measures with classification based on feature distributions. *Pattern Recogn* 29:51–59
23. Ahonen T, Hadid A, Pietikainen M (2004) Face recognition with local binary patterns. In: *8th European conference on computer vision*, pp 469–481
24. Shan C, Gong S, McOwan P (2005) Conditional mutual information based boosting for facial expression recognition. In: *Proceedings of the British machine vision conference*
25. Ferrari V, Marin-Jimenez M, Zisserman A (2008) Progressive search space reduction for human pose estimation. In: *CVPR*, June 2008

Chapter 41

Illicit Image Detection Using Erotic Pose Estimation Based on Kinematic Constraints

Mofakharul Islam, Paul Watters, John Yearwood, Mazher Hussain
and Lubaba A. Swarna

Abstract With the advent of the Internet along with sophisticated digital image processing technology, the Internet quickly became the principal medium for the distribution of pornographic content favouring pornography to become a drug of the millennium. With the advent of GPRS mobile telephone networks, and with the large scale arrival of the 3G networks, along with the cheap availability of latest mobile sets and a variety of forms of wireless connections, the internet has already gone to mobile, drives us toward a new degree of complexity. The detection of pornography remains an important and significant research problem, since there is great potential to minimize harm to the community. In this paper, we propose a novel approach to investigate and implement a pornography detection technique towards a framework for automated detection of pornography based on most commonly found erotic poses. Compared to the results published in recent works, our proposed approach yields the highest accuracy in recognition.

41.1 Introduction

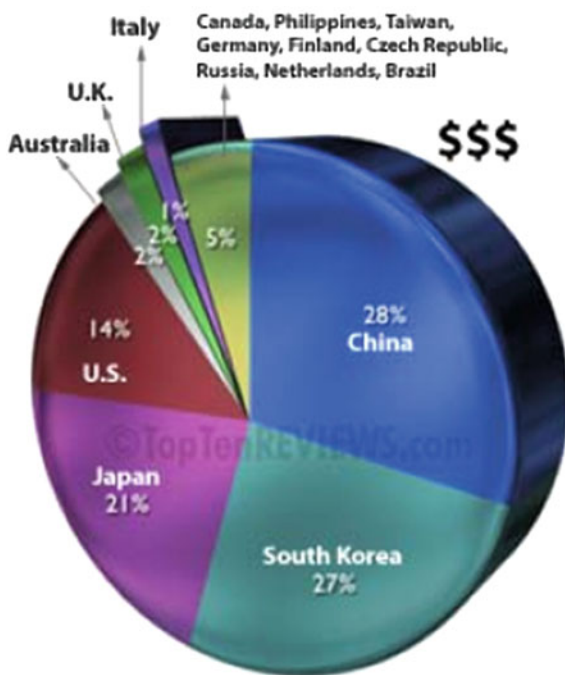
Internet access to the computers and a range of other communicating devices are increasingly a modern day must have for children and young people around the world. Not only are they establishing new cultural norms, they are also becoming

M. Islam (✉) · P. Watters · J. Yearwood
Internet Commerce and Security Laboratory, School of Science, Information Technology
and Engineering, University of Ballarat, Ballarat, Australia
e-mail: mlrj@mmmi.sdu.dk

M. Hussain
Melbourne Institute of Technology, Melbourne, Australia

L. A. Swarna
IBM Australia, Ballarat, Australia

Fig. 41.1 Worldwide pornographic revenue (courtesy: top ten reviews)



mainstream within education. Majority of adolescent worldwide spend a great majority of their time online to locate their educational stuff and as a source of perfectly innocent fun and games. But every child and young people who uses the Internet will almost certainly be exposed at some stage to material that will shock and possibly harm them, or they will come into contact with organization or individuals who mean them injury. To follow potentially catastrophic consequences throughout a children's lifetime, he/she needs only one such encounter to go wrong. It is therefore, no surprise that the Internet as well as the social networking sites pose a serious threat to our children who are more or less vulnerable, either some or all of the time while they are on the Internet. In fact, potential outcome of the Internet use is questionable as a little is understood yet of the potential problems and benefits associated with it, and the overall impact on the society from social benefit point of view that may arise. The risks posed by criminals who attempt to share, exchange, consume and produce child exploitation material, however, are fairly clear, and law enforcement is faced with the difficult task of trying to deal with the sheer volume of material (and offences) in a streamlined and systematic way.

Research [1] shows some of the statistics on pornography as shown in the pie-chart below (Fig. 41.1).

Apart from the exposure of age inappropriate content to minors, dealing with pornography in the workplace is a serious challenge for many large organizations but employing a block-all-images Internet browsing and email policy no longer

Table 41.1 Internet pornography statistics

| |
|---|
| * Every second—\$3,075.64 is being spent on pornography |
| * Every second—28,258 internet users are viewing pornography |
| * Every second—372 internet users are typing adult search terms into search engines |

provides a viable solution. Web and email are more media-based than ever before, and it is common for business document and mail to contain images such as logos, publicity shots etc. In a paper written by McGuire et al. [2] found that the viewing of pornography can serve as a source of a paraphilic “vivid sexual fantasy” which, when contemplated during masturbation, may condition men into perversion. Further, research also established a direct link between pornography consumption and crimes like sexual abuse, rape, violence, and child molestation [3–5].

Therefore, it is obvious that exposure of pornography and age-inappropriate materials, either legal or illegal have a devastating impact most often turns into a complex, often compulsive, psychosexual disorder with profound implications on us and our children who all use the internet or will do soon (Table 41.1).

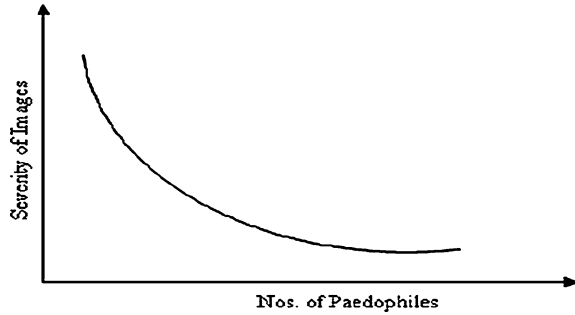
Research showed that most of the audience belongs to the lower levels that gradually progress towards higher levels. If we are able to put a stopping mechanism somewhere in this low level area, it will protect people from entering the exiting world of arousal and being pornography addicted.

Contemporary research established a link between child pornography and adult pornography. From that point of view, it will not only prevent people from viewing the most erotic and bestial images but also put an end for the potential paedophiles who are currently belongs to less severe groups to progress further towards most severe groups that eventually make them professional paedophiles at some stage. From the studied literatures, we find a progression model for pornography consumer as shown in Fig. 41.2.

So, there is a fundamental demand in real time classification of attached images to block inappropriate contents in a business or commercial environment. The problem therefore is the detection and prevention of certain types of images. Real-time detection of pornographic images can effectively prevents pornographic images from entering into the workplace via email and Internet browsing.

Prevention of explicit sexual content is an extreme challenge that paved the way to a growing industry aimed at blocking and filtering such contents. The software industry suggests two types of solutions to combat with the problem of pornography—collecting the adult web site addresses and collecting key words of adult web pages. While the former allows the user to access the requested page upon verifying its IP address, the latter performs the content analysis for providing access to the user. Here, the filtering programs block the pages from viewing upon finding any related key word of the web page in user’s computer. The drawback of these systems is that it is an uphill task to collect all adult web site addresses because hundreds of new web sites are launching every day. The drawback of the latter system is that it tends to block the sex education pages due to the key words used on these pages.

Fig. 41.2 Progression model for pornography consumer



In order to prevent access to pornographic sites, many commercial systems are available in the market. The commercial softwares/tools like Net-Nanny, CyberSitter, CyberPatrol and ChildWebGuardian allows access upon comparison of IP addresses/URLs and key words based on a long list of pornographic site IP address/URL and key words in their system databases. Although it is effective in blocking well-known pornographic sites and pages of pornographic links, it miserably fails in blocking pages containing pornographic image galleries since these most often do not contain links to other pages or objectionable text.

Other commercial systems such as ScreenShield, Snitch, System Recon and Enologic NetFilter Home in fact, deter the viewing of objectionable images rather than blocking where these systems consider the percentage of skin within an image to classify the image as to whether or not pornographic. Such an approach has been found not very accurate in practice and requires inspection by a human observer to make conclusion.

We are interested in the computational foundations of vision that helps us designing machine vision systems with applications to pornographic image detection. Our proposed project aims to find out the solutions for some fundamental questions in computer vision. How can we recognize persons, guns, cars, boats and many other suspicious categories of objects in cluttered pictures? How can we be trained these categories in the first place? Can we provide machines with ability akin to this?

This project primarily aimed at detection of pornography using erotic pose based on kinematic constraints. Erotic images focus on themes with either indicative, erotic or sensual scenes or subjects, sometimes with depictions of human nudity and lovemaking, but not always of an extremely explicit, gratuitous or pornographic nature. These kinds of films often appeal to the emotions of the viewer, with their emphasis on pleasure, physical desire, and human companionship.

The remainder of this paper proceeds as follows: In [Sect. 41.2](#) we present previous work in this specific field. In [Sect. 41.3](#), we describe our proposed pornography detection model, which estimates body part appearance models for pictorial structures using latent relationships between the appearances of different body parts. Experimental results demonstrating the accuracy and efficiency of the

proposed approach are discussed in [Sect. 41.4](#) and finally in [Sect. 41.5](#), we present our conclusion and future work.

41.2 Related Work

While significant research have already been conducted on pornography or illicit image detection most often with miscellaneous low and high level features, and other visual cues in fact, no research has looked closely at the subject.

In fact, all the pornography detection approaches are based on skin detection module that defines a decision rule based on which skin and non-skin pixels will be discriminated. A decision rule based on the distance of the pixel color to skin tone using a metric plays the crucial role in the discrimination process as this metric is defined by the skin color modelling technique. The final goal of skin-color detection is to choose a classifier that will discriminate between a skin and a non-skin. From classification point of view, skin-color detection can be viewed as a two class problem: skin pixel versus non-skin pixel.

Literature suggests different classifiers are employed by different authors for skin detection based on their respective contexts, where data distributions or data modelling played a crucial role. Our literature review suggests three specific types of skin modelling techniques—Explicitly defined skin region, Non-parametric distribution modelling, Parametric distribution modelling, and Adaptive modelling. In Explicitly Defined Skin Region techniques, threshold is a parameter that stipulates the values a pixel can be if it is to be considered as skin. In Nonparametric distribution modelling, Normalized Lookup Table (LUT), Na Bayes Classifier, and Self Organizing Map (SOM) are most often used as skin classifier. Single Gaussian, Mixture of Gaussians, Multiple Gaussian Clusters, and Elliptical Boundary Model are mostly used as Parametric Model for skin color detection. Adaptive skin model employs two or more skin-color models from the models as discussed above in combination or in sequence to make the technique robust against the varying conditions. In other word, it is called ‘Hybrid Model’.

Most of the adult image or pornographic detectors employed these skin color models based on different color space along with some other high level features like shape, color moment, color mean, region descriptors for their classification schemes.

A mathematical approach of grouping the images using the detected skin regions and extracted high level region features from these regions that might have link to the pornography is called pornographic classification, which is mainly focussed on classification of the skin region into benign or pornographic. We can broadly categorize pornographic image classifier into three groups: Supervised Machine Learning, Geometric Classifier, and Boosting Classifier. Supervised Machine Learning can be further categorized here under into four groups. Supervised Machine Learning can be further categorized here under into four groups—Support Vector Machine [6–8], Neural Network, Decision Tree [9], and

K-Nearest Neighbour [10]. Fleck et al. [11] demonstrated an affine-invariant model that define human as a set of rules describing how to assemble possible girdles and spine-thigh groups, where both the individual geometry of the body parts and the relationships between parts are constrained by the geometry of the skeleton which eventually provides an appropriate and effective model for human body recognition. Bootstrapping induces a classifier on a small set of labelled data and a large set of unlabeled data. Lee et al. [12] reported an increase of sensitivity from 81.74 to 86.29 % with boosting algorithm.

Cusano et al. reported better performance of SVM than multiple decision trees [6]. Bosson et al. [13] found that neural networks (83.9 specificity) gave slightly better results to that of k-NN and SVM. Bosson et al. attained 94.7 % sensitivity and 95.1 % specificity using NN with MPEG-7 Descriptors. Zheng et al. [8] found that DT is able to provide better accuracy than NN and SVM [13]. Xu et al. demonstrated superior performance of k-NN over NN [10].

Another approach aimed at robust pornography detection proposed by the same group of authors where pornography specific contextual constraints are employed to pornography detection but no comparisons are made as both of them are contemporary work and presented in the same conference [14].

Almost all the existing techniques proposed region descriptors that are in fact, not directly aimed at capturing contextual constraints representatives of pornography rather based on some heuristic high level statistical and geometrical features on skin region like area, size, shape, location, and their relationships in terms of some ratios, resulting poor performance in finding the actual pornographic stuff on images and thus failing to utilize its full potential in areas like detection of pornography. Further, the extracted features as employed in the existing work are found to be vulnerable under the inherent imaging artifact like intensity inhomogeneity, scale, rotation, transformation, occlusion, and camera viewpoint. These are the void in the literature that this research is attempting to fill.

41.3 Our Proposed Approach

Although the identification and detection of pornographic activity in an image is easy to a human observer, automatic and accurate identification is difficult and complex. Obtaining satisfactory detection results depends mainly on the ability of the descriptors in characterizing pornographic contextual constraints. There are sufficient reasons to believe that the pornographic images would not be of highest standard as these images were neither taken in ideal conditions nor taken by any skilled photographer. In addition, fears, tension, and panic associated with these sorts of heinous activity or crime are not conducive at all to have quality photographs, resulting poor quality images that suffer numerous imaging artifacts originated from intensity inhomogeneity, shadow, occlusion and unorthodox camera manoeuvrings. As a result, noisy images are the final output at pornographer's end that would eventually turns into input in our proposed project. Pre

processing i.e., noise removal, enhancement, and reconstruction are not practicable at all on multiple terabyte of image data on the Internet and being found in confiscated hard drives and other storage media by the LEAs. These noisy data poses a great challenge for the descriptors to be able to extract contextual contexts, representative of any specific context. In order to deal with such difficulty, we will employ erotic pose estimation to detect pornography rather than applying classification purely on extracted low or high level features.

Detecting humans pose in an image is noteworthy due to numerous reasons. Pose and/or pose sequence are extremely useful in depicting human's attitude and action.

Here, we propose a novel approach for estimating part appearance models from a single image. We employ generic detector to determine an approximate location in terms of location distribution and scale reference frame on the object as done in [15–17]. The basic motivation of our proposed approach is based on two observations—(i) some parts have rather stable location relative to the reference frame, (ii) different parts are statistically related in the appearance model. This implies that the appearance of some parts can be predicted from the appearance of other parts.

Over the last few years Pictorial Structure (PS) [1, 18, 19] has guided most of the research on articulated pose estimation paradigm. PS [1, 14, 16, 18, 19], are usually used for humans pose estimation though it is equally effective in any articulated object class like aeroplane, car, cow, lion etc. PS is basically a probabilistic model that portrays objects made of parts tied together by pair wise potentials encoding contextual constraints as prior. In addition, a unary potential measures part's position to generate an appearance model of the Part. Finally, MAP classifier infers spatial configuration of the parts (the pose of the object).

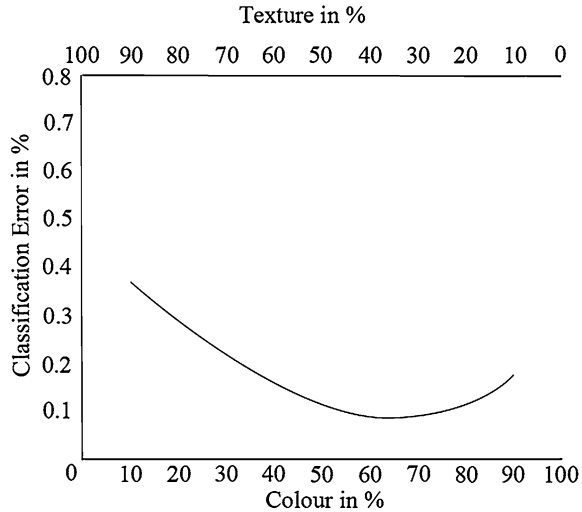
The relative location distribution of parts with respect to the reference frame and the dependencies between the appearances of different body parts are learned from training images with ground-truth pose annotated by a stickman. These relations are exploited generating appearance models for body parts given a new image. In determining the appearance model for more mobile parts (e.g. lower arms, lower legs), parts (torso) having higher location distribution with respect to the reference frame plays a crucial role.

A human body parts are represented by the general framework of pictorial structure, where body parts are tied together in conditional random field. Typically, parts l_i are rectangular image patches and their position is parameterized by location (x, y) , orientation θ , and scale s . The posterior of a configuration of parts $L = \{l_i\}$ given an image I is

$$P(L|I, \theta) \propto \exp\left(\sum_{i,j \in E} \psi(l_i, l_j) + \sum_i \phi(l_i|I, \theta)\right) \quad (41.1)$$

where, the $\psi(l_i, l_j)$ is a pair wise potential term correspond to a prior on the relative position of parts based on kinematic constraints (e.g. the upper arms and upper legs must be attached to the torso). The term $\phi(l_i|I, \theta)$ is an unary potential depicts the

Fig. 41.3 Mixing weights of colour and texture



local image evidence for a part in a particular position. How parts should look like entirely depends on appearance models θ that measure the dissimilarity between the image patch at l_i and the appearance model for part i . The appearance models are in fact, parameters of the PS and must be provided by an external mechanism.

We extend the model (41.1) by adding some new low level features to make it more accurate and robust under diverse imaging artifacts. Model (41.1) is primarily based on edge features and RGB colour features, which appears to be vulnerable while applied on unconstraint and cluttered images. We add texture features in addition to the existing color features to extend its ability to extract similarity of body parts appear on cluttered background. So, in addition to the color, texture is another feature that can be employed as a cue for image analysis. Texture features are constructed in such a way that they characterize local variation in intensity or color within the neighbourhood surrounding the pixel. As a result, a value is assigned to the pixels with a given texture. Research suggests, color and texture based analysis are in combination capable of producing more accurate results than either the color or texture feature used independently.

However, mixing weight of colour and texture, which is pertinent to accuracy in capturing image appearance throws an open problem. Up to date research, computer vision and image processing in particular do not provide any authenticated relationship between these two while applied in combination, which incite us to run an experiment to find out an optimal weight ratios of these two. In our experiment, we have sourced 100 pornographic images freely available on the Internet and run an Maximum Likelihood Estimation (MLE) based segmentation scheme on these images to segment skin regions on them. Misclassification rate are plotted against both colour and texture component to find a cogent mixing weight (Fig. 41.3).

Percentage of colour and texture components are taken at 5 % interval starting from 10 to 90 % and the graph is plotted on the mean classification error based on the segmentation results as obtained for the 100 images as chosen earlier. The plotted curve showing its minimum at 65 % colour and 35 % texture. So, we use this mixing weight of colour and texture in our work.

While the YC_rC_b space is employed as colour feature, Haar wavelet transforms are employed as texture features. Literature suggests a clear dominance of YC_rC_b , reason why we prefer it. Similarly, wavelet transform outperforms its other counterparts while applied in cluttered image having complex background. The wavelet transform has the capability to capture subtle texture information on objects appear on cluttered background more precisely than other texture descriptors [19].

YC_rC_b is the most popular choice, where ‘Y’ represents luminance computed as a weighted sum of RGB values, and chrominance values ‘Cb’ and ‘Cr’ computed by subtracting the luminance component from ‘B’ and ‘R’ values, offering us a significant advantage for utilising pure color information in the detection process. Using pure color information, a skin model or detector gets an ability to be able to adapt to the changes in the lighting and the viewing environment. The key advantage of being a lighting invariant color spaces made YC_rC_b space the most popular choice to the research community working on skin detection. Although there is a strong debate on whether or not the luminance component ‘Y’ playing any significant role in the detection process we discard it considering the luminance invariance approach of our proposed algorithm.

In this paper, our improvement mainly focussed on the unary potential, where we introduce a texture histogram in addition to the existing colour histogram to generate appearance model. So, unary potential takes the form;

$$\sum_i \phi(l_i|I, \theta) = \sum_i \delta * \phi_C(l_c|C, \theta_c) + \sum_i (1 - \delta) * \phi_T(l_t|T, \theta_t) \quad (41.2)$$

where, δ is a colour component weight, C is colour image, T is texture image, θ_c and θ_t are describing how parts should look like in terms of colour and texture respectively, l_c and l_t are image patch and texture patch respectively, and ϕ_C and ϕ_T are local evidence for a part in respect of colour and texture respectively.

A person detector’s detection windows is the input to our method [20].

Two observations are the main motivation of this particular approach: (i) some parts have relatively stable location in the detection window, $W = (x; y; s)$. For example, the torso is typically in the middle of an upper-body detection window; (ii) the different body parts appearances are related (e.g. the upper-arms often have the same color as the torso).

Now, two statistical observations—location prior of the body parts and appearance transfer mechanism are learned, where location prior holds distribution of the body part locations relative to the detection window and appearance transfer mechanism improves the models derived from the location prior by combining models for different body parts. The training data consists of images with ground-

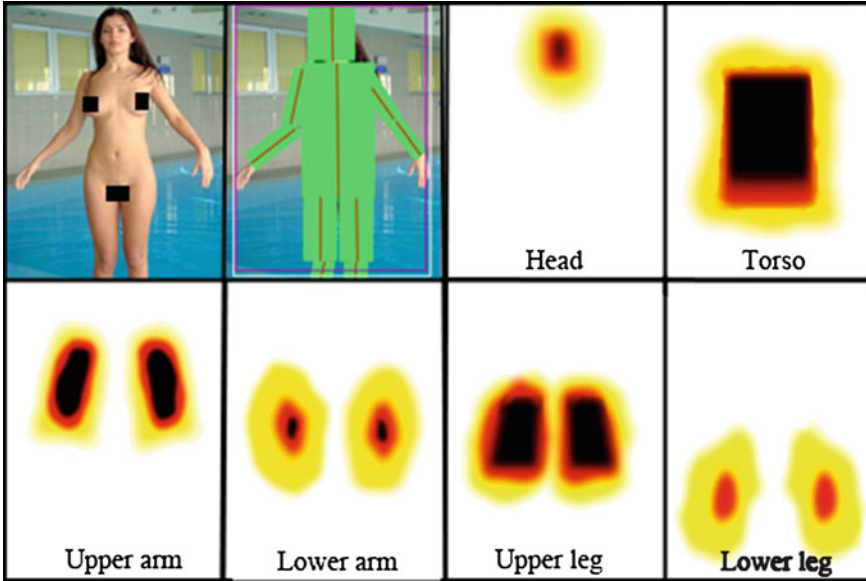


Fig. 41.4 Learning location priors. From top left **a** a training image **b** detection windows (magenta), window associated to the stickman (white), body part rectangles (green) obtained by widening the stickman line segments (red), **c** learnt location priors for different body parts

truth pose annotated by a stickman, i.e. a line segment for each body part (Fig. 41.4). Given a detection window and the learnt location priors, initial appearance models are estimated and then subsequently refined by the appearance transfer mechanism.

41.3.1 Learning Location Prior from the Training Images

For each body part i , we learn the prior probability

$LP_i(x, y) \in [1, 0]$ for a pixel (x, y) to be covered by the part before considering the actual image data (Fig. 41.4). These are in fact, location prior of the body parts with respect to the detection window/frame.

The LPs are learnt from training images with ground-truth pose annotated by a stickman (Fig. 41.4a). A generic person detector is initially run to find a person detection window. Now, association of stickmen to detection windows are made as shown in (Fig. 41.4b). Then all the training stickmen are projected with the common coordinate frame aligning the location and scale roughly. Finally, the maximum likelihood estimate is employed to learn the LPs for every pixels in the window. Example LPs are presented in (Fig. 41.4c) LPs for the head and torso are found quite sharp and localized, while LPs for the arms and legs are more diffuse.

As the lower arms and lower legs can move around freely, the location of these body parts appears very uncertain a priori.

41.3.2 *Transferring Appearance Models Between Body Parts Training Phase*

In our case, i.e., pornographic image, all the body parts have same colour due to existence of skin on them offering us to exploit this intrinsic relation between the appearances in prediction. Being inspired by the influence of the above relations, here we employ a transfer mechanism to combine the appearance models of different body parts, where the input appearance models are derived from estimated LPs. The final output of this transfer mechanism is a new appearance model of a part as a linear combination of the input appearance models of all parts. The newly estimated appearance model requires a mixing weight of part i , in the combination of part p . The appearance model is defined as;

$$AM_{pc}^{TM} = \sum_i w_{pc} AM_c^{LP} \quad (41.3)$$

where, AM_c^{LP} is the initial colour appearance model derived earlier from the colour location prior.

Similarly, we get the texture appearance model from the texture location prior based on the similar assumptions that all the body parts have similar skin texture;

$$AM_{pt}^{TM} = \sum_i w_{pt} AM_t^{LP} \quad (41.4)$$

We learn the mixing weights w_{pc} and w_{pt} by minimizing the squared difference between the appearance models produced by the transfer mechanism (AM_{pc}^{TM} and AM_{pt}^{TM}) and those derived from the ground-truth stickmen (AM^{GT}) as done in [21].

41.3.3 *Estimation of Appearance Model for a Test Image*

Estimation of good appearance models composed of two steps—colour & texture model and soft-segmentation based on the colour & texture models.

The colour & texture model estimation has three distinct steps. First, a standard coordinate frame is transformed from the detection window W cropping out of the input image and rescaling it to a fixed size to estimate the LPs. The next step is estimation of colour & texture models from the LPs. Finally, refined colour and texture models $\Pr(c \mid fg)$ and $\Pr(t \mid fg)$ are learned applying appearance transfers as in Eqs. (41.3) and (41.4) respectively.

Characterize the appearance of the body parts are estimated using the aforesaid colour & texture models. In addition, we also estimate here a background model $\Pr(c | bg)$ and $\Pr(t | bg)$ for each body part following Ramanan [22]. The learned foreground and background colour and texture models are employed to derive the posterior probability for a pixel to belong to a part i using Bayes theorem, assuming equal prior probability for foreground and background colour and texture.

$$P_i(fg|c) = \frac{P_i(c|fg)}{P_i(c|fg) + P(c|bg)} \quad (41.5)$$

$$P_i(fg|t) = \frac{P_i(t|fg)}{P_i(t|fg) + P(t|bg)} \quad (41.6)$$

Based on the posterior foreground probabilities for both colour and texture, a soft-segmentation for each body part is done on the image to generate a cue for the modified unary term in Eq. (41.2).

41.4 Experimental Results and Discussion

We have sourced the most common categories of erotic poses from uncontrolled pornographic images freely available on the Internet to run our experiment. This data is challenging due to the inherent imaging artifacts as mentioned in Sect. 41.1. Cluttered images, noise due to electro-magnetic interference, often dark illumination, persons appearing at a wide range of scales are some of the many difficulties adhered to image data. We have annotated 400 pornographic images having erotic poses, where roughly upright and approximately frontal images are only chosen. A person is annotated by a 10-part stickman (head, torso, upper and lower arms, upper and lower legs). The whole person must be visible in image. Now, LPs and mixing weights are computed as described earlier in this section.

Once LPs and mixing weights are computed from the training images, the proposed model is ready to estimate good appearance models for new test images.

For testing purpose, we have downloaded another 200 pornographic images of erotic categories from the Internet. Another 200 non-nude images from categories as sports, including aquatic sports, arts, and others with clothed people pictures having assorted poses are also downloaded from the Internet for testing purpose. Pose estimation procedure are followed as:

1. Detection of person using detection windows suggested in [16];
2. Estimation of part-specific color and texture models as described in previous section;
3. Detection window along with the detected person is subjected to image parsing engine [22] using directly our color and texture models in the unary potential to estimate the person's pose.

Table 41.2 Comparison of performance between the proposed and Belem and Cavalcanti [23] approach

| Method | TP | FP | FN | TN | Pre | Rec | Acc |
|---------------------------|-----|----|----|-----|------|------|------|
| Belem and Cavalcanti [23] | 161 | 61 | 39 | 139 | 0.73 | 0.81 | 0.75 |
| Proposed | 189 | 32 | 19 | 168 | 0.86 | 0.95 | 0.89 |

A popular measure, called Precision and Recall (Eq. 41.7), based on True Positive (TP), True Negative, False Positive (FP), and False Negative (FN) is applied on both—our experimental results and the results obtained from another pornography image detector proposed by Belem and Cavalcanti [23] to evaluate the performance of our novel approach. We selected Belem and Cavalcanti [23] for performance comparison as its authors reported more than 90 % accuracy. Apart from the Precision and Recall, we also compared accuracy (Eq. 41.8) of our approach with Belem and Cavalcanti [23] (Table 41.2).

$$Precision = \frac{TP}{TP + FP} \quad (41.7)$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{P + N} \quad (41.8)$$

We experienced a bit lower performances with Belem and Cavalcanti [23] than what claimed by the authors. The low performance might be attributed to the use of unconstraint images in this particular work.

41.5 Future Work and Conclusion

We have proposed a novel pornography detection approach, where erotic poses works in tandem with some low level features on human body images to detect pornography with maximum accuracy. We primarily base on Eichen and Ferrari's work [20], which primarily aimed at human pose estimation, but extended it with addition of texture feature on top of colour feature to foster the effectiveness of the proposed technique while dealing with unconstraint and cluttered images. While Eichen and Ferrari employed RGB colour features, we apply YC_bC_b space to make it more appropriate for skin detection. Further, luminosity component, 'Y' is discarded considering illumination or luminosity invariance. To the best of our knowledge, this is the first of its kind which is able to recognize pornography using erotic pose effectively with highest accuracy and thus paves the way for research in this area to not only help pornography detection, but also to contribute

significantly in security & surveillance, human body movement and pose identification, medical imaging and much more. The model may be extended further for detection of pornography based on other erotic poses (non-frontal and non-upright) that are not considered in this particular work.

References

1. Ropelato J (2009) Internet pornography statistics <http://internet-filter-review.toptenreviews.com/internet-pornography-statistics.html> Accessed March 2009
2. McGuire RJ, Carlisle JM, Young BG (1964) Sexual deviations as conditioned behaviour: a hypothesis. *Behav Res Ther* 2:185–190
3. Silbert MH, Pines AM (1984) Pornography and sexual abuse of women. *Sex Roles* 10: 857–868
4. Carter DL, Prentky RA, Knight RA, Vanderveer PL, Boucher RJ (1987) Use of pornography in the criminal and developmental histories of sexual offenders. *J Interpers Violence* 2: 196–211
5. Alexy EM, Ann WB, Robert AP (2009) Pornography use as a risk marker for an aggressive pattern of behavior among sexually reactive children and adolescents. *J Am Psychiatr Nurses Assoc* 14(6):442–453
6. Ruiz-del-Solar J, Castañeda V, Verschae R, Baeza-Yates R, Ortiz F (2005) Characterizing objectionable image content (pornography and nude images) of Speci_cWeb segments: chile as a case study. In: *Proceedings of the third Latin American web congress (LA-WEB'05, IEEE)*
7. Jeong C, Kim J, Hong K (2004) Appearance-based nude image detection. In: *Proceedings of the 17th international conference on pattern recognition (ICPR'04)*, IEEE Computer Society, Washington, DC, USA
8. Zheng QF, Zeng W, Wang WQA, Gao W (2006) Shape-based adult image detection. *Int J Image Graphics (IJIG)* 6(1):115–124
9. Chai D, Bouzerdoun A (2000) A Bayesian approach to skin color classification in YCbCr color space. In: *IEEE TENCON00*, vol 2, pp 421–424
10. Soriano M, MartinKauppi JB, Huovinen S, Laaksonen M (2003) Adaptive skin color modeling using the skin locus for selecting training pixels. *Pattern Recognit* 36(3):681–690
11. Fleck M, Forsyth A, Bregler C (1996) Finding naked people. In: *Computer Vision—ECCV'96*, 4th European conference on computer vision, Cambridge, UK, vol II, pp 593–602
12. Lee J-S, Kuo Y-M, Chung P-C, Chen E-L (2007) Naked image detection based on adaptive and extensible skin colour model. *Pattern Recognit* 40(8):2261–2270
13. Wang Y, Wang W, Gao W (2005) Research on the discrimination of pornographic and bikini images. In: *Proceedings of the 7th IEEE international symposium on multimedia (ISM'05)*
14. Islam M, Watters P, Yearwood J, Hussain M, Swarna L (2011) Illicit image detection: an MRF model based stochastic approach. In: *CISSE'2011*, Bridgeport, USA
15. Andriluka M, Roth S, Schiele B (2008) People-tracking-by-detection and people-detection-by tracking. In: *CVPR*
16. Ferrari V, Marin-Jimenez M, Zisserman A Progressive search space reduction for human pose estimation. In: *CVPR*, Jun 2008
17. Gammeter S, Ess A, Jaeggli T, Schindler K, Van Gool L (2008) Articulated multi-body tracking under egomotion. In: *ECCV*
18. Felzenszwalb P, Huttenlocher D (2005) Pictorial structures for object recognition. *IJCV* 61(1):55–79

19. Islam M (2008) Unsupervised color image segmentation using Markov random fields model. Master by Research Thesis, Graduate School of Information Technology and Mathematical Sciences, University of Ballarat, Australia
20. Ferrari V, Marin-Jimenez M, Zisserman A (2008) Progressive search space reduction for human pose estimation. In: CVPR, Jun 2008
21. Eichner M, Ferrari V (2009) Better appearance models for pictorial structures. In: British machine vision conference, London, UK
22. Ramanan D (2006) Learning to parse images of articulated bodies. In NIPS
23. Belem RJS, Cavalcanti JMB (2005) SNIF: a simple nude image finder. In: Proceedings of the third Latin American web congress (LA-WEB'05), IEEE

Chapter 42

Energy Efficient Public Key Cryptography in Wireless Sensor Networks

Vladimir Cervenka, Dan Komosny, Lukas Malina
and Lubomir Mraz

Abstract As Wireless Sensor Networks are getting to be more applied in commercial solutions (e.g. Smart Grid and Energy Saving Outdoor Lighting) the more important issue, is to provide proper security. This paper gives a comprehensive study of threats faced by Low Rate Wireless Personal Area Networks defined by IEEE 802.15.4. Although several security protocols have been developed to address this issue most of them are aimed just for a particular layer or application. Furthermore, they usually introduce great overhead in terms of energy and communication. Our effort is not just to point out security problems, but to propose energy efficient solution. We introduce and analyze energy efficient system providing symmetric key cryptography and public key cryptography so that both are possible to compute with help of advanced encryption standard hardware accelerator. We show that this hardware based solution is more than 100 times more energy efficient than purely software solution.

42.1 Introduction

Wireless Sensor Network (WSN) is becoming very popular as it has been plentifully deploying in miscellaneous applications taking advantages of their numerous features. Devices in these Low Rate Wireless Personal Area Networks

This paper was prepared within the framework of No. FR-TI2/571 grant project of the Ministry of Industry and Trade of the Czech Republic.

V. Cervenka (✉) · D. Komosny · L. Malina · L. Mraz
Department of Telecommunications, Brno University of Technology,
Brno, Czech Republic
e-mail: cervenka.v@phd.feec.vutbr.cz

(LR-WPAN) are characterized by short range, low bit rate, low power and low cost. The low power requirements allow utilizing of a wide range of alternative energy sources thus nodes could be easily placed almost everywhere not only for their small size. Many of these devices are limited in their computational power, memory and energy availability. Particular nodes communicate among themselves via radio links based on the IEEE 802.15.4-2006 standard. Moreover, 6LoWPAN protocol (defined in RFC 4944) provides the ability of TCP/IP communication so that it is possible to reach also IP enabled devices [1]. The common security mechanisms such as IPsec or TLS are generally not very appropriate due to their complexity and large computation and communication overhead.

42.1.1 Security Requirements

Cryptography ensures features or requirements such as data confidentiality, authentication, data integrity and non-repudiation. Generally, network security distinguishes three basic security features: data confidentiality, data integrity and availability. But security in wireless sensor networks needs to meet more specific requirements and thus adds: data freshness, time synchronization, secure localization, privacy or self-organization [2].

The set of security features depends on the purpose and intended architecture. The meaning of the basic additional security requirements in WSNs are listed below:

- Data confidentiality: only authorized recipients should understand the content of messages.
- Authentication: the identity of communicating entity must be correct and the one that it claims to be. Also data authentication exists and is defined as originality of data created by trusted entity.
- Data integrity: data which traverse to recipient should not be altered.
- Non-repudiation: a node cannot deny the authenticity of its signature.
- Availability: WSN services must be available at all times.
- Sequential freshness: old messages cannot be replayed by any adversary again.
- Time synchronization: the secure mechanism should be time-synchronized especially for collaborative WSN.
- Secure localization: each sensor should be secured properly against the false localization attack.
- Privacy: in a special case, sensors require the confidentiality of their identity.
- Self-organization: each sensor should be self-organizing and self-healing after topology changes or attempts of attacks.

Confidentiality is not an essential requirement for most applications; instead we are looking for a message authenticity. Let's assume a fire alarm system for example. The fact that fire broke out is not obviously a secret, but we need to be assure that the information comes from a trustworthy source. In other example, someone can be interested in economic benefits in terms of Smart Grids. People can try to tamper

power meters to reduce measured consumption. That is why the data authentication is needed.

Cryptographic techniques can be divided into two generic types, described in the next section.

42.1.2 Security Mechanisms

(1) *Symmetric key cryptography—symmetric cryptosystems*

Symmetric Key Cryptography (SKC) applies encryption and decryption operations for data confidentiality. Both communication participants must hold their keys in secret, because the encryption key is the same as the decryption key, and is easily derivable. Moreover, the secret keys must be securely distributed to all participants. Symmetric key cryptography is already defined in the MAC layer of IEEE 802.15.4 standard which uses advanced encryption standard (AES). Nevertheless, the secure distribution of the secret keys provided by key management system (KMS) is still a challenging task.

(2) *Public key cryptography—asymmetric cryptosystems*

Public key cryptography (PKC) essentially differs from the symmetric cryptosystem in key mechanism. The asymmetric cryptosystems use for encryption and decryption operations public key and different secret private key. Deriving the private key is computationally very unfeasible as discrete logarithm problem (DSA, Diffie Hellman protocol) or factoring problem (RSA) is often used. Given this fact, only private key could be kept in secret and the public key can be known for everyone. However, PKC is not very suitable for WSNs due to the fact that decryption operations could take a long time on constrained devices and have a great power consumption. It is well known that Elliptic Curve Cryptography (ECC) is better suited for the resource constrained devices because it can achieve the same level of security with a far smaller key size, thus reducing processing and communication overhead. For example the Elliptic Curve Digital Signature Algorithm (ECDSA) with 160-bit keys provides comparable security to RSA with 1,024-bit keys [3].

Although, the symmetric cryptography is highly suitable for WSN, a mechanism for the key establishment and authentication needs to be always considered. However, contemporary key managements are still very energy demanding. In our paper, we show the energy efficient method allowing use of both symmetric and asymmetric cryptography.

The remainder of the paper is organized as follows. In [Sect. 42.2](#), possible solutions for security in sensor networks and related work are discussed. [Section 42.3](#) presents an analysis of many security vulnerabilities of current protocols. In [Sect. 42.4](#), a review on platforms for wireless sensor networks is presented along with brief description of the chosen platform. [Section 42.5](#) describes the architecture of energy efficient cryptosystem that allows for

symmetric and asymmetric cryptography. The comparison of energy consumption of software and hardware based cryptography algorithms with obtained results are shown and discussed in Sect. 42.6. Finally, Sect. 42.7 concludes the paper results and outlines the future work.

42.2 Security Solutions in WSN and Related Work

The main problem in terms of security in WSN is the shared wireless medium. Wireless channel is wide open to attackers and any interested party within communication range can listen to the transmitted messages. Furthermore, low-cost requirements limit the physical safeguards and it is relatively easy to capture a node. There is no simple “protocol” solution so that additional hardware would not be required. Thus, the final cost would rise. The physical safeguard is the most challenging task which needs to be investigated more.

42.2.1 *The IEEE 802.15.4 Standard*

As the most WSN solutions are based on IEEE 802.15.4 standard (ZigBee, 6LoWPAN and partly WirelessHART and ISA100) it is important to consider its security capabilities. The IEEE 802.15.4 standard defines physical (PHY) and the Medium Access Control (MAC) protocols and interconnections of devices via radio communication in a LR-WPAN [4]. The maximum frame size is fixed to 127 bytes while 25 bytes are reserved for a frame overhead at MAC layer. Supported over-the-air data rates reaches up to 250 kbps for 2.4 GHz and other physical layers on 915 and 868 MHz are available as well. A channel access method utilizes both slotted and unslotted CSMA/CA.

The MAC sub-layer operates in one of three security modes: unsecured mode, secured mode or access control list (ACL) mode [5]. Frame security is a set of optional services that may be provided by the MAC; nevertheless, not every application needs them. The available security services are:

- Access control: allows MAC to select devices to communicate with and deny communication with other devices.
- Data encryption: offers symmetric cipher so only devices with the same key can decrypt the message. Encryption is provided to data payloads, beacon payloads and command payloads. Encryption of acknowledgement messages is not supported.
- Frame integrity: uses a message integrity code (MIC) to ensure that the message was not modified by other parties without cryptographic key. Frame integrity is provided, again, to data payloads, beacon payloads and command payloads.

Table 42.1 IEEE 802.15.4 security suites [5]

| Suite | Bits of integrity protection | Access protection | Encryption | Integrity | Sequential freshness |
|-------------------|------------------------------|-------------------|------------|-----------|----------------------|
| 1 AES-CTR | 0 | Yes | Yes | No | (Optional) |
| 2 AES-CCM-128 | 128 | Yes | Yes | Yes | (Optional) |
| 3 AES-CCM-64 | 64 | Yes | Yes | Yes | (Optional) |
| 4 AES-CCM-32 | 32 | Yes | Yes | Yes | (Optional) |
| 5 AES-CBC-MAC-128 | 128 | Yes | No | Yes | No |
| 6 AES-CBC-MAC-64 | 64 | Yes | No | Yes | No |
| 7 AES-CBC-MAC-32 | 32 | Yes | No | Yes | No |

- Sequential freshness: optional service, which offers very basic replay protection by appending sequence of values to the message.

These services are available in three security modes:

(1) *Secured mode*

There are seven security suites (see Table 42.1) which employ any of four security services (access control, encryption, integrity and freshness). As we suggested earlier, standard take advantage of AES with 128 bit keys and 128 bit block size. The AES-CTR stands for AES Counter mode whereas the AES-CCM utilizes AES-CTR and AES-CBC. Security suites 1–4 offer, optionally, sequential freshness results in 5 bytes of overhead. AES cipher blocks with integrity codes of 128, 64 and 32 in length are available for suites 2–7.

(2) *Access control list mode*

This mode provides an authentication service according to the access control list of devices allowed to communicate with. However, there is no cryptographic security and message source address can be easily spoofed.

(3) *Unsecured mode*

No security services are provided in this mode. This mode is suitable for special cases such as testing or operating within large secure areas.

A great advantage is that by reusing AES in a smart way, we are able to achieve all security services with only one cryptographic algorithm. However, IEEE 802.15.4 does not define the processes of key establishment/distribution or authentication. One possible solution is introduced in Sect. 42.5.

42.2.2 Pairwise Secret Key

The simplest solution of secret key establishment is to use a wide shared key, but a compromised node means a compromised network. Revealed secret key, even from a single node, can be used for decryption of any network traffic. A slightly

better solution is to use a wide shared key to establish a set of link pairwise keys and then erase the wide shared key. This technique, however, does not allow deploying of additional nodes after initialization. Other interesting solutions are based on variations of random key pre-distribution schemes. The very basic idea is to distribute random subset of the pool to each node. Two nodes have to find a common key in their pool in order to establish a session key and communicate with each other. The problem is when an adversary compromise a sufficient number of nodes the complete key pool can be easily reconstructed.

Furthermore, many advanced techniques were described in [6]. Liu and Ning presented framework for pairwise key establishment based on polynomial-based key pre-distribution scheme and probabilistic key pre-distribution scheme, which use polynomial pool instead of a key pool. Secret on each sensor node is generated from a subset of polynomials in the pool. Only nodes with the secret generated from the same polynomial can establish a pairwise key.

They also introduced novel schemes like a random subset assignment scheme or a hypercube-based scheme, which arranges polynomials in a hypercube space, assigns each sensor node to a unique coordinate in the space, and gives the secrets generated from the polynomials related to the corresponding coordinate [6]. Each sensor node can then identify whether it can directly establish a pairwise key with another node, or what intermediate nodes it can contact to indirectly establish the pairwise key. Advantages are resilience to node compromise and high probability to establish a pairwise key between non-compromised nodes, even if some nodes are compromised.

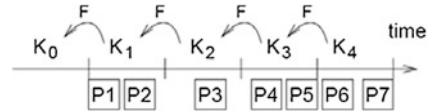
Nevertheless, two neighbor nodes are able to establish pairwise key only with some certain probability. When two neighbor nodes cannot establish a direct key, they need to find one or more intermediate nodes to help them. Hence, this technique may introduce additional communication with even more than two nodes. Moreover, after adding a new node to the network, the setup server performs the pre-distribution process for the new node and then it has to inform the deployed nodes which leads, again, to extra communication overhead. Pre-distribution schemes keys are not, still, a realistic option. When plenty of different devices need to communicate an automated key management has to be considered.

42.2.3 μ TESLA and Its Derivates

Many proposed solutions utilize asymmetric digital signatures for the authentication, but they suffer from high communication overhead of 50–1,000 bytes per packet [7]. One-time signatures based on symmetric cryptography are also challenge. Timed, Efficient, Streaming, Loss-tolerant Authentication Protocol (TESLA protocol) was originally aimed to provide efficient authenticated broadcast with regular desktop workstations. It has an overhead of approximately 24 bytes per packet and it authenticates the initial packet with a digital signature [7].

However, it was not designed for such constrained devices as in WSNs. That has given rise to μ TESLA design [7]. To authenticate a packet μ TESLA simply

Fig. 42.1 Time-released key chain for source authentication [7]



computes a message authentication code on the packet P_i with a key K_T which is a key of a key chain, generated by a public one-way function F such as MD5. The interesting part is that at the time when packet is received by node, the message authentication code key is known only by the sender (base station). The node stores the packet in a buffer until the base station broadcasts the verification key. Time is divided into time intervals and every interval is associated with one key of a key chain so that packets sent within one time interval are authenticated with the same key. To achieve this, nodes need to be loosely time synchronized and have one authentic key. The message authentication code key K_T can be prospectively computed by any subsequent key e.g. K_{T+2} by applying $K_T = F(F(K_{T+2}))$. Simultaneously, each receiver is also able to authenticate subsequent keys of the one-way key chain by the same procedure. Time-released key chain is visible in the Fig. 42.1.

However, μ TESLA protocol is computationally very expensive and it has high energy requirements as disclosed keys need to be broadcasted to all receivers and each key have to be re-computed. Also initial key chain commitments is unicast-based distributed and so introduce high communication overhead and we are not sure even about the robustness to compromised sensors. The simplest way to disrupt the communication is just to jam the communication channel only when disclosed keys are being transmitted.

Although, many μ TESLA modifications have been proposed, they still suffer from particular issues. In spite of some modifications trying to solve these issues (Multi-level μ TESLA, Tree-Based μ TESLA), they still do not solve the problem as a whole [6]. Multi-level μ TESLA introduce lower overhead, tolerance of message loss and resistance to denial-of-service (DoS) attacks. Tree-based broadcast authentication supports multiple senders and allows for revoking the broadcast capabilities. All the μ TESLA derivatives, however, need to be loosely time synchronized, which is not a trivial task to implement but very easy to disrupt.

42.3 The Analysis of Security Vulnerabilities of Current Protocols

42.3.1 Vulnerability of the Physical Layer

Goodspeed et al. [8] proposed Packet-in-Packet technique to remotely inject raw frames into wireless networks by exploitation of the PHY layer, inspired by Orson Welles authored attack in 1938. As an example, let's assume that Alice is a sender and Bob is a receiver. An adversary just needs only prepend its own packet to Alice's packet with preamble and sync, and then retransmit several times.

There are three possibilities at the receiver's side. In the first case, Bob receives every symbol and interprets the packet correctly. In the second case, adversary's packet is dropped along with the Alice's due to failed checksum. The third case makes this attack possible. When a symbol error occurs before the body part, then there is no checksum to cause the packet to be dropped and inner preamble plus sync will determine the start of the frame. Symbol errors within digital radio packets actually occur with relatively high frequency.

This sort of injection has been successfully tested on IEEE 802.15.4 and ANT + standards [8]. Although, there are number of complications which have to be taken into account in this approach, it has clearly showed a serious lack of security on the PHY layer. Almost any available cryptography is quite likely the best solution to injection attacks.

42.3.2 ZigBee Protocol

Many of ZigBee vulnerabilities have been mentioned and practically demonstrated along with tools for exploiting ZigBee and IEEE 802.15.4 which allows for eavesdropping on ZigBee networks, packet decoding, injecting and replaying traffic [9].

Similarly, numerous weaknesses such as data interception, masquerading by the third party as well as man-in-the-middle (MItM) attack against ZigBee derived Radio Frequency for Consumer Electronics (RF4CE) standard have been reported [10]. Shon et al. have proposed enhanced key agreement method for this standard. Their approach with two-phase key seed distribution is based on the device authentication of the IEEE 802.16 standard. First, the controller receives target's certificate and verify whether the certificate is provided from an authenticated target through the verification process of Certificate Authority (CA). Then, the controller sends its own certificate along with Unique ID (UID) and Seed Encryption Key Random number (SEEK_R). The UID and SEEK_R are encrypted by target's public key with the signature of a controller after applying hash-function. In the second phase, one of two proposed mode (quick or main) is performed to distribute 255 seed values in order to generate the link encryption key.

However, the proposed system has the additional computation cost of supporting PKC and high energy demands of transferring 255 seed values.

42.3.3 Pseudorandom-Number Generator Implementation

Although, architecture of security system might be well designed there are still many particular issues which should be considered such as the proper pseudorandom-number generator (PRNG) for cryptographic signatures and session keys.

Finnigin et al. [11] has presented a brute-force attack on an ECC system with a poor random-number generator (TinyOS). The results have shown that 50 % of the node's address space leads to a private key compromise in 25 min. An extreme example of a poorly implemented random-number generator causes ECC vulnerability has been presented by Goodspeed in [12]. The PRNG data generated in Z-Stack ZigBee Smart Energy Profile (version 2.2.2-1.30) initializing the ECC functions are actually predictable, repeated in one of two 32 kB cycles. It allows for eavesdropping on encrypted communications and cracking the AES key for symmetrical communication.

42.4 Platforms for Wireless Sensor Networks

42.4.1 Common Platforms

Probably the most used WSN platforms are TelosB/TmoteSKY, MICAz and IRIS. They have been designed to meet the LR-WPAN requirements, which means low bit rate, low power and low cost. They are equipped with only 8/16-bit microcontrollers operating at the maximum frequency of 8 MHz. This speed of processing is sufficient for most applications, and with help of the native AES support, it can offer quite reasonable cryptographic security.

Even though, there are efforts to implement PKC, particularly ECC, to these platforms like TinyECC [13], NanoECC [14], WM ECC [15] or [16] the proposed implementations are still very energy demanding due to long processing times. Moreover, microcontrollers usually work at the maximum clock speed (8 MHz) which require 3.6 V power supply [17].

The hardware architecture of the whole system is also very important, which outlines the results from [18]. The authors found out that even when the same chip is used, the computational time can be radically different. MICAz and TmoteSKY use the same Chipcon CC2420 transceiver with the support of AES. MICAz is, however, about 15 times faster than TmoteSKY.

42.4.2 Chosen Platform

For our implementation we chose the microcontroller EFM32G890F128 as the best options because of its extremely low power consumption, very powerful processing and hardware support for AES-128 and AES-256. This microcontroller is based on ARM Cortex-M3 core, includes a 32-bit RISC processor which can achieve as much as 1.25 DhrystoneMIPS/MHz, 128 kB Flash and 16 kB RAM [19]. In spite of these advantages, a unit price is not higher than the one of the microcontrollers mentioned in Table 42.2.

Table 42.2 Common WSN platforms

| Platform | MCU | Radio chip | AES support |
|----------|------------|------------|-------------|
| TelosB | TI MSP 430 | CC2420 | Yes |
| TmoteSKY | TI MSP 430 | CC2420 | Yes |
| MICAz | TI MSP 430 | CC2420 | Yes |
| IRIS | ATMega128 | AT86RF230 | No |

42.5 Cryptosystem Architecture

With all these pieces of knowledge in mind, we decided to make use of great possibilities of AES services defined in IEEE 802.15.4 standard, particularly the AES-CCM-64 with only 8 Bytes of overhead. However, as we mentioned earlier, IEEE 802.15.4 does not define the processes of key distribution or node authentication. Taking advantage of ECC on constrained devices (Sect. 42.2) Ephemeral Elliptic Curve Diffie-Hellman (ECDHE) could be used to establish shared keys and ECDSA (particularly the ECC-based TLS handshake) for authentication.

Noting the process power of 32-bit ARM processor on one hand, and the hardware support for AES-256 on the other, the AES-CCM ECC Cipher Suites for TLS proposed in draft [20] seems to be the most efficient solution for WSN field. The cipher suites use ECDHE as its key establishment mechanism and can be used with DTLS. They are based on the authenticated encryption with associated data (AEAD) algorithm defined in RFC 5116. Moreover, the AEAD_AES_128_CCM algorithm, used in this cipher suite, actually uses AES-128 as a block cipher. That is the key element which allows utilization of an AES hardware accelerator. The chosen cipher suite TLS_ECDHE_ECDSA_WITH_AES_128_CCM requires additional support for secp256r1 curve and SHA-256 hash algorithm.

Although, there are more ECC-based algorithms optimized for 32-bit processors available like a CompactECC [21, 22] but they are supposed to be proceeded by processor as whole. Instead, the AES-CCM ECC Cipher Suites make use of AES algorithm, which is highly advantageous in such a constrained environment as WSN. Furthermore, it is possible to utilize AES co-processor and so achieve extremely low power consumption.

42.6 Evaluation

The hardware AES operations on EFM32G890F128 are available in Energy Mode 0 (EM0) and Energy Mode 1 (EM1). The main processor is on, during the EM0 and can process other data, whilst in the EM1 the processor is in sleep mode, but the AES support is still available. Figure 41.2 shows the dependency of current consumption on processing time for both possible modes.

A comparison of the energy consumption of hardware and software based AES solutions is available in Fig. 42.3 and Table 42.3. For this comparison the software

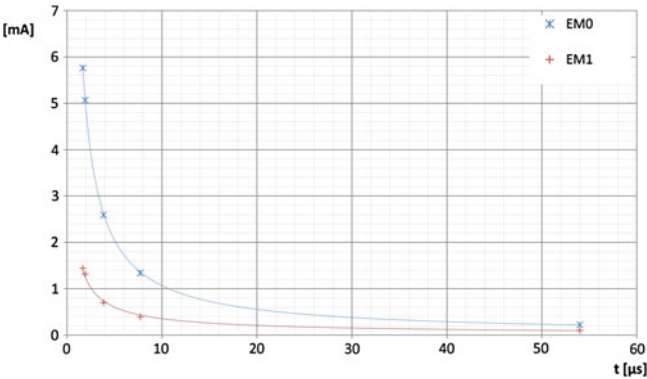


Fig. 42.2 Current consumption of hardware AES implementation on microcontroller EFM32G890F128. Encryption/decryption one 128-bit data block with 128 bit key. *EM0* energy mode 0 – run mode, *EM1* energy mode 1 – sleep mode. Measured for $V_{DD} = 3.0\text{ V}$

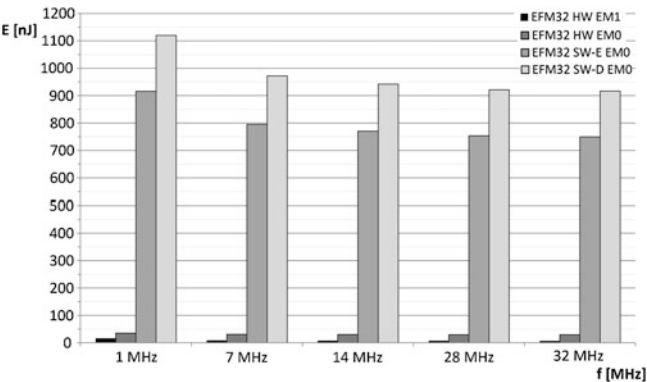


Fig. 42.3 Comparison of energy consumption: hardware versus software AES implementation. Encryption/decryption one 128-bit data block with 128 bit key. *EM0* energy mode 0 – run mode, *EM1* energy mode 1 – sleep mode, *HW* hardware encryption/decryption, *SW-E* software encryption; *SW-D* software decryption

implementation was based on [22]. This algorithm is specially optimized for architecture of ARM Cortex M-3. Even if we compare the best results for software encryption (749,520 nJ) and hardware encryption (7,290 nJ), at 32 MHz, it is clearly visible that hardware implementation reduces energy consumption more than by a factor of 100. For further comparison, the hardware implementation of AES encryption on MICAz consumes 1.83 and 14.30 μJ on TelosB [23]. The software implementation then consumes 39.08 and 28.16 μJ on MICAz and TelosB, respectively [23].

Table 42.3 Comparison of energy consumption

| Process | I [μ A] | V _{DD} [V] | P [W] | t [μ s] | E [nJ] |
|-------------------------|--------------|---------------------|----------|--------------|-----------|
| EFM32 HW EM1 @ 1 MHz | 103 | 3 | 0.000309 | 54.00 | 16,686 |
| EFM32 HW EM1 @ 7 MHz | 392 | 3 | 0.001176 | 7.71 | 9,072 |
| EFM32 HW EM1 @ 14 MHz | 700 | 3 | 0.002100 | 3.86 | 8,100 |
| EFM32 HW EM1 @ 28 MHz | 1,316 | 3 | 0.003948 | 1.93 | 7,614 |
| EFM32 HW EM1 @ 32 MHz | 1,440 | 3 | 0.004320 | 1.69 | 7,290 |
| EFM32 HW EM0 @ 1 MHz | 220 | 3 | 0.000660 | 54.00 | 35,640 |
| EFM32 HW EM0 @ 7 MHz | 1,337 | 3 | 0.004011 | 7.71 | 30,942 |
| EFM32 HW EM0 @ 14 MHz | 2,590 | 3 | 0.007770 | 3.86 | 29,970 |
| EFM32 HW EM0 @ 28 MHz | 5,068 | 3 | 0.015204 | 1.93 | 29,322 |
| EFM32 HW EM0 @ 32 MHz | 5,760 | 3 | 0.017280 | 1.69 | 29,160 |
| EFM32 SW-E EM0 @ 1 MHz | 220 | 3 | 0.000660 | 1,388.00 | 916,080 |
| EFM32 SW-E EM0 @ 7 MHz | 1,337 | 3 | 0.004011 | 198.29 | 795,324 |
| EFM32 SW-E EM0 @ 14 MHz | 2,590 | 3 | 0.007770 | 99.14 | 770,340 |
| EFM32 SW-E EM0 @ 28 MHz | 5,068 | 3 | 0.015204 | 49.57 | 753,684 |
| EFM32 SW-E EM0 @ 32 MHz | 5,760 | 3 | 0.017280 | 43.38 | 749,520 |
| EFM32 SW-D EM0 @ 1 MHz | 220 | 3 | 0.000660 | 1,697.00 | 1,120,020 |
| EFM32 SW-D EM0 @ 7 MHz | 1,337 | 3 | 0.004011 | 242.43 | 972,381 |
| EFM32 SW-D EM0 @ 14 MHz | 2,590 | 3 | 0.007770 | 121.21 | 941,835 |
| EFM32 SW-D EM0 @ 28 MHz | 5,068 | 3 | 0.015204 | 60.61 | 921,471 |
| EFM32 SW-D EM0 @ 32 MHz | 5,760 | 3 | 0.017280 | 53.03 | 916,380 |

Encryption/decryption one 128-bit data block with 128 bit key. *EM0* energy mode 0 – run mode, *EM1* energy mode 1 – sleep mode, *HW* hardware encryption/decryption, *SW-E* software encryption; *SW-D* software decryption

42.7 Conclusion and Future Work

In this paper we described the current security issues in WSNs and outlined possible solution for energy demanding key establishment. We investigated the efficiency of hardware accelerator for 128/256-bit AES encryption and decryption. Furthermore, we presented a method how to efficiently reuse this accelerator to achieve confidentiality, data integrity and even the key establishment with data origin authentication via Ephemeral Elliptic Curve Diffie-Hellman and Elliptic Curve Digital Signature Algorithms, while taking advantage of AES-CCM ECC Cipher Suites. We demonstrated that the hardware based cryptographic solution is more than 100 times more energy efficient than software based solution. Moreover, we used only one microcontroller so that no additional hardware like special hardware was needed. Since the memory requirements were not objective in this study, it was not considered for the investigation.

As a future work we plan to evaluate our system with other implementations. Also more complete solution based on behavior monitoring and trust management would be desirable.

References

1. Cervenka V, Komosny D, Kathiravelu G (2011) IETF 6LoWPAN and sensor networking. In: Proceedings of the 6th international conference on teleinformatics, pp 69–73
2. Sen J (2009) A survey on wireless sensor network security. *Int J Commun Netw Inf Secur* 1(2):3–10
3. Wander AS, Gura N, Eberle H, Gupta V, Shantz SC (2005) Energy analysis of public-key cryptography for wireless sensor networks. In: Proceedings of the 3rd IEEE international conference on pervasive computing and communication
4. Ludovici A et al (2009) Implementation and evaluation of the enhanced header compression (IPHC) for 6LoWPAN. In: EUNICE, pp 168–177
5. Gutiérrez JA, Callaway ED, Barrett RL (2003) Low-rate wireless personal area networks: enabling wireless sensor with IEEE 802.15.4. Standards Information Network IEEE Press, New York
6. Liu D, Ning P (2007) Security for wireless sensor networks. Springer, New York
7. Perring A et al (2002) SPINS: security protocols for sensor networks. *Wirel Netw* 8(5):521–534
8. Goodspeed T et al (2011) Packets in packets: Orson Welles in-band signaling attacks for modern radios. In: Proceedings of the 20th USENIX security symposium
9. Wright J (2009) KillerBee: practical zigbee exploitation framework. Presented at the 11th ToorCon conference, San Diego
10. Shon T et al (2011) A secure and robust connectivity architecture for smart devices and applications. *EURASIP J Wirel Commun Netw* 2011:200
11. Finnigin KM (2007) Cryptanalysis of an elliptic curve cryptosystem for wireless sensor networks. *Int J Secur Netw* 2(3–4):260–271
12. Goodspeed T (2011) PRNG vulnerability of Z-stack zigbee SEP ECC. <http://www.travisgoodspeed.blogspot.com/2009/12/prng-vulnerability-of-z-stack-zigbee.html>, 10 Sept 2011
13. Liu A, Ning P (2008) TinyECC: a configurable library for elliptic curve cryptography in wireless sensor networks. In: Proceedings of the 7th international conference on information processing in sensor networks, pp 245–256
14. Szczechowiak P et al (2008) NanoECC: testing the limits of elliptic curve cryptography in sensor networks. In: Proceedings of the 7th international conference on information processing in sensor networks, pp 305–320
15. Wang H, Li Q (2006) Efficient implementation of public key cryptosystems on mote sensors. In: Proceedings of the international conference on information and communication security, pp 519–528
16. Kargl A et al (2008) Fast arithmetic on ATmega128 for elliptic curve cryptography. International Association for Cryptologic Research Eprint archive, Oct 2008
17. Texas Instruments (2011) MSP430F16x, MSP430F161x mixed signal microcontroller datasheet, Oct 2002. Revised Mar 2011
18. Healy M, Newel T, Lewis E (2007) Efficiently securing data on a wireless sensor network. *J Phys Conf Ser* 76(1):012063
19. Energy Micro (2011) EFM32G890 datasheet. Revised May 2011
20. McGrew D et al (2011) “AES-CCM ECC cipher suites for TLS” draft-mcgrew-tls-aes-ccm-ecc-02
21. Aydos M et al (2000) An high-speed ECC-based wireless authentication protocol on an ARM microprocessor. In: Proceedings of the 16th annual computer security applications conference
22. Ekelund Ø (2009) Low energy AES hardware for microcontroller. MA thesis, Norwegian University of Science and Technology, Norway
23. Jongdeog Lee M et al (2010) The price of security in wireless sensor networks. *Comput Netw* 54(17):2967–2978

Chapter 43

Implementation of VLSB Stegnography Using Modular Distance Technique

Sahib Khan and Muhammad Haroon Yousaf

Abstract This work proposes a spanking new technique, Modular Distance Technique, to implement Variable Least Significant Bits Stegnography, in spatial domain, providing twofold security. It is an overriding and secure data embedding technique having low data hiding capacity of with least distortion. This is much immune to Steganalysis providing a large Key Size. This technique can be implemented with Euclidean, Chess Board and City Block distances with same data hiding capacity for each and contributing significantly to the key size. The key size of modular distance technique is almost 27 times of the size of the square of cover image. Low distortion made it difficult for intruder to detected hidden information and large key size make it difficult to extract the hidden information. This technique is contributing a data hiding capacity of 12.5–56.25 % with SNR ranging from 29.7 to 8 db. The hiding capacity and SNR varies with changing reference pixel, base of Mod and type of distance.

43.1 Introduction

This Stegnography is an art of concealed writing i.e. to hide useful information inside other risk-free cover file in a way that does not allow any snooper to even detect that there is a hidden information present [1] and transmitting hidden

S. Khan (✉)

Department of Electrical and Computer Engineering, Kohat University of Science and Technology, Taxila, Pakistan
e-mail: engr.sahib@kohat.edu.pk; engrsahib_khn@yahoo.com

M. H. Yousaf

Department of Computer Engineering, University of Engineering and Technology Taxila, Taxila, Pakistan
e-mail: haroon.yousaf@uettaxila.edu.pk

messages through innocuous cover carrier in such a manner that the existence of the embedded messages is imperceptible [2].

The history of Stegnography is very old; starting from Greeks till today it is used in a variety of applications. Greeks used it writing message on some material and later covering it with wax, tattooing messages on bald head, later growing hair to cover it up. In World War II invisible inks were used to write messages in between the lines of normal text message [3]. World War II saw the use of microdots by Germans. In microdots technology, photograph of secret message taken was reduced to size of a period. This technology was called “the enemy’s master piece of espionage” by FBI director Edgar Hoover [4].

Now a day’s Stegnography has found many applications [5, 6] and become an emerging research area, encompassing copyright protection, water marking [7, 8], fingerprinting and data hiding. A broad overview of data embedding and water-marking methods is available in [9]. Additional resources related to Stegnography and watermarking are obtainable at [10].

To implement Stegnography both in spatial domain or transform domain many different techniques have been introduced [11, 12]. Discrete Cosine Transform (DCT) is used and data is hidden by exploiting the coefficient of DCT of the cover image [13]. Wavelet Transform is also used for data hiding. The most important technique implemented in spatial domain is least significant bits (LSB) Stegnography utilizing the least significant bits of cover file elements (pixels, samples etc.). The 4 Least Significant (4LSB) Stegnography is one of the well-known techniques of this family. In 4LSB Stegnography four least significant bits of cover file elements are used for data embedding [14]. The 4LSB method is implemented for color bitmap images (24 bit and 8 bit i.e. 256 color palette images) and wave files as the carrier media. The 4LSB Stegnography [15, 16] is having a good enough data hiding capacity of 50 % and is used for the exchange of secret information over public channel in a safe way. All algorithms employed for any type of format have pros and cons and depend upon the environments used [17].

43.2 VLSB Stegnography

Although 4LSB Stegnography is very effective [18] but this technique is very insecure. Once a snooper come to know that hidden information is present it is very easy to retrieve the hidden information. To overcome this problem a new data hiding method is proposed called Variable Least Significant Bits (VLSB) Stegnography having variable data hiding capacity and distortion. This technique is used to hide variable amount of data in cover file in more secure way [19], instead of a fix data as in 4LSB Stegnography. To hide variable amount of data in cover file in more secure way using VLSB Stegnography a variable amount of data is hidden in every individual pixel of each sector of the cover file. VLSB Stegnography can be implemented in various ways depending on the algorithm devised. A proper algorithm is needed to decide that how much data should be hidden in which pixel or

group of pixels of cover image. VLSB Steganography has been implemented with Decreasing Distance Decreasing Bits Algorithm [20] with significantly increased hiding capacity, distortion and key size. The distortion created in cover image using DDDDB algorithm is non uniform over the cover the stego image and is very significant for larger hiding capacity.

To get a uniform and non-detectable distortion with a large key size a new technique, Modular Distance Technique, is presented in this work. The proposed technique is having a small data hiding capacity with negligible distortion and large key size and provides twofold security mechanism. Low distortion and large SNR made the hidden message imperceptible and large key size make it hard to retrieve the hidden information.

43.3 Modular Distance Technique

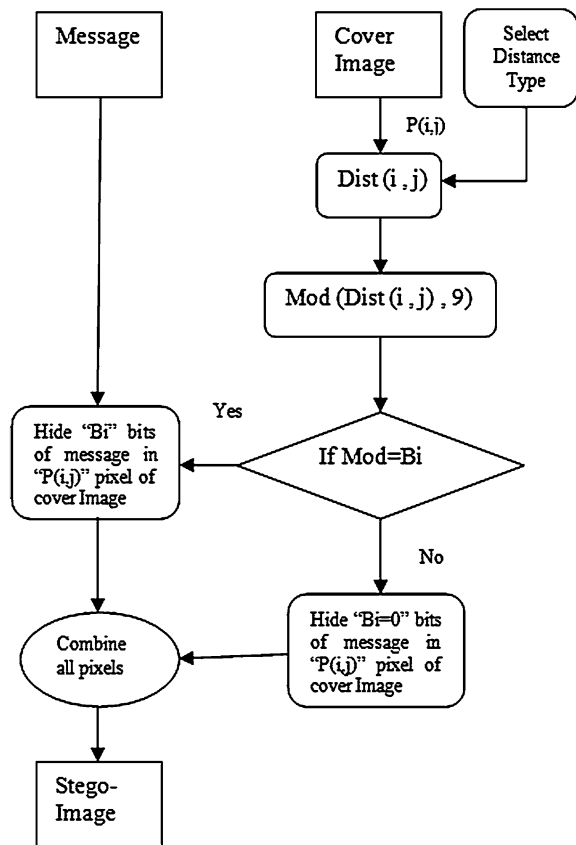
The main theme of VLSB Steganography is to hide variable amount of data in a cover file/image and an appropriate algorithm/technique is needed to implement it. Modular distance Technique (MDT) is one of such methodologies to implement VLSB Steganography with small data hiding capacity and large key size and signal to noise ratio (SNR).

MDT is a distance based technique it can be implemented with Euclidean, Chess Board and City Block distance. The provision of three types of distances is very important and contribute much to the key size because the snoopers is not aware of the type of distance used on the sender side in data hiding process.

To implement VLSB Steganography with MDT first of all a reference pixel/point usually the centre point is selected. Then the distance between the reference pixel and the pixel under process is calculated by using either Euclidean, Chess Board or City Block distances. MDT then takes the modulus of the calculated distance and on the bases of the value of modulus the number of bits “Bi” to be embedded in that pixel is decided. The number of bits to be substituted varies from 0 to 8 and any number of bits within this range can be embedded in a specific pixel. How much bits “Bi” are embedded in a pixel is the key to retrieve the hidden information. This contributes much to the security of the information. After all pixels of a cover image or the whole message are processed, all the pixels are combined to get a stego image. The stego image should be much closed to the original cover file in resemblance so that the snoopers doesn’t suspect about the presence of the hidden information.

The implementation of Variable Least Significant Bits Steganography with Modular Distance Technique is shown in block diagram in Fig. 43.1.

Fig. 43.1 VLSB steganography with modulus distance



43.3.1 Hiding Capacity

The data hiding capacity of VLSB Steganography using Modular Distance Technique varies with reference pixel, type of distance and the base of modulus. The gray scale image is a 2-D array of pixels having “R” number of row and “C” number of columns, so the total number of pixels “N” is:

$$N = R \times C \quad (43.1)$$

And each pixel’s intensity is represented by 8 bits. So the total size of the image in bits is:

$$\text{Size}_{\text{total}} = N \times 8 \quad (43.2)$$

Modular Distance technique provides us the liberty to embed any number of bits “Bi” ranging from 0 to 8 in a pixel of a cover image. The total data to be hidden in a cover image is dependent on the number of bits “Bi” substituted in each pixel. So the total bits embedded “Be” in the cover image are:

$$B_e = \sum_{i=1}^N B_i \quad (43.3)$$

The data hiding capacity “C” of VLSB Steganography using Modular Distance technique is:

$$C = \frac{B_e}{\text{Size}_{\text{total}}} \times 100 \quad (43.4)$$

43.3.2 Key Size

Modular Distance Technique is much secure method for the implementation of VLSB Steganography. This much immune to Steganalysis; because of it's of built-in encryption. The aim of Steganography is to hide data in a cover file in non-perceivable manner but if snoopers comes to know about the presence of secret, the snoopers has to try “K” different combinations to extract the hidden data exactly. The key size of MDT depends on the size of cover image and number of types of distances. An image with rows “R” and columns “C” will have a total of “N” pixels in the image. Using MDT we can hide a number of bits “Bi” ranging from 0 to 8 bits each pixel so there are 9 possible values for a single pixel and three choices of distance type in MDT. So the total key size “K” of Modular Distance Technique is:

$$K = 3 * (R * C) * C_1^9 \quad (43.5)$$

$$K = 3 * (R * C) * 9 \quad (43.6)$$

$$K = 27 * (R * C) \quad (43.7)$$

$$K = 27 * N \quad (43.8)$$

Where N: Size of cover image.

R: Number of rows of cover image

C: Number of columns of cover image

K: Number of possible keys (Key Size)

The reference point/pixel also contribute a lot to the key size as there are “N” number of pixels/point is cover image and any of the pixel can be used as a reference. The reference point used on the sender side during data hiding process should be kept secret without the knowledge of an exact reference the retrieval of data/message is impossible. Now if the reference point is considered then the key size will be:

$$K = 27 * N * N \quad (43.9)$$

$$K = 27 * N^2 \quad (43.10)$$

$$K = 27 * (R * C)^2 \quad (43.11)$$

So there are a total “K” number of possible ways to hide data in a cover image. Larger the value of “K” more difficult it would be for an unauthorized person to extract data from the Stego-Image even if the snoopers came to know that some data is hidden in the image he must have to try “K” various combinations to retrieve data exactly.

43.3.3 SNR and PSNR

The aim of Stegnography is to hide information in a cover image in such a manner that no one detects the presence of hidden data. For a best Stegnographic technique the stego should be closely resemble to the cover image. The quality of the stego-image is measured quantitatively by calculating signal to noise ratio (SNR) [21, 22] and peak signal to noise ratio (PSNR) [21, 22]. SNR for a Stego image in Decibels is calculated as:

$$SNR = -10 \log \left[\frac{\sum((Cover - Stego)^2)}{\sum((Cover)^2)} \right]^{-1} \quad (43.12)$$

And PSNR for a stego image in Decibels is calculated as:

$$PSNR = -10 \log(Mean((Cover - Stego)^2)) \quad (43.13)$$

43.4 Implantation

The results presented in this paper is obtained by implementing the VLSB Stegnography using Modular Distance Technique (MDT) by selecting centre pixel of cover image as reference pixel and calculating the distances of all the pixels are calculated with respect to the reference point. Then mod 8 of the distance of each pixel is taken. If the mod 8 of the distance is 0, 8 bits are used for data hiding in that pixel if mod 8 are 1, 2, 3, 4, 5, 6 and 7, then 7, 6, 5, 4, 3, 2 and 1 bits are used for data hiding respectively. The data hiding capacity, SNR and PSNR of the stego image are calculated. Then the reference point was changed the same combination was applied to calculate Capacity, SNR and PSNR. The results obtained using different type of distances; bases of mod and reference point are given in detail in experimental results section.

43.5 Experimental Results

As I discussed in the earlier section that Modular Distance Technique can be implemented using three different types of distance i.e. Euclidean distance, Chess Board distance and City Block distance. In this section the results of Variable Least Significant Bits Steganography using MDT method with different bases of mod function and with different reference point/pixel are presented one by one in details.

43.5.1 Results of MDT Using Euclidean Distance

To scrutinize the effect of change in the reference point/pixel on the Hiding Capacity, SNR and PSNR, experiments were performed. Using MDT base of the Mod is kept constant and reference point is changed. The results are obtained for Mod 8 and with reference point centre, (0, 0), (1, 10), (10, 1), (100, 50) and (500, 500) and the stego images obtained are shown in Fig. 43.2(a, b, c, d, e and f) respectively. The hiding capacity, SNR and PSNR are listed in Table 43.1. It has been observed from the experimental results that the parameters (Capacity, SNR and PSNR) vary with changing reference points.

Now to evaluate the outcome of VLSB Steganography Implemented with MDT for changing the value of Base of Modulus of distance. To get Hiding Capacity, SNR and PSNR, experiments were performed. Using MDT base of the reference point is kept fixed i.e. (10, 1) and Base of the Modulus is altered. The results are obtained for Base 16, 8, 4, 2 and 1 with a fixed reference point (10, 1) and the stego images obtained are shown in Fig. 43.3(a, b, c, d and e) respectively. The hiding capacity, SNR and PSNR are listed in Table 43.2. It has been observed from the experimental outcome that the Hiding Capacity increases with decreasing Base of Mod while SNR and PSNR decreases.

43.5.2 Units Results of MDT Using Chess Board Distance

To observe the effect of change in the reference point/pixel on the Hiding Capacity, SNR and PSNR, experiments were performed. Using MDT base of the Mod is kept constant and reference point is changed. The results are obtained for Mod 8 and with reference point centre, (0, 0), (1, 10), (10, 1), (100, 50) and (500, 500) and the stego images obtained are shown in Fig. 43.4(a, b, c, d, e and f) respectively. The hiding capacity, SNR and PSNR are listed in Table 43.3. It has been observed that all the parameters vary with varying reference point.

Now to evaluate the effect of change in the value of Base of Modulus of distance on the Hiding Capacity, SNR and PSNR, experiments were performed. Using MDT

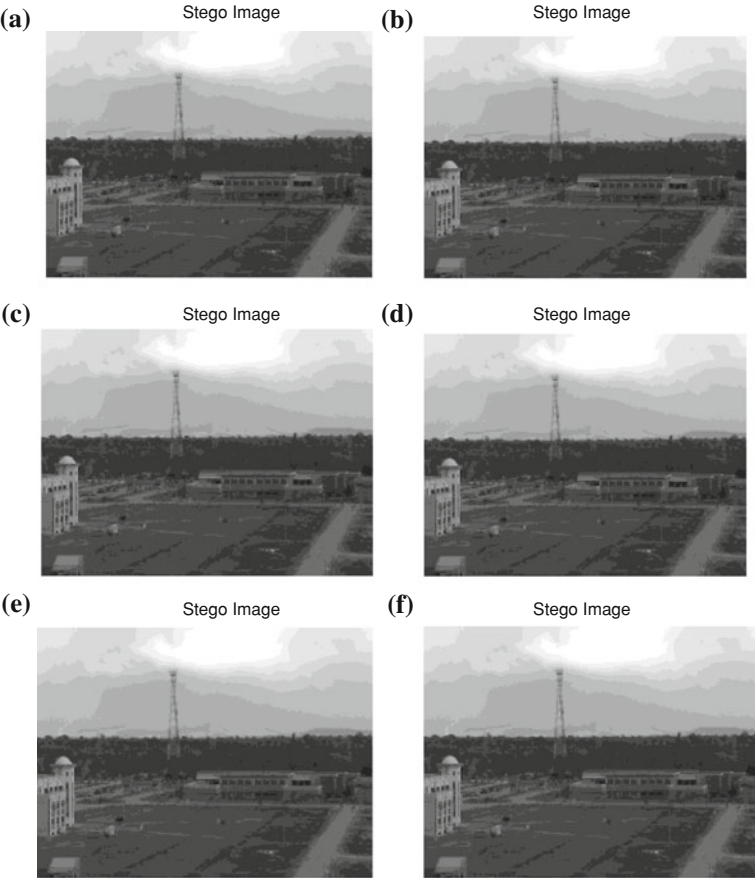


Fig. 43.2 The resulted stego images from VLSB steganography using MDT with Euclidean distance for fixed base of modulus and varying reference point/pixel **a** Stego image with centre as a reference point **b** Stego image with (0, 0) as a reference point **c** Stego image with (1, 10) as a reference point **d** Stego image with (10, 1) as a reference point **e** Stego image with (100, 500) as a reference point **f** Stego image with (500, 500) as a reference point

| Table 43.1 Capacity, SNR and PSNR using Euclidean distance with varying reference points | | | | | |
|---|-----|------------|-----------|---------|--------|
| Sr. No. | Mod | Reference | Capacity | SNR | PSNR |
| 1 | 8 | Centre | 12.5000 % | 29.7496 | 5.6852 |
| 2 | 8 | (0, 0) | 12.5070 | 29.6667 | 5.6022 |
| 3 | 8 | (1,10) | 12.5168 | 29.4023 | 5.3378 |
| 4 | 8 | (10, 1) | 12.5168 | 29.4014 | 5.3369 |
| 5 | 8 | (100, 50) | 12.5169 | 29.3964 | 5.3319 |
| 6 | 8 | (500, 500) | 12.5200 | 29.5388 | 5.4744 |

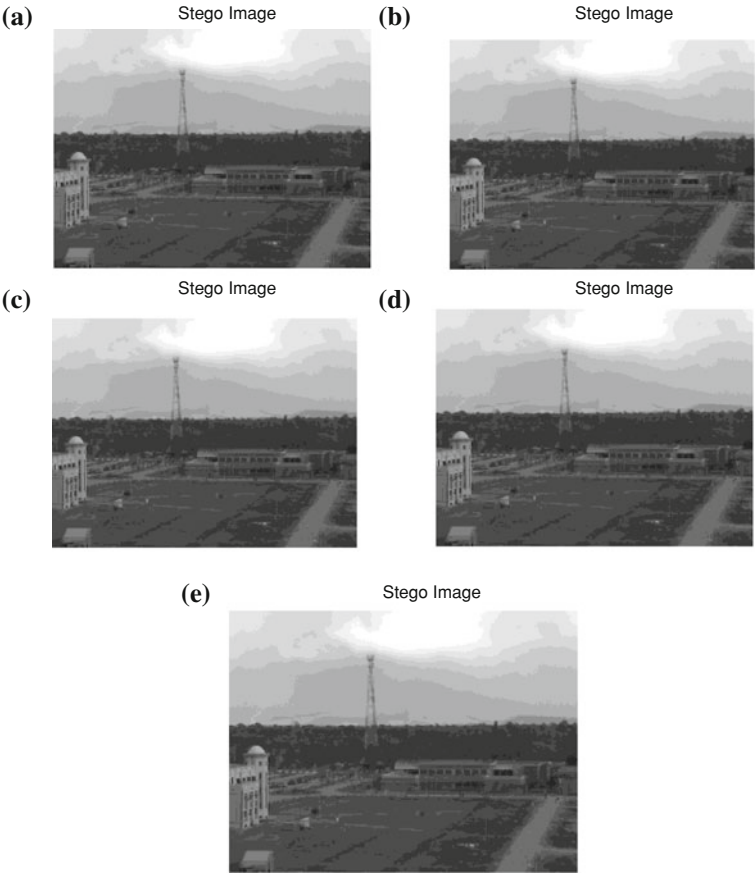


Fig. 43.3 The resulted stego images from VLSB Steganography using MDT with Euclidean distance for fixed reference point (10, 1) and varying bases of Mod **a** Stego image with Mod 16 **b** Stego image with Mod 8 **c** Stego image with Mod 4 **d** Stego image with Mod 2 **e** Stego image with Mod 1

| Table 43.2 Capacity, SNR and PSNR using Euclidean distance with varying base of mod | | | | | |
|--|-----|-----------|--------------|---------|--------|
| Sr. No. | Mod | Reference | Capacity (%) | SNR | PSNR |
| 1 | 16 | (10, 1) | 12.5081 | 29.5760 | 5.5115 |
| 2 | 8 | (10, 1) | 12.5168 | 12.5168 | 5.3369 |
| 3 | 4 | (10, 1) | 12.5350 | 29.0572 | 4.9927 |
| 4 | 2 | (10, 1) | 12.5734 | 28.4121 | 4.3476 |
| 5 | 1 | (10, 1) | 12.6537 | 27.3269 | 3.2624 |

base of the reference point is kept fixed i.e. (10, 1) and Base of the Modulus is changed. The results are obtained for Base 16, 8, 4 and 3 with a fixed reference point (10,1) and the stego images obtained are shown in Fig. 43.5(a, b, c and d)

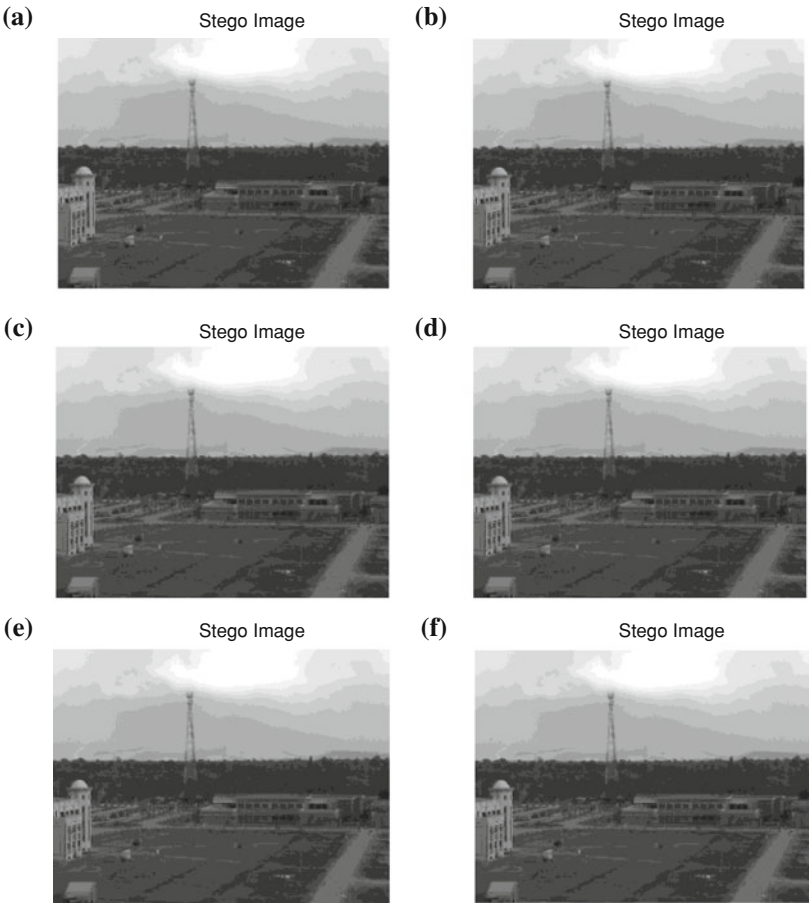


Fig. 43.4 The resulted stego images from VLSB steganography using MDT with Chess Board distance for fixed base of modulus and varying reference point/pixel **a** Stego image with centre as a reference point **b** Stego image with (0, 0) as a reference point **c** Stego image with (1, 10) as a reference point **d** Stego image with (10, 1) as a reference point **e** Stego image with (100, 500) as a reference point **f** Stego image with (500, 500) as a reference point

Table 43.3 Capacity, SNR and PSNR using Chess Board distance with varying reference points

| Sr. No. | Mod | Reference | Capacity (%) | SNR | PSNR |
|---------|-----|------------|--------------|---------|----------|
| 1 | 8 | Centre | 12.5 | 29.7496 | 5.6852 |
| 2 | 8 | (0, 0) | 23.4670 | 13.9030 | −10.1615 |
| 3 | 8 | (1, 10) | 23.4122 | 13.9070 | −10.1575 |
| 4 | 8 | (10, 1) | 23.4123 | 13.9064 | −10.1580 |
| 5 | 8 | (100, 50) | 23.4252 | 13.8950 | −10.1694 |
| 6 | 8 | (500, 500) | 23.4333 | 13.8959 | −10.1686 |

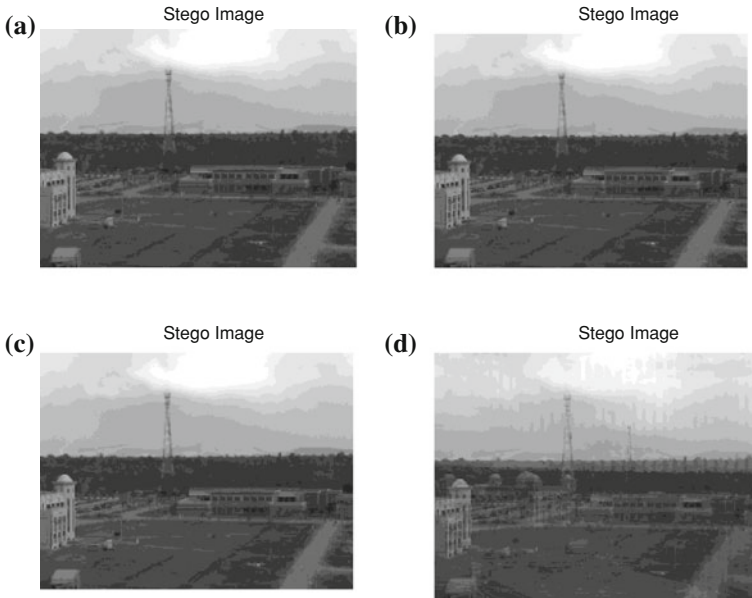


Fig. 43.5 The resulted stego images from VLSB steganography using MDT with Euclidean distance for fixed reference point (10, 1) and varying bases of Mod **a** Stego image with Mod 16 **b** Stego image with Mod 8 **c** Stego image with Mod 4 **d** Stego image with Mod

respectively. It is observed experimentally that Mod 2 and Mod 1 create lots of distortion with very large data hiding capacity. Due the no affordable distortion Mod 2 and Mod 1 can't used for Stegnographic purposes in this proposed using Chess Board distance. The hiding capacity, SNR and PSNR are listed in Table 43.4. It has been observed that all the parameters vary with varying the value of Base.

43.5.3 Equ Results of MDT Using City Block Distance

To observe the effect of change in the reference point/pixel on the Hiding Capacity, SNR and PSNR, experiments were performed. Using MDT base of the Mod is kept constant and reference point is changed. The results are obtained for Mod 8 and with reference point centre, (0, 0), (1, 10), (10, 1), (100, 50) and (500, 500) and the stego images obtained are shown in Fig. 43.6(a, b, c, d, e and f) respectively. The hiding capacity, SNR and PSNR are listed in Table 43.5. It has been observed that all the parameters vary with varying reference point.

Now to evaluate the effect of change in the value of Base of Modulus of distance on the Hiding Capacity, SNR and PSNR, experiments were performed. Using MDT base of the reference point is kept fixed i.e. (10, 1) and Base of the Modulus is changed. The results are obtained for Base 16, 8, 4, 2 and 1 with a fixed

Table 43.4 Capacity, SNR and PSNR using Chess Board distance with varying base of mod

| Sr. No. | Mod | Reference | Capacity (%) | SNR | PSNR |
|---------|-----|-----------|--------------|---------|---------|
| 1 | 16 | (10, 1) | 17.9561 | 16.7826 | −7.2819 |
| 2 | 8 | (10, 1) | 23.4123 | 13.9064 | −10.158 |
| 3 | 4 | (10, 1) | 34.3582 | 10.9521 | −13.112 |
| 4 | 3 | (10, 1) | 41.6555 | 9.7148 | −14.349 |

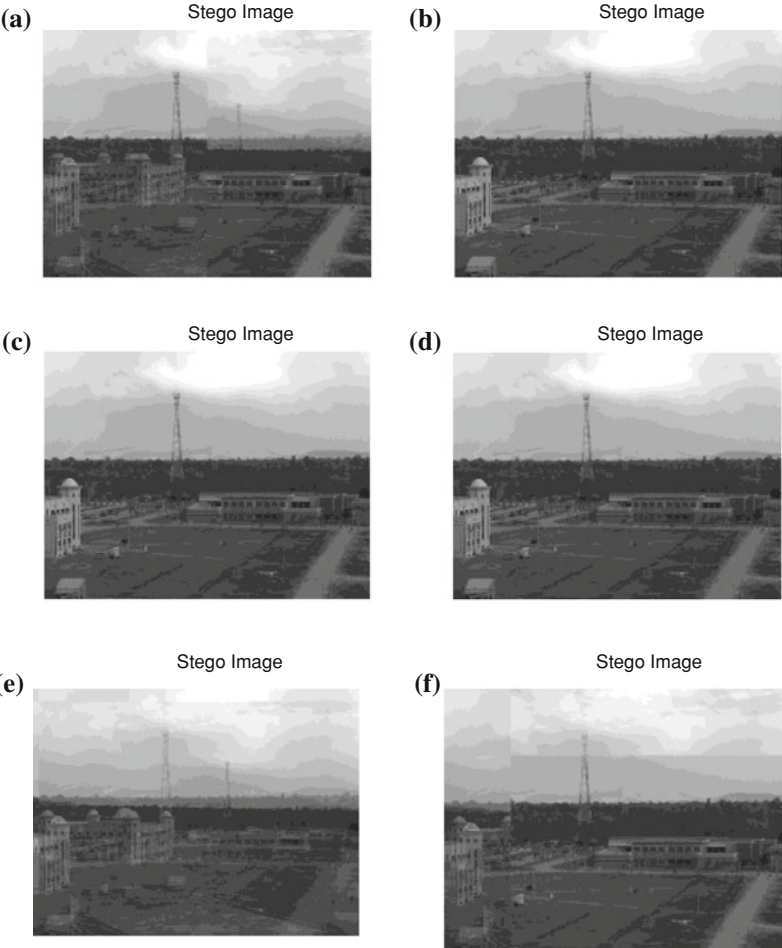


Fig. 43.6 The resulted stego images from VLSB steganography using MDT with City Block distance for fixed base of modulus and varying reference point/pixel **a** Stego image with Centre as a reference point **b** Stego image with (0, 0) as a reference point **c** Stego image with (1, 10) as a reference point **d** Stego image with (10, 1) as a reference point **e** Stego image with (100, 500) as a reference point **f** Stego image with (500, 500) as a reference point

Table 43.5 Capacity, SNR and PSNR using City Block distance with varying reference points

| Sr. No. | Mod | Reference | Capacity (%) | SNR | PSNR |
|---------|-----|------------|--------------|---------|----------|
| 1 | 8 | Centre | 23.4375 | 13.9057 | −10.1588 |
| 2 | 8 | (0, 0) | 23.4375 | 13.9052 | −10.1593 |
| 3 | 8 | (1, 10) | 23.4375 | 13.9020 | −10.1625 |
| 4 | 8 | (10, 1) | 23.4375 | 13.9019 | −10.1626 |
| 5 | 8 | (100, 50) | 23.4375 | 13.9026 | −10.1619 |
| 6 | 8 | (500, 500) | 23.4374 | 13.9046 | −10.1599 |

Table 43.6 Capacity, SNR and PSNR using City Block distance with varying base of mod

| Sr. No. | Mod | Reference | Capacity (%) | SNR | PSNR |
|---------|-----|-----------|--------------|---------|----------|
| 1 | 16 | (10, 1) | 17.9688 | 16.8023 | −7.2622 |
| 2 | 8 | (10, 1) | 23.4375 | 13.9019 | −10.1626 |
| 3 | 4 | (10, 1) | 34.3750 | 10.9495 | −13.1150 |
| 4 | 2 | (10, 1) | 56.2500 | 7.9680 | −16.0965 |

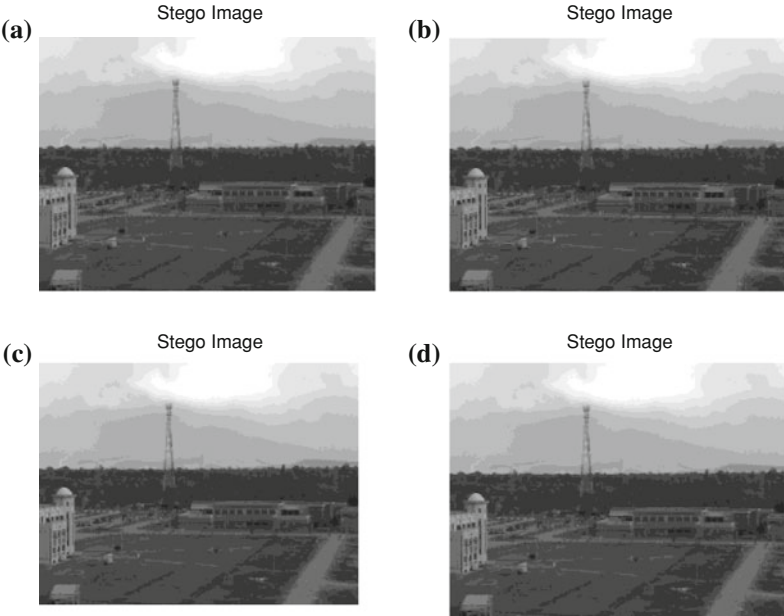


Fig. 43.7 The resulted stego images from VLSB stegnography using MDT with City Block distance for fixed reference point (10, 1) and varying bases of Mod **a** Stego image with Mod 16 **b** Stego image with Mod 8 **c** Stego image with Mod 4 **d** Stego image with Mod 2

reference point (10, 1) and the stego images obtained are shown in Fig. 43.7(a, b, c and d) respectively. It is observed experimentally that Mod 1 creates lots of distortion with very large data hiding capacity. Due the no affordable distortion Mod 1 can't be used for Stegnographic purposes in this proposed using City Block

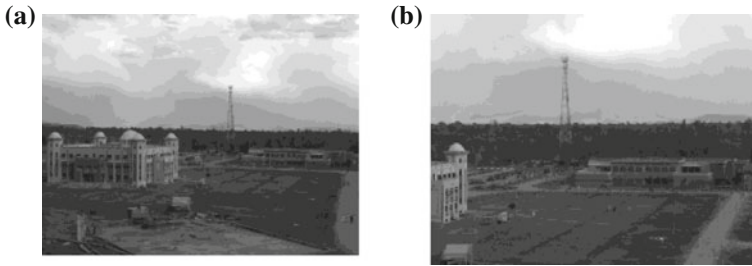


Fig. 43.8 Cover image used in steganographic process and the stego image resulted by 4LSB steganography. **a** Cover image **b** Stego image of 4LSB

distance. The hiding capacity, SNR and PSNR are listed in Table 43.4. It has been observed that all the parameters vary with varying the value of Base.

Now on comparison of all the stego images with original cover image and Stego image of 4LSB Steganography shown in Fig. 43.8 (a and b), it can be seen that all the stego images obtained by VLSB Steganography using Modular Distance Technique closely resemble the cover image and the quality of the stego images is much better than that of 4LSB Steganography's Stego image. But at the same time it can also be seen that for some specific values of reference point and base of Mod, the stego images are much distorted even with low data hiding capacity as shown in Figs. 43.5d and 43.6 (e and f). The SNR and Capacity vary with the input parameters. The appropriate selection of Reference Pixel and Base for Mod is the key to fine results.

43.6 Conclusion

After In this paper I have proposed an efficient technique to implement VLSB Steganography. The proposed scheme has a large key size and Signal to Noise ratio. The stego image quality is very high, closely resembles to the cover image, so that hidden information are invisible. Due to good enough SNR and large Key size the MDT is providing twofold security. The hiding capacity and SNR varies with both base of Mod and location of reference point/pixel. The selection of appropriate values of both of these parameters is very important and plays a vital role in the whole Steganographic process.

References

1. Moon SK, Kawitkar RS (2007) Data security using data hiding, International conference on computational intelligence and multimedia applications 2007, pp 247–251
2. Dumitrescu S, Wu WX, Memon N (2002) On steganalysis of random LSB embedding in continuous-tone images. In: Proceedings of international conference on image processing, Rochester, pp 641–644

3. Cedric T, Adi R, Mcloughlin I (2000) Data concealment in audio using a nonlinear frequency distribution of PRBS coded data and frequency-domain LSB insertion. In: Proceedings of IEEE international conference on electrical and electronic technology, Kuala Lumpur, Malaysia, pp 275–278
4. Kahn D (1967) The codebreakers. Macmillan, New York
5. Moon SK, Vasnik VN Application of steganography on image file, National conference on recent trends in electronics, pp 179–185
6. Morkel T, Eloff JHP, Olivier MS An overview of image steganography, information and computer security architecture (ICSA) research group, Department of Computer Science, University of Pretoria
7. Zheng J-B, Feng DD, Zhao R-C (2005) A multi-channel framework for image watermarking, Proceedings of the fourth international conference on machine learning and cybernetics, Guangzhou, 18–21 Aug 2005
8. Uccheddu F, Corsini M, Barni M (2004) Wavelet-based blind watermarking of 3D models, Proceedings of ACM multimedia and security workshop, pp 143–154
9. Mehboob B, Faruqi RA (2008) A steganography implementation. IEEE 2008
10. Swanson MD, Kobayashi M, Tewfii AH (1998) Multimedia data embedding and watermarking technologies. Proc IEEE 86(6):1064–1087
11. Anderson R (ed.) (1996) Information hiding: first international workshop, Cambridge, Lecture notes in computer science, vol. 1174, Springer-Verlag, Berlin Heidelberg, New York, 1996
12. Bender W, Gruhl D, Morimoto N, Lu A (1996) Techniques for data hiding, IBM systems journal, vol. 35, no. 3k4, MIT Media Lab, pp 313–336
13. <http://isise.gmu.edu/~njohnson/steganography>
14. Ogihara T, Nakamura D, Yokoya N Data embedding into pictorial images with less distortion using discrete cosine transform
15. Ker A (2004) Improved detection of LSB steganography in grayscale images. Proceedings of 6th information hiding workshop. vol. 3200, Springer LNCS, pp 97–115
16. Fridrich J, Goljan M, Du R (2001) Detecting, LSB Steganography in color and gray –scale images, Magazine of IEEE multimedia, Special issue on security, Oct-Nov issue, 2001, pp 22–28
17. Habes A (2006) Information transmissions in computer network. Information hiding in bmp image implementation analysis and evaluation (Jan 2006)
18. Moon SK, Kawitkar RS, Data security using data hiding international conference on computational intelligence and multimedia applications 2007
19. Khan S, Haroon Yousaf M (2011) Variable least significant bits steganography. Accepted for publication in IJCSI vol. 8, Issue 5, Sept 2011
20. Khan S, Haroon Yousaf M (2011) Implementation of variable least significant bits steganography using decreasing distance decreasing bits algorithm. Accepted for publication in IJCSI, vol. 8, Issue 6, Nov 2011
21. <http://www.mathworks.com>
22. <http://www.math.ucla.edu/>

Chapter 44

A Graphic User Interface for H-Infinity Static Output Feedback Controller Design

J. Gadewadikar, K. Horvat and O. Kuljaca

Abstract In this work a MATLAB/SIMULINK based graphic user interface Design Scheme is described. H-Infinity Static Output Feedback controllers are very easy to implement once designed. The original design process to find the controller is a mathematically rigorous process; in this work we suggest a simpler process using a graphic user interface with a few simplifications introduced.

44.1 Introduction

The use of output feedback allows flexibility and simplicity of implementation. Moreover, in practical applications, full state measurements are not usually possible. The restricted-measurement static output-feedback (OPFB) problem is of extreme importance in practical controller design applications including flight control [1], manufacturing [2], and elsewhere where it is desired that the controller have certain pre-specified desirable structure, e.g., unity gain outer tracking loop and feedback only from certain available sensors. A survey of OPFB design results is presented [3]. Finally, though many theoretical conditions have been offered for the existence of OPFB, there are few good solution algorithms. Most existing

J. Gadewadikar (✉)

Systems Research Institute, Alcorn State University, Lorman, USA
e-mail: jyotirmay@gmail.com

K. Horvat · O. Kuljaca

Brodarski Research Institute, Zagreb, Croatia
e-mail: kruno@hrbi.hr

O. Kuljaca

e-mail: okuljaca@hrbi.hr

algorithms require the determination of an initial stabilizing gain, which can be extremely difficult.

It is well known that the OPFB optimal control solution can be prescribed in terms of three coupled matrix equations [4], namely two associated Riccati equations and a spectral radius coupling equation. A sequential numerical algorithm to solve these equations is presented [5]. However several problems are still open, most of the solution algorithms are hard to implement, are difficult to solve for higher order systems, may impose numerical problems and may have restricted solution procedures such as the initial stabilizing gain requirements. In this paper a MATLAB and simulink based graphic user interface is described to help users design an H-Infinity Static Output Feedback Controller using the tools presented in [6]. This paper documents a recent invited presentation at Mathwork's Robust Control System Group. The theme of this work is how to present an easy to use graphic user interface to let the control system designer use mathematically rigorous methods like H-Infinity. The paper is organized as follows, in Sect. 44.2 a brief introduction of the design task is given along with necessary and sufficient conditions of H-Infinity Static Output Feedback; Sect. 44.3 describes the GUI Interface design process steps. Section 44.4 explains the realization of the graphic user interface; Sect. 44.5 demonstrated GUI with a model, and finally a conclusion is given.

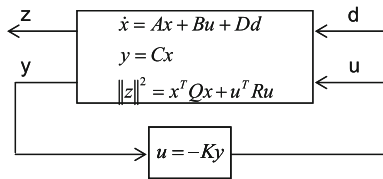
44.2 Design Task

The design task is to create a Simple prototype for a GUI-Based tool to help customers design multivariable H-Infinity Static Output Feedback controllers in a Simulink Environment. One of the goals of this exercise is that the graphic user interface must demonstrate user work flow. The overall objective is to create a tool to let users take advantage of H-Infinity Static Output Feedback Control Methodology; final requirement is to demonstrate the process through a simple problem.

44.2.1 System Definition and Design Objectives

A brief description of the Necessary and sufficient conditions is given in the next section which is followed by an explanation of how these can be converted into relevant design objective, constraints and additional requirements for a graphic user interface.

Fig. 44.1 System description



44.2.1.1 Necessary and Sufficient Conditions: H-Infinity Static Output Feedback

Consider the linear time-invariant system of Fig. 44.1 with control input $u(t)$ output $y(t)$, and disturbance $d(t)$ given by

$$\dot{x} = Ax + Bu + Dd, y = Cx, \quad (44.1)$$

and a performance output $z(t)$ that satisfies $\|z(t)\|^2 = x^T Q x + u^T R u$, $y = Cx$, for some positive matrices for some positive matrices $Q \geq 0$ and $R > 0$. It is assumed that C has full row rank, a standard assumption to avoid redundant measurements.

By definition the pair (A, B) is said to be stabilizable if there exists a real matrix K such that $A - BK$ is (asymptotically) stable. The pair (A, C) is said to be detectable if there exists a real matrix such that $A - LC$ is stable. System (44.1) is said to be output feedback stabilizable if there exists a real matrix K such that $A - BKC$ is stable. The System L_2 gain is said to be bounded or attenuated by γ if

$$\frac{\int_0^\infty \|z(t)\|^2 dt}{\int_0^\infty \|d(t)\|^2 dt} = \frac{\int_0^\infty (x^T Q x + u^T R u) dt}{\int_0^\infty (d^T d) dt} \leq \gamma^2 \quad (44.2)$$

(a) Bounded L_2 Gain Design Problem

Defining a constant output-feedback control as

$$u = -Ky = -KCx \quad (44.3)$$

It is desired to find a constant output-feedback gain K such that system is stable and the L_2 gain is bounded by a prescribed value γ .

Theorem 1 *Necessary and Sufficient Conditions for H-Infinity Static Output Feedback Control*

Assume that $Q \geq 0$ and (A, \sqrt{Q}) is detectable. Then system defined by Eq. (44.1) is output-feedback stabilizable with L_2 gain bounded by γ , if and only if

- i. (A, B) is stabilizable and (A, C) is detectable;
- ii. There exist matrices K^* and L such that

$$K^* C = R^{-1}(B^T P + L) \quad (44.4)$$

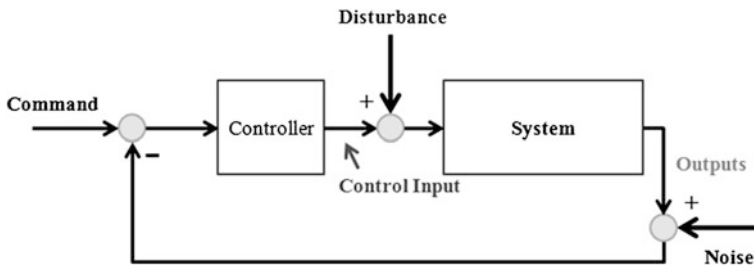


Fig. 44.2 System description

where $P > 0$, $P^T = P$, is a solution of

$$PA + A^T P + Q + \frac{1}{\gamma^2} P D D^T P - P B R^{-1} B^T P + L^T R^{-1} L = 0 \quad (44.5)$$

44.2.1.2 Gui Design Objective, Constraints and Additional Requirements

Objective of the graphic user interface is to present user a simple and more intuitive schematic to find the H-Infinity Static Output Feedback gain for a system modeled in simulink as shown in Fig. 44.2 such that output feedback controller is found using an iterative design algorithm inbuilt in the GUI. Here the objective is to Design Controller to reduce the disturbance effect with a constraint to use static gains only.

Additional Requirements in this exercise is to have a Relative control over how much Control Input to be used, how much Outputs to be made faster, and how much Disturbance to be rejected. Here the controller gain matrix K as described in Eq. 44.3 can be found using an algorithm which solves Eqs. 44.4 and 44.5 simultaneously. Further the solution depends on matrices Q and R , and the system L_2 gain. In this GUI design exercise matrices Q , and R will be assumed as identity matrices initially and the user can change these matrices using an interface. Increasing the weights of elements in the design matrix Q will make the states faster, and increasing the weights in the design matrix R will make the states slower with less control input being used. Parameter γ can be assumed as disturbance gain, lower it is lower is the effect of the disturbance on the outputs.

44.3 Graphic User Interface Design

This section describes the GUI Interface design process steps. Sections 44.3.1 and 44.3.2 describes the steps users must make, Sect. 44.3.3 describes GUI outputs and Sect. 44.3.4 described user controller selection and further testing.

44.3.1 User Workflow

In this section a user work flow is defined. The first step is to specify closed-loop I/Os in the model. Here it is assumed that a simulink model as shown in Fig. 44.2 is already available for the plant including a controller, and system block. In this step specific inputs and outputs are indicated using standard simulink procedure. An arrow pointing on a small circle downwards on a signal is representation of an input signal, and an arrow pointing upwards is a representation of the output as shown in Fig. 44.4. The next step is to specify the blocks or subsystems representing the controller. This stage is important as the model need to know the controller block. Later in this paper explained is the procedure how the GUI automatically replaces this controller after finding the H-Infinity controller using the user inputs. The next step is to specify the performance and stability goals for the inputs and outputs. This is to decide in weighing of the Q , and R matrices and the disturbance rejection gain γ . The last stage is to design the controller and observe the response of the controller and replace the controller block identified earlier with the identified one. All of the steps defines the user workflow process.

44.3.2 User Inputs

As described earlier the user has to provide the Q matrix, R matrix and the disturbance rejection gain γ . In the graphic user interface the Q and R matrices are initially selected as the identity matrices and a scaling parameter multiplying the respective matrices is used and given to user to design the static output feedback gain. Disturbance rejection gain γ can also be changed by the user. In order to make it more intuitive for the user the Q multiplication factor is called as system response slider; R multiplication factor is called as control input slider, and the γ as the disturbance gain slider. User can change these slider values using the mouse, lower system the response slider is slower is the system response. Lower control input slider means higher control efforts, better disturbance rejection and faster system response. For the Disturbance Gain Slider lower the value better is the disturbance rejection.

44.3.3 GUI Outputs

This section defined the GUI outputs, namely four outputs will be used. Three outputs will be the direct representation of the input sliders, these will be system response representative value, control input representative value, and disturbance gain representative value. This way user can see the values he/she is varying and can adjust the controller accordingly. There will be one more output in the

graphical form which will plot all the system outputs for the inputs and the disturbance values. For example of the system has two disturbance signals, two noise signals, two command inputs and two outputs. The graphical response will include the effect of all the inputs individually to all the outputs and there will be a twelve figure window for the user to effectively see the response at individual channel. User can look at these outputs and change the controller till he/she get the one best suitable for his/her needs.

44.3.4 User Controller Selection and Further Testing

For user controller selection and further testing there is a user controller selection pushbutton “Replace controller” to return satisfactory multivariable controller to the Simulink. User than can use Simulink Interface to see the response of the controller, user can also vary disturbance values and come back to GUI to design new controller for the changed disturbance values.

44.4 GUI Realization

As shown in the Fig. 44.3, clicking on the GUI button “Click to run GUI”, opens the Simulink Model, gets current linearization I/O points and linearizes the model while removing the contribution of the controller. The next step automated step is to obtain System Response, Actuator Output, and Disturbance Rejection Guidelines from user. Once the guidelines are obtained the GUI calculates H-Infinity Static Output Feedback Gain. After calculating the gain GUI combines the system with the controller, provide visual feedback to the user in terms of responses and finally, let user apply design to Simulink block diagram. Next section describes various functions and callbacks used to realize the GUI.

44.4.1 Functions and Callbacks

Various functions and callbacks are used towards realization of this graphic user interface. GUI opening function will open whenever “Click to run GUI” button in the simulink will be clicked, the GUI Opening Fun is to set the default controller and plot the output for the default controller. At this stage GUI interface will open where user can change various slider values. As explained earlier there will be three sliders, namely System Response, Control Input, and Disturbance Gain. Whenever the slider values are changes a function called `linear_sim` will be called. The `linear_sim` will act as an interface between the GUI and the simulink, it first will retrieve the plant removing the existing controller. To calculate the control

Fig. 44.3 Quanser 2 degrees of freedom system



gain only the plant model data is needed (namely matrices A , B , C , and D). The linear_sim than will retrieve all the current slider values and define Q , R , and γ values from the slider. Function linear_sim than will call the H-Inf Static Output Feedback algorithm and will return the controller to linear_sim, at the final stage the outputs using the returned controlled will be shown in the GUI output window. Once the user is satisfied with the design push button in the GUI window “Replace Controller” will replace the controller in simulink and plot the simulink outputs in GUI. Next section describes the H-Infinity Iterative algorithm used by the sub-routine H-Inf SOFB algorithm.

44.4.1.1 H-Infinity Static Output Feedback Algorithm

1. Initialize: Fix $\gamma \geq \gamma^*$. Set $n = 0$, $L_0 = 0$. Solve a standard (e.g., LQR) Riccati equation for given Q and R and obtain a stabilizing SVFB gain as initial gain $K_{s(0)}$. Define closed-loop matrix $\tilde{A}_0 = A - BK_{s(0)}$.
2. n th iteration: Solve ARE for P

$$P_n(\tilde{A}_n) + (\tilde{A}_n)^T P_n + Q + K_{s(n)}^T R K_{s(n)} + \frac{1}{\gamma^2} P_n D D^T P_n = 0, \quad (44.6)$$

update K , project onto nullspace perpendicular of C

$$K_{s(n+1)} = R K_{s(n)} - B^T P_n, \quad (44.7)$$

update L

$$L_{n+1} = R K_{s(n)} - B^T P_n, \quad (44.8)$$

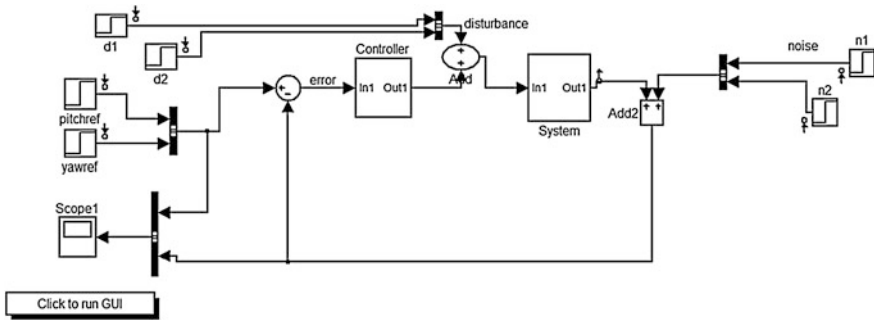


Fig. 44.4 Simulink system setup

update closed-loop system matrix

$$\tilde{A}_{n+1} = A - BK_{s(n+1)}, \quad (44.9)$$

3. Check convergence. If converged, go to step 4 otherwise set $n = n + 1$ go to step 2.
4. End. Set OPFB gain $K = K_s V_1 (S_0)^{-1} U^T$.

44.4.2 MATLAB/SIMULINK Inbuilt Functions Used

Various MATLAB/SIMULINK functions used in this exercise are briefly described, for details please refer to the Mathworks documentation for various tool boxes [7] for more details. The function “getlinio” is to get linearization I/O settings for Simulink model, “linlft” is to receive the plant without controller from Simulink, “lft” is used to combine the plant with the multivariable controller, “assignin” to define the controller parameters in workspace, “get_param” to receive Simulink step input values, and “set_param” to set Simulink parameters.

44.5 GUI Demonstrations

This section briefly demonstrates the design using an example model, a model available in the laboratory is used to show the simulations. In this exercise we have used a quanser two degree of freedom helicopter model, but the user can use any other plant using the schematic presented in the paper.

In this helicopter model there are two DC motors actuating two propellers affecting pitch and yaw angles, this system is a good candidate for showing H-Infinity Static Output Feedback approach using a graphic user interface.

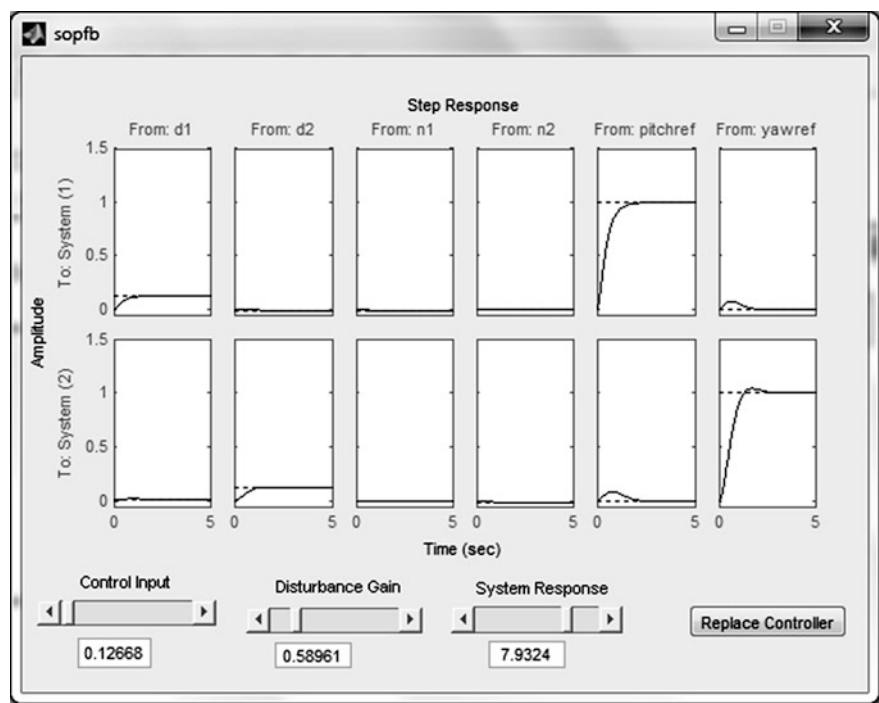


Fig. 44.5 Graphic user interface

In Fig. 44.4 one can observe the disturbance in outs getting on the control input channel, there are two set command inputs, pitchref and yawref for pitch and yaw respectively, there is some noise in the output channel defined by n1 and n2 blocks. “Click to run GUI” will open the default GUI and one can change the slider values and can see the system outputs as shown in Fig. 44.5.

44.6 Conclusion

In this paper a MATLAB/SIMULINK based graphic user interface design procedure is explained. The GUI can be easily modified to change the plant and controller values. The procedure allows user to directly change the design parameters and see the effect without going into mathematical details.

Acknowledgments The first author would like to acknowledge support of department of Homeland Security’s Scientific Leadership Award 2009-ST-062-000024.

References

1. Stevens BL, Lewis FL (2003) Aircraft control and simulation, 2nd edn. Wiley Interscience, New York
2. Kim YH, Lewis FL (1998) High-level feedback control with neural networks. World Scientific, Singapore, pp 55–75.
3. Syrmos VL, Abdallah C, Dorato P (1994) Static output feedback: a survey. In: Proceedings of 33rd IEEE conference on decision and control, Orlando, pp 837–842
4. Lewis FL, Syrmos VL (1995) Optimal control, 2nd edn. Wiley-Interscience, New York
5. Moerder DD, Calise AJ (1985) Convergence of a numerical algorithm for calculating optimal output feedback gains. IEEE Trans Autom Cont 30(9):900, 903
6. Gadewadikar J (2007) H-infinity output feedback: application to unmanned aerial vehicles. PhD dissertation, University of Texas at Arlington, May 2007
7. <http://www.mathworks.com>

Chapter 45

Active Contour Texture Segmentation in Modulus Wavelet Feature Spaces

Ashoka Jayawardena and Paul Kwan

Abstract In this paper we discuss a model that is able to segment textures using active contours. Our technique is based on active contour techniques using curve evolution. We build our model on properties of human vision, in that we segment the textures in a certain feature space. We will show the advantages of using modulus feature spaces. Wavelet coefficients are shown to exhibit local features both in space and frequency domains. We will implement our model in modulus wavelet subbands.

45.1 Introduction

The idea behind the active contour image segmentation is that a contour evolves subject to constraints imposed by the image such as image gradient. The active contour image segmentation algorithms can be implemented using classical snakes [1, 2] or level sets. In both these implementations active contours are energy minimizing curves and hence are formulated as energy minimization problems.

The curve evolution models are particular interest to us. When the curve (or the front) evolves in the normal direction of the curve we arrive at the following evolution scheme [3]:

$$\frac{\partial \phi}{\partial t} + F|\nabla \phi| = 0 \quad (45.1)$$

A. Jayawardena (✉) · P. Kwan
School of Science and Technology, University of New England,
Armidale, NSW 2351, Australia
e-mail: ashoka@turing.une.edu.au

P. Kwan
e-mail: kwan@turing.une.edu.au

The speed F is normally modelled by mean curvature thus resulting in mean curvature motion. The mean curvature evolution equation is given by [3, 4]

$$\frac{\partial \phi}{\partial t} + |\nabla \phi| \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) = 0 \quad (45.2)$$

The mean curvature motion has been extensively used to model geometric flow. In the level sets implementation the evolving curve is normally embedded in the zeroth level set. They have shown to be able to undergo automatic topologic changes.

Active contour segmentation algorithms have been developed where object mean can be used to discriminate textures [4]. However mean fails to discriminate many textures in the presence of high variances. Julesz [5] has proposed a statistical description of texture that is consistent with human visual perception which is now recognized as the Julesz conjecture.

A texture is defined as a homogeneous random field (RF) $u(x, y)$ on a finite lattice $(x, y) \in L \subset \mathbb{Z}^2$. The Julesz conjecture quoted in [6] states that: there exists a set of functions $f_k(u)$ such that samples drawn from any two RFs that are equal in expectation over this set are visually indistinguishable under some fixed comparison conditions. Mathematically,

$$\begin{aligned} \mathcal{E}(f_k(u)) &= \mathcal{E}(f_k(v)), \\ \forall k &\Rightarrow \text{samples of } u \text{ and } v \text{ are perceptually equivalent.} \end{aligned} \quad (45.3)$$

Thus as long as we find the right feature functions f we can use the expectation as the discriminatory variable. However, we use feature functions which are evaluated at each pixel location to enable active contour segmentation.

45.2 Energy Minimization Model

Lets define $u : D_I \rightarrow R$ be the image which we want segment into two partitions. Lets define the feature function $\mathbf{f} : D_I \rightarrow R^N$, where N is the number of feature dimensions, be the feature space where the feature function is evaluated at each image location, i.e. pixels.

We use the following external energy model:

$$\begin{aligned} E_1(C, \mathbf{c}_1) + E_2(C, \mathbf{c}_2) &= \int_{\text{inside}(C)} (\mathbf{f}(x, y) - \mathbf{c}_1)^T \mathbf{D} (\mathbf{f}(x, y) - \mathbf{c}_1) dx dy \\ &+ \int_{\text{outside}(C)} (\mathbf{f}(x, y) - \mathbf{c}_2)^T \mathbf{D} (\mathbf{f}(x, y) - \mathbf{c}_2) dx dy \end{aligned} \quad (45.4)$$

where \mathbf{c}_1 and \mathbf{c}_2 are constants and \mathbf{D} is a diagonal matrix with positive values.

It can be seen that when \mathbf{c}_1 is the mean of the feature space \mathbf{f} inside the contour C and \mathbf{c}_2 is the mean of the feature space \mathbf{f} outside the contour C , $E(C, \mathbf{c}_1, \mathbf{c}_2) = E_1(C, \mathbf{c}_1) + E_2(C, \mathbf{c}_2)$ achieves its minimum for the given contour C . To see this we calculate the partial derivatives of $E(C, \mathbf{c}_1, \mathbf{c}_2)$ with respect to \mathbf{c}_1 and \mathbf{c}_2 .

$$\frac{\partial(E(C, \mathbf{c}_1, \mathbf{c}_2))}{\partial \mathbf{c}_1} = -2\mathbf{D} \int_{\text{inside}(C)} (\mathbf{f}(x, y) - \mathbf{c}_1) dx dy$$

$$\frac{\partial(E(C, \mathbf{c}_1, \mathbf{c}_2))}{\partial \mathbf{c}_2} = -2\mathbf{D} \int_{\text{outside}(C)} (\mathbf{f}(x, y) - \mathbf{c}_1) dx dy$$

When \mathbf{c}_1 and \mathbf{c}_2 are the mean of inside and outside regions of the contour C , the above two partial derivatives vanish. If we further assume that for each feature dimension the feature value of a texture object is approximately constant, it is clear that when C is the contour separating the objects $E(C, \mathbf{c}_1, \mathbf{c}_2)$ achieves its global minimum. We have the following result.

Theorem 1 *Let $u : D_I \rightarrow R$ be an image function and $f : D_I \rightarrow R$ be a feature function of the image. Let the feature image f consists of two homogeneous random fields R_1 and R_2 . Let D_1, D_2 be a disjoint partition of D_I resulted from the two random fields such that $D_1 \cup D_2 = D$. Let R_1 has μ_1 mean and R_2 has μ_2 mean. Then if $\mu_1 \neq \mu_2$,*

$$E(C, c_1, c_2) = \int_{\text{inside}(C)} (f(x, y) - c_1)^2 dx dy$$

$$+ \int_{\text{outside}(C)} (f(x, y) - c_2)^2 dx dy$$

achieves its infimum at the object boundary.

Proof Let the D_I is arbitrarily partitioned into $D_A \cup D_B$ such that D_A and D_B are disjoint. Let R_A and R_B are the random fields corresponding of D_A and D_B . Let R_1 and R_2 has variances σ_1 and σ_2 and density functions $p_1(\mu_1, \sigma_1)$ and $p_2(\mu_2, \sigma_2)$ respectively. Since R_1 and R_2 are homogeneous random fields, it is clear that the densities of R_A and R_B are given by

$$p_A(\mu_A, \sigma_A) = k_A p_1(\mu_1, \sigma_1) + (1 - k_A) p_2(\mu_2, \sigma_2)$$

and

$$p_B(\mu_B, \sigma_B) = k_B p_1(\mu_1, \sigma_1) + (1 - k_B) p_2(\mu_2, \sigma_2)$$

respectively where k_A and k_B are given by

$$k_A = \frac{\int_{D_A \cap D_1} dx dy}{\int_{D_A} dx dy}$$

and

$$k_B = \frac{\int_{D_B \cap D_1} dx dy}{\int_{D_B} dx dy}.$$

Now it can be shown that

$$\begin{aligned}\mu_A &= k_A \mu_1 + (1 - k_A) \mu_2 \\ \mu_B &= k_B \mu_1 + (1 - k_B) \mu_2 \\ \sigma_A &= k_A \sigma_1 + (1 - k_A) \sigma_2 + k_A (1 - k_A) (\mu_1 - \mu_2)^2 \\ \sigma_B &= k_B \sigma_1 + (1 - k_B) \sigma_2 + k_B (1 - k_B) (\mu_1 - \mu_2)^2.\end{aligned}$$

Now the error E is given by

$$\begin{aligned}E &= \sigma_A \int_{D_A} dx dy + \sigma_B \int_{D_B} dx dy \\ &= \sigma_1 \int_{D_A \cap D_1} dx dy + \sigma_2 \int_{D_A \cap D_2} dx dy \\ &\quad + \frac{\int_{D_A \cap D_1} dx dy \int_{D_A \cap D_2} dx dy}{\int_{D_A \cap D_1} dx dy + \int_{D_A \cap D_2} dx dy} (\mu_1 - \mu_2)^2 \\ &\quad + \sigma_1 \int_{D_B \cap D_1} dx dy + \sigma_2 \int_{D_B \cap D_2} dx dy \\ &\quad + \frac{\int_{D_B \cap D_1} dx dy \int_{D_B \cap D_2} dx dy}{\int_{D_B \cap D_1} dx dy + \int_{D_B \cap D_2} dx dy} (\mu_1 - \mu_2)^2 \\ &= \sigma_1 \int_{D_1} dx dy + \sigma_2 \int_{D_2} dx dy \\ &\quad + \frac{\int_{D_A \cap D_1} dx dy \int_{D_A \cap D_2} dx dy}{\int_{D_A \cap D_1} dx dy + \int_{D_A \cap D_2} dx dy} (\mu_1 - \mu_2)^2 \\ &\quad + \frac{\int_{D_B \cap D_1} dx dy \int_{D_B \cap D_2} dx dy}{\int_{D_B \cap D_1} dx dy + \int_{D_B \cap D_2} dx dy} (\mu_1 - \mu_2)^2\end{aligned}$$

Thus it is clear that the error of any other partition is larger than the error of object partition $\sigma_1 \int_{D_1} dx dy + \sigma_2 \int_{D_2} dx dy$. \square

Let the step function H and the dirac impulse function δ are given by

$$H(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

and

$$\delta(t) = \frac{d}{dt}(H(t)) \text{ respectively.}$$

Then we can write the external energy Eq. 45.4 as follows:

$$\begin{aligned} E(C, \mathbf{c}_1, \mathbf{c}_2) = & \int_{D_I} (\mathbf{f}(x, y) - \mathbf{c}_1)^T \mathbf{D}(\mathbf{f}(x, y) - \mathbf{c}_1) H(\phi(x, y)) dx dy + \\ & \int_{D_I} (\mathbf{f}(x, y) - \mathbf{c}_2)^T \mathbf{D}(\mathbf{f}(x, y) - \mathbf{c}_2) (1 - H(\phi(x, y))) dx dy \end{aligned} \quad (45.6)$$

In order to perform the gradient decent of the energy equation, we can decompose the evolution equation of $\phi_t(x, y)$ as:

$$\frac{\partial \phi}{\partial t} = \frac{\partial \phi}{\partial t_{\text{external}}} + \frac{\partial \phi}{\partial t_{\text{internal}}}$$

Now external energy component of the evolution equation is given by

$$\begin{aligned} \frac{\partial \phi}{\partial t_{\text{external}}} & \stackrel{\text{def}}{=} \frac{d(E(C))}{d\phi} \\ & = \delta(\phi) \left(-(\mathbf{f}(x, y) - \mathbf{c}_1)^T \mathbf{D}(\mathbf{f}(x, y) - \mathbf{c}_1) \right. \\ & \quad \left. + (\mathbf{f}(x, y) - \mathbf{c}_2)^T \mathbf{D}(\mathbf{f}(x, y) - \mathbf{c}_2) \right) \end{aligned}$$

It is clear from the above equation that if the image domain consists only a single texture object then external energy contribution to the evolution equation is zero. Therefore, only locally discriminative features contribute to the evolution equation.

Using similar steps as in [4] we arrive at the following level set formulation of the curve evolution:

$$\begin{aligned} \frac{\partial \phi}{\partial t} = & \delta_\epsilon(\phi) \left[\mu \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) - v \right. \\ & - \lambda_1 (\mathbf{f}(x, y) - \mathbf{c}_1)^T \mathbf{D}(\mathbf{f}(x, y) - \mathbf{c}_1) \\ & \left. + \lambda_2 (\mathbf{f}(x, y) - \mathbf{c}_2)^T \mathbf{D}(\mathbf{f}(x, y) - \mathbf{c}_2) \right] \end{aligned}$$

45.3 Modulus Wavelet Feature Space

Since $\int \psi(x, y) dx dy = 0$, the $E(w)$, i.e. the expectation of a wavelet coefficient, is zero. However variance of the wavelet coefficients depends on the texture. Since coefficient mean is the discriminating criteria of our active contour segmentation model, we need to transform the wavelet coefficients to a certain feature space where variance energy is transferred into mean. The modulus feature space does exactly that. To quantify, to what extent variance energy gets transferred into the mean, let's assume that wavelet coefficients are zero mean gaussian process with variance σ . It can be shown that [7]

$$\begin{aligned} E(|x|) &= \sigma \sqrt{\frac{2}{\pi}} \\ E(|x|^2) &= \sigma^2 \end{aligned}$$

Therefore,

$$\begin{aligned} E((|x| - E(|x|))^2) &= E(|x|^2) - (E(|x|))^2 \\ &= \sigma^2 \left(\frac{\pi - 2}{\pi} \right) \end{aligned}$$

Thus the variance is reduced and transferred into mean.

45.3.1 Wavelet Subband Pre-Processing

Since our classification algorithm is based on the deviation from mean, each subband may respond to a particular object differently, i.e. may respond with a higher object mean than the global mean or lower object mean than the global mean. When using more than one feature space, the expectation of feature values for each texture objects is critical since even though each feature space may discriminate the texture objects when combined they may cancel out discriminatory features in terms of expectation and variance of feature values. This is indeed the case for horizontal and vertical modulus wavelet subbands since those subbands are orthogonal. When this occurs we need transform some subbands into negative images to have the same classification response for the same object. We apply the following transform:

$$-abs(f(x, y)) + k \tag{45.7}$$

where $abs(.)$ is the absolute value function and k is some constant. This pre-processing step improves the feature space correlation resulting in improved homogeneity of the feature space for a particular object.

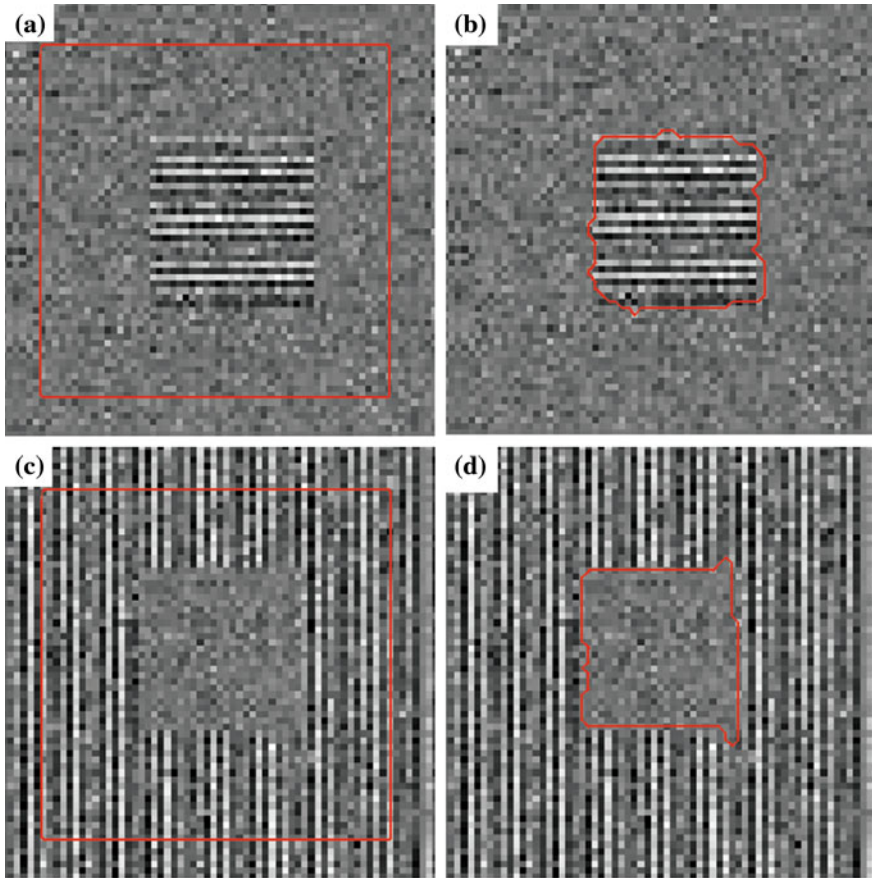


Fig. 45.1 **a** is the horizontal wavelet subband of the original texture image. **b** is the segmented image of the image **(a)**. **c** is the vertical wavelet subband of the original texture image. **d** is the segmented image of the image **(c)**

45.4 Results and Discussion

The active contour algorithm of Chan and Vese [4] fails to segment the texture images in wavelet domain since the mean of the wavelet coefficients zero.

We have only used horizontal and vertical subbands of the wavelet transform. The diagonal subband has been omitted since it contains mixture of both horizontal and vertical directional features. The low pass subband has been omitted since our choice of texture images has the same mean for both texture objects.

We have used $\max_{(x,y)}(\text{abs}(f(x,y)))$ for the value of k in 7. Separable wavelet subbands are capable of extracting horizontal and vertical features of the image. When texture objects can be discriminated using horizontal and vertical features individual subbands are sufficient for active contour segmentation as illustrated in

Fig. 45.2 The segmentation in the combined horizontal and vertical subbands

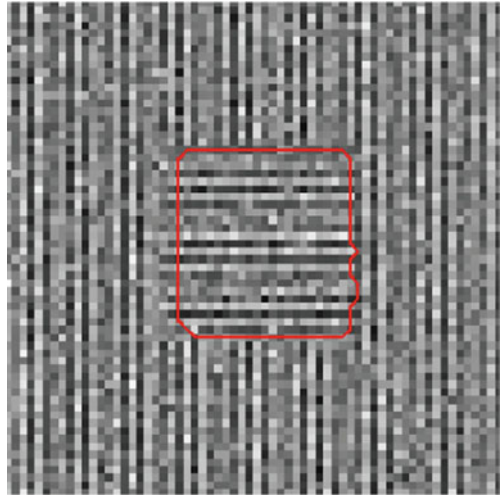


Fig. 45.1. The Fig. 45.2 illustrates the active contour segmentation with both horizontal and vertical subbands.

References

1. Kass M, Witkin A, Terzopoulos D (1988) Snakes: active contour models. *Int J Comput Vis* 1:321–331
2. Caselles V, Catte F, Coll T, Dibos F (1993) A geometric model for active contour models in image processing. *Numer Math* 66:1–31
3. Osher S, Sethian J (1988) Front propagating with curvature dependent speed: algorithms based on hamilton-jacobi formulation. *J Comput Phys* 79:12–49
4. Chan TF, Vese LA (2001) Active contours without edges. *IEEE Trans Image Process* 10: 266–277
5. Julesz B (1963) Visual pattern discrimination. *IRE Trans Inf Theory*, IT-8:84–92
6. Portilla J, Simoncelli EP (2000) A parametric texture model based on joint statistics of complex wavelet coefficients. *Int J Comput Vis* 40(1):49–71
7. Papoulis A (1991) Probability, random variables, and stochastic processes, 3rd edn. WCB/McGraw-Hill, New York

Chapter 46

A Framework for Verification of Fuzzy Rule Bases Representing Clinical Guidelines

M. Esposito and D. Maisto

Abstract The increase of expert knowledge is characterizing medical domain and determining a constantly growing and interacting number of relevant standardized specifications for care known as clinical guidelines. However, most clinical guidelines, especially when expressed in the form of condition-action recommendations, embody different kinds of structural errors that compromise their effectiveness. With this respect, this paper presents a framework to represent condition-action clinical recommendations as “IF-THEN” fuzzy rules and to verify the presence of some structural anomalies. In particular, we propose a method to detect redundancy, inconsistency and contradictoriness—a structural anomaly introduced in this paper for the first time—in a very simple and understandable way by using the concept of similarity between antecedents and consequents. Formalization in fuzzy degrees for these anomalies can be straightly interpretable as measurements suggesting how to suitably modify the clinical rules to eliminate or mitigate undesired effects. The framework has been assessed on a relevant sample set identified from the clinical literature with profitable results.

M. Esposito (✉) · D. Maisto
ICAR-CNR, Institute for High Performance Computing and Networking,
Via P. Castellino 111, 80131 Naples, Italy
e-mail: massimo.esposito@na.icar.cnr.it

D. Maisto
e-mail: domenico.maisto@na.icar.cnr.it

46.1 Introduction

The increase of expert knowledge is characterizing medical domain determining, as a consequence, an increase of the specialization phenomenon. Specialization, in its turn, produces a constantly growing number of clinical guidelines.

Clinical guidelines are standardized specifications for care, developed by a formal process incorporating the best scientific evidence of effectiveness with experts' opinions [9]. Their aim is promoting more consistent, effective, and efficient medical practices and improving healthcare [22].

These forms of specialized knowledge can be provided by clinical Decision Support Systems (DSSs), i.e., computer applications designed to support clinical decision making [19].

In general, DSSs built from clinical practice guidelines need a suitable representation—ontologies or condition-action rule bases—able to encode the medical knowledge into a logical formalism that can be effectively used by a reasoning engine to plan healthcare activities.

In particular, condition-action clinical rules, represented as “IF-THEN” rules, can encode elementary care recommendations, by specifying one or at most a few conditions linked to specific actions [18], permitting to cover most diagnostic and therapeutic guidelines.

Up to now, several decision support implementations have been widely studied [1, 11], by focusing on condition-action clinical recommendations. However, actually, one prerequisite for their broad acceptance and efficient application to medical settings is the guarantee of a high level of quality and reliability [5]. Moreover, as shown in numerous studies, most condition-action clinical recommendations embody different kinds of structural *anomalies* that compromise their practical value, such as inconsistency, i.e., some contradictory conclusions can be derived from two or more condition-action clinical rules, and redundancy, i.e., a useless and ineffective repetition of semantic information by two or more condition-action clinical rules [5, 18].

Anomalies can be thought as characteristics of the clinical guideline knowledge, independently of the knowledge coding formalism. However, each decision support implementation encodes clinical rules in a specific representation formalism, with its own syntax and semantics.

Further, verification and validation for correcting such anomalies are never definitive since guidelines are often subject both to evolution and changes.

Some studies, indeed, have shown that guideline medical knowledge generally becomes outdated in few years [17]. Therefore, a released guideline needs to undergo periodic reviews and revisions in order to maintain its connection with evidence-based practice [15]. Moreover, it is confirmed that a revision of generic setting-independent clinical guidelines, with the aim of adapting them to local clinical contexts, improves care providers' adherence to them [14]. The above should make clear that medical knowledge encoded in guidelines is never totally defined and bounded, also because clinical practice changes over time.

Furthermore, clinical guidelines do not need every clinical parameter is considered: just some parameters are useful to characterize a certain diagnosis or treatment. In some situations, it is interesting to compare two different guidelines with the intention of merging them, for example. This entails the necessity to assess the existence of structural anomalies between rules either having just some condition variables in common, or that do not share any condition variables, although concurrently working on the same action variable.

For this reason, a DSS modeling clinical guidelines should have, among its main characteristics, the ability of managing incomplete information, e.g., some missing variable value.

Starting from such considerations, this paper presents a framework for verifying the reliability of condition-action clinical recommendations encoded in the form of fuzzy rules, with the final aim of determining inconsistency and redundancy anomalies in a simple and understandable fashion.

Similarly, this paper introduces a novel anomaly, denoted as *contradictoriness*, which shows a potential inconsistency stemming from the incompleteness of the adopted rule base.

The proposed framework establishes a representation of clinical guidelines in a fuzzy-rule base and defines the measures for the aforementioned anomalies in terms of similarity between antecedents and consequents.

A key issue relies on the formalization of fuzzy degrees for these anomalies that can be simply interpreted by final users as measurements suggesting the modifications to be applied to the clinical rules in order to eliminate or mitigate the existing undesired effects. The method has been profitably assessed on a sample set of clinical rules identified from the relevant clinical literature.

The remainder of the paper is structured as follows. In [Sect. 46.2](#), both the motivations for adopting the fuzzy logic formalism to encode clinical rules and an overview of existing methods for verifying fuzzy rules are reported. In [Sect. 46.3](#), the proposed method for verifying clinical rules encoded by exploiting the fuzzy logic formalism is presented. A clinical scenario for validating the framework is described in [Sect. 46.4](#). A brief discussion concludes the paper in [Sect. 46.5](#).

46.2 Related Work

In the last few years, the Computer Science community has supported the transition to evidence-based medicine by creating the technologies necessary to make clinical knowledge more accessible, manageable and updatable [\[3\]](#).

In more detail, from an information management perspective, a condition-action clinical rule must be encoded into a decision support implementation. With this respect, the most critical activity is the selection of a mediating representation in a computable format [\[18\]](#).

Three specific issues affecting this activity are: (1) the inherent vagueness of the natural language of the clinical rules; (2) the need of managing incomplete

knowledge and successive guideline revisions to maintain its own evidence-based integrity; (3) the uncertainty arising because patients with highly similar clinical characteristics might receive very different recommendations from the guideline [7].

Additionally, measured and collected data lying on borderline cannot be strictly or clearly defined [7]. On the other hand, a physician may make mistakes or may misinterpret the indications of a guideline because the boundary between normal and abnormal status is not clearly set, or fail to carry out a complete test for diagnosis [21].

In this work, Fuzzy Logic has been identified as the most suitable approach to describe vagueness and imprecision, since it enables to explicitly represent natural vagueness rather than abolishing it, by means of a precise mathematical language. The most notable advantage in the adoption of fuzzy logic for encoding condition-action clinical rules is that they can be written in a form similar to natural language and provide recommendations consistent with the human thinking. A consequence of this transparency is that fuzzy rules are easy to develop and understand.

Representing condition-action clinical recommendations in the form of fuzzy rules implies the adoption of verification methods specifically devised and developed to work on fuzzy DSSs. Until now, different approaches have been adopted to study the problem of verifying a fuzzy rule base. They can be divided into global and local approaches.

Global (or dynamic) methods rely on the idea that any anomaly can be detected by involving the rule evaluation in the analysis [23, 24]. If an anomaly occurs within a fuzzy rule base, then this imposes the existence of some constraint on the results of the inference. Thus, the inference process is tightly involved in the verification.

Local (or static) methods try to detect anomalies in a fuzzy rule base by using similarity, affinity or matching measures [2, 10]. These methods examine subsets (e.g., a couple) of rules within the rule base and assess the degree of overlapping between the fuzzy sets associated to the propositions forming either conditions or conclusions.

The conceptual differences between the two verification approaches have as consequence the achievement of different results [20]. This is the major outcome we need to deal with if we compare local and global approaches to the verification. In general, the choice to use a specific approach depends on the particular application case analyzed [20].

For what concerns verification of condition-action clinical rules, global approaches do not result the most appropriate since the encoding of condition-action recommendations into fuzzy logic does not generate rule chaining. Rule chaining expresses the dependencies among rules, namely it expresses, more specifically, dependencies among the actions of a rule and the conditions of other rules. Condition-action clinical rules essentially involve 1-level fuzzy rules without chaining, since the action inferred by a rule is just a suggestion to report to clinicians and it cannot generate any feedback for activating other clinical rules.

As a result, local approaches result the most suitable to verify clinical condition-action rules. However, it seems approaches proposed in literature can hardly lead to a practical verification of fuzzy rule bases due to the lack of generality and the poor understandability of the suggested methods.

Further, to the best authors' knowledge, other verification methods need to work on rules involving the same variables. Practically, these methods do not envisage the possibility that in some rule there can be certain variables not expressed.

In such a direction, the proposed framework essentially performs a local verification based on a general definition of the inconsistency, redundancy and contradictoriness concepts for fuzzy clinical rules in terms of similarity between antecedents and consequents. A gradual hint is integrated in these definitions, in accordance with the imprecise character of fuzzy DSSs. The strength of this method relies on the formalization of fuzzy degrees of structural anomaly presence, which can be simply interpreted by final users as measurements oriented towards modifications of the clinical rules in order to eliminate or mitigate undesired effects potentially caused by contradictory or redundant knowledge.

46.3 A Framework for Fuzzy Rule Base Verification

The proposed framework has been developed to detect structural anomalies within a fuzzy Rule Base System (RBS) encoding condition-action clinical recommendations.

In the presented approach, structural anomalies are detected by means of a similarity measure defined for both the antecedents and consequents of rules stored into the RBS. Such measures, in their turn, derive from similarity degree between fuzzy sets representing the terms adopted in the rules.

Differently from classical systems, two or more sets can be said 'similar' in fuzzy logic if there is an overlapping between their membership functions. As a result, it is possible to define a fuzzy similarity measure that associates to every couple of fuzzy sets their degree of overlapping [16]. Among several definitions of fuzzy similarity measure [16], we have chosen to adopt a very common definition based on set-theoretic considerations. Given any two fuzzy sets A and B defined in a universe U , the fuzzy similarity measure used is defined by:

$$\sigma(A, B) = \frac{M(A \cap B)}{M(A \cup B)} \quad (46.1)$$

where $M(\cdot)$ associates to any fuzzy set the integral on U of its own membership function. σ , as it is possible to see by (46.1), varies in $[0, 1]$ and it is 0 or 1 when the fuzzy sets are not at all or completely overlapped, respectively.

Obviously, the set-theoretic similarity measure σ can be applied just to fuzzy sets. However, it is possible to extend the concept of similarity measure to both the

antecedents and the consequents of the rules of a fuzzy system. To this aim, several methods have been proposed and each of them needs the rules satisfy some specific requirements [8].

In this paper we provide, as a part of the local verification approach designed, a method to assess the similarity of fuzzy rules with antecedents in Conjunctive Normal Form (CNF).

In the CNF, the rule antecedents are formed by conjunctions of a set of propositions, each of them composed of a disjunction of a set of linguistic terms defined for an input variable. This representation has been chosen because of its high degree of compactness and knowledge synthesis.

In addition, rule antecedents can have an arbitrary subset of input variables expressed in describing the considered database. The introduction of this hypothesis tries to formalize the fact that not every variable present in a rule base may be present in each rule as well. Contrarily to what is typically proposed for RBSs [13], we manage this missing information by adhering to Open World Assumption (OWA) [4]. This choice depends of the fact that Closed World Assumption (CWA) assumes as false every proposition that cannot be proven true. As a result, under CWA any statement not expressed is considered false, while, under OWA it has a null value. Thus, in this framework, no information is inferred by variables with value not explicitly asserted.

Finally, this verification method is independent of the specific implementation used for the connectives and it can be applied to Zadeh, Lukasiewicz, Gödel and product logic.

Given the sets of linguistic variables $\{x_u\}$ and $\{y_v\}$, with $u = 1, \dots, n$ and $v = 1, \dots, m$ —representing the inputs and the actions of the RBS, respectively—, and their relative fuzzy sets A_u and B_v , let us consider a generic fuzzy rule R_i involving the subset of variables $\{x_u\}_i \subseteq \{x_u\}$ and the action $y_v \in \{y_v\}$:

$$R_i : \text{IFA}_i(\{x_u\}_i) \text{ THEN } B_i(y_v) \quad (46.2)$$

The condition $A_i(\{x_u\}_i)$ is written as:

$$A_i(\{x_u\}_i) \equiv \text{AND}_{x_u \in \{x_u\}_i} [A_i(x_u)]$$

with

$$A_i(x_u) \equiv \text{OR}_j [(x_u \text{ is } A_{iu}^j)]$$

where A_{iu}^j are terms of the linguistic variables x_u in the clauses $A_i(x_u)$.

On the other hand, the consequent is defined on a single action variable y_v to which is associated a term B_{iv} :

$$B_i(y_v) \equiv (y_v \text{ is } B_{iv})$$

The statement in (46.2) can also be expressed in a more compact way through the implication connective as follows:

$$R_i : A_i(\{x_u\}_i) \rightarrow B_i(y_v)$$

With the help of the similarity measure for fuzzy sets, it is possible to introduce the definition of similarity for rule antecedents (*SRA*) and similarity for rule consequents (*SRC*). Let us consider two fuzzy rules:

$$\begin{aligned} R_i &: A_i(\{x_u\}_i) \rightarrow B_i(y_v) \\ R_k &: A_k(\{x_u\}_k) \rightarrow B_k(y_v) \end{aligned}$$

then *SRA* and *SRC* of these rules are defined as follows:

$$SRA(i, k) = T\left(\left\{S\left\{\sigma(A_{iu}^j, A_{ku}^l)\right\}_{j,l}\right\}_u\right) \quad (46.3)$$

$$SRC(i, k) = \sigma(B_{iv}, B_{kv}) \quad (46.4)$$

where $T(\cdot)$ and $S(\cdot)$ are the T -norm and the T -conorm (or S -norm) implementing the logical connectives AND and OR. Practically, we propose to compute the *SRA* by firstly measuring, for each input variable, the similarity between the terms in the disjunctive clause and making, subsequently, their S -norm; secondly, by calculating the T -norm of the S -norm values achieved over the input variables. According to OWA, variables not expressed in both compared rules are not evaluated. Differently, to calculate *SRC*, we simply apply the (46.1) to the terms used in the consequents of the compared rules.

To better illustrate the proposed similarity measures we provide a simple instance. Let us consider the rules

$$R_1 : ((x_1 is A_{11}^1) \text{ OR } (x_1 is A_{11}^2)) \text{ AND } (x_2 is A_{12}^1) \rightarrow (y_1 is B_{11})$$

$$R_2 : (x_1 is A_{21}^1) \text{ AND } (x_2 is A_{22}^2) \text{ AND } (x_3 is A_{23}^1) \rightarrow (y_1 is B_{21})$$

and let us suppose that connectives are implemented according to Zadeh logic [25]. By applying the (46.3) and the (46.4), we have:

$$\begin{aligned} SRA(1, 2) &= \min\{ \max[\sigma(A_{11}^1, A_{21}^1), \sigma(A_{11}^2, A_{21}^1)], \sigma(A_{12}^1, A_{22}^2) \} \\ SRC(1, 2) &= \sigma(B_{11}, B_{21}) \end{aligned}$$

It worth making to note that the variable x_3 is present in rule R_2 but not in rule R_1 . As a consequence, the similarity measure $\sigma(\text{NULL}, A_{23})$ is not defined.

At this point, we can describe how to estimate the degree of some structural anomalies of two fuzzy condition-action clinical rules. Such a method allows to conciliate the satisfaction of the fuzzy formalism with the intuition and the common sense of human beings.

In general, fuzzy condition-action clinical rules are considered inconsistent if they have similar antecedents, but dissimilar consequents. It is worth noting that, differently from the classic logic, two rules can be also inconsistent if their premise parts and consequence parts are not necessarily the same and different ones,

respectively. On the other hand, two rules could contradict each other if their premises have little similarity but their consequences are pretty similar.

A definition of inconsistency that seems to satisfy the discussed requirements has been provided by [8]. Given two fuzzy rules R_i and R_k , their inconsistency is defined as:

$$Inc(R_i, R_k) = 1 - \exp \left\{ - \frac{\left(\frac{SRA(i,k)}{SRC(i,k)} - 1 \right)^2}{\left(\frac{1}{SRA(i,k)} \right)^2} \right\} \quad (46.5)$$

Analogously to the (46.5), we propose a new definition for the redundancy [6]. Two fuzzy condition-action clinical rules are considered redundant if both their antecedents and consequents are similar. However, this structural anomaly can occur in different degrees based on the relationships among the antecedents. Therefore, the degree of redundancy of two fuzzy rules R_i and R_k , can be computed by means of:

$$Red(R_i, R_k) = \exp \left\{ - \frac{\left(\frac{SRA(i,k) + SRC(i,k)}{2} - 1 \right)^2}{(SRA(i,k) \cdot SRC(i,k))^2} \right\} \quad (46.6)$$

From the aforementioned considerations, the concepts of consistency and redundancy are not concrete in fuzzy logic and they can only be described by a degree. Moreover, both (46.5) and (46.6) are defined if only if rule premises are comparable, i.e., if and only if the antecedents of the examined rules have some condition variable in common ($\{x_u\}_i \cap \{x_u\}_k \neq \emptyset$).

The third structural anomaly we introduce is a direct consequence of OWA premised in our framework. It can happen that two rules R_i and R_k have not comparable premises, i.e., with input variables $\{x_u\}_i$ and $\{x_u\}_k$ not matching at all ($\{x_u\}_i \cap \{x_u\}_k = \emptyset$), but consequences involving the same action variable y_v . In this case, while antecedents are not comparable by similarity, it is possible to evaluate just the similarity degree of their consequents. If the consequents are dissimilar then R_i and R_k could determine a contradiction. In fact, their premises can be contemporaneously verified and assert different values for the same action variable, under specific values assumed by the antecedent attributes.

According to these considerations, we propose calling this anomaly, describing a sort of inconsistency due to the incompleteness of the rule base, as *contradictoriness* and calculating its measure as:

$$Cont(R_i, R_k) = \exp \left\{ - \left(\frac{SRC(i,k)}{1 - SRC(i,k)} \right)^2 \right\} \quad (46.7)$$

defined if and only if ($\{x_u\}_i \cap \{x_u\}_k \neq \emptyset$).

Table 46.1 The set of clinical rules for classifying the COPD stage

| | |
|-----------------|---|
| I: Mild | $FEV_1/FVC < 70 \%$ $FEV_1 \text{ predicted} > 80 \%$ |
| II: Moderate | $FEV_1/FVC < 70 \%$ $50 \% < FEV_1 \text{ predicted} < 80 \%$ |
| III: Severe | $FEV_1/FVC < 70 \%$ $30 \% < FEV_1 \text{ predicted} < 50 \%$ |
| IV: Very Severe | $FEV_1/FVC < 70 \%$ $FEV_1 \text{ predicted} < 30 \% \text{ or } FEV_1 \text{ predicted} < 50\%$ plus CRF |

Definitions (5) and (6) and (7) have some fundamental properties characterizing them as measures.

Equations (46.5) and (46.6) are applicable when rule antecedents are comparable; they have value 1 if the rules have exactly both the same antecedents and consequents. When the rules have both the antecedents and consequents totally different, inconsistency and redundancy reaches its lowest value, i.e., 0.

If antecedents are totally different and consequents are exactly the same, then inconsistency and redundancy go to 0. In the opposite case, i.e., when antecedents are exactly the same but the consequents are totally different, redundancy is 0 but inconsistency has the value 1. In all the other cases, when both antecedents and consequents are similar in some degree, inconsistency and redundancy ranges in [0, 1.0].

Equation (46.7) is defined just when rule antecedents are not comparable; it assumes values in [0, 1.0] and it is equal to 0 when consequents are perfectly similar and is equal to 1 when they are completely dissimilar.

46.4 An Applicative Scenario: The Gold Guidelines

The framework above described has been validated by its application on a practical sample case. For this purpose, we have considered some clinical recommendations extracted from the GOLD guideline [12] for diagnosis, management and prevention of Chronic Obstructive Pulmonary Disease (COPD).

COPD is a preventable and treatable disease with some significant extra-pulmonary effects. In its severe forms, COPD leads to respiratory failure, hospitalization and eventually suffocation. Spirometry is the standardized and reproducible test adopted to establish the presence of airflow obstruction, indispensable in confirming the diagnosis of COPD.

Characteristically, parameters measured by spirometry are the forced expiratory volume in one second (FEV_1) and the forced vital capacity ratio (FEV_1/FVC).

Starting from the values of FEV_1 and FEV_1/FVC , and in addition to the presence of other chronic respiratory failures, the GOLD guideline has developed a set of clinical rules for classifying the COPD stage, as reported into Table 46.1.

These clinical rules have been modified according to a smooth fashion and encoded in terms of fuzzy rules. More in detail, three linguistic variables have been assigned to FEV_1 predicted, FEV_1/FVC ratio and Chronic Respiratory Failure (CRF), respectively, for modeling condition parts of the rules. Additionally, the linguistic variable *Stage* has been defined to model the action part of the clinical rules.

At this point, physicians in accordance with the knowledge formalized into Table 46.1, have identified some linguistic terms for which we have defined membership functions. For each considered linguistic variable, the used linguistic terms and their membership functions are reported in Fig. 46.1. A trapezoidal membership function has been used for the variables FEV_1 predicted, FEV_1/FVC ratio and *Stage*, whereas a singleton membership function has been applied to *CRF*, since it models a categorical concept.

By employing these linguistic variables and terms, we have written a set of eight "IF-THEN" rules aimed at identifying the appropriate COPD stage, as shown in the Table 46.2. This set of rules has been intentionally encoded wrongly, by including some inconsistent, redundant and contradictory rules.

Subsequently, the set of rules shown in Table 46.2 has been examined by means of proposed verification method assuming, to calculate the similarity measures, the Zadeh implementation for the connectives AND and OR.

By using the (46.5), the fifth rule is detected as inconsistent with the first, the second and the third ones with an inconsistency degree equals to 1.

By applying the definitions (46.6), it attains that the third rule is redundant with respect to the second one and the eighth one is redundant with respect to the seventh one, both with a redundancy degree equals to 1.

Finally, through the (46.7), we find out that the eighth rule has contradictoriness degree equals to 1 with the first and the second one, and equals to 0.999 with the fourth one.

As a result, by exploiting such indications provided in terms of inconsistency, redundancy and contradictoriness degrees, the fuzzy rule base has been updated by deleting from Table 46.2 the third, the fifth and the eighth rule.

Thus, the verification process has allowed to detect structural anomalies in the former rule base (Table 46.2) and, successively, to fix it in such a way that it is coherent with respect to the GOLD guidelines.

46.5 Conclusions and Future Works

In this paper, a framework to encode clinical guidelines in DSS fuzzy rule bases with its verification is presented.

In order to do that, guidelines are formalized in "IF-THEN" fuzzy rules with antecedents in CNF form involving an arbitrary number of linguistic variables.

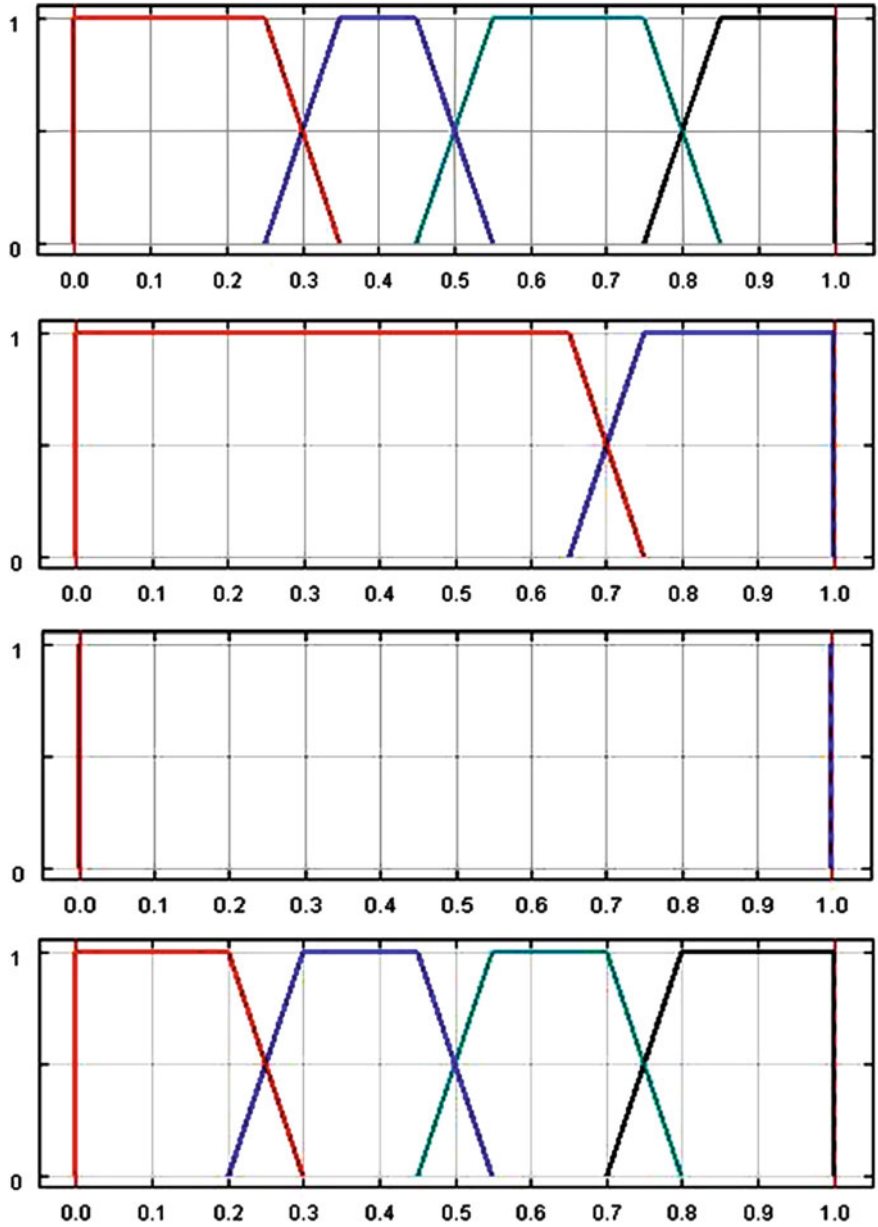


Fig. 46.1 From the *top* to the *bottom*: membership functions for FEV_1 predicted (Low (in red), Moderate (in blue), High (in cyan), Very High (in black)), FEV_1/FVC ratio (Low (in red), High (in blue)), CRF (Absent (in red), present (in blue)) and Stage (Mild (in red), Moderate (in blue), Severe (in cyan), Very Severe (in black))

Table 46.2 Redundant, inconsistent and contradictory “IF-Then” rule base aimed at identifying the appropriate COPD stage

| | |
|----|---|
| 1) | IF [<i>FEV₁/FVC</i> Ratio is Low AND <i>FEV₁</i> predicted is Very High] THEN [<i>Stage</i> is Mild] |
| 2) | IF [<i>FEV₁/FVC</i> Ratio is Low AND <i>FEV₁</i> predicted is High] THEN [<i>Stage</i> is Moderate] |
| 3) | IF [<i>FEV₁/FVC</i> Ratio is Low AND <i>FEV₁</i> predicted is High AND <i>CRF</i> is Absent] THEN [<i>Stage</i> is Moderate] |
| 4) | IF [<i>FEV₁/FVC</i> Ratio is Low AND <i>FEV₁</i> predicted is Moderate] THEN [<i>Stage</i> is Severe] |
| 5) | IF [<i>FEV₁/FVC</i> Ratio is Low AND <i>FEV₁</i> predicted is High AND <i>CRF</i> is Absent] THEN [<i>Stage</i> is Very Severe] |
| 6) | IF [<i>FEV₁/FVC</i> Ratio is Low AND <i>FEV₁</i> predicted is Low AND <i>CRF</i> is Absent] THEN [<i>Stage</i> is Very Severe] |
| 7) | IF [<i>FEV₁/FVC</i> Ratio is Low AND <i>FEV₁</i> predicted is (Low OR Moderate) AND <i>CRF</i> is Present] THEN [<i>Stage</i> is Very Severe] |
| 8) | IF [<i>CRF</i> is Present] THEN [<i>Stage</i> is Very Severe] |

Subsequently, a verification method to detect redundancy, inconsistency and contradictoriness—a novel anomaly presented in this work for the first time—has been introduced.

Redundancy, inconsistency and contradictoriness have been formulated in terms of degrees by taking in account the definition of similarity between antecedents and between consequents for couples of rules.

The method has been profitably assessed on a sample set of condition-action clinical recommendations known as GOLD guidelines and identified from the relevant clinical literature.

The results demonstrate how the verification method proposed is able to suggest rule base modifications in order to eliminate or mitigate undesired errors, eventually caused either by an incorrect formulization of the clinical guideline or by their successive revision/extension.

In the future, we are planning to test the application of this framework on larger guidelines and we are going to study its computational characterization (cost, scalability etc.).

References

1. Brokel J, Shaw M, Nicholson C Expert clinical rules automate steps in delivering evidence-based care in the electronic health record
2. Cho S, Lehto M (1992) A fuzzy scheme with application to expert classification systems for computing the degree of match between a rule and an assertion. *Cybern Syst* 23(1):1–27
3. Coiera E (1996) Artificial intelligence in medicine: the challenges ahead. *J Am Med Inf Assoc* 3(6):363
4. Drummond N, Shearer R (2006) The open world assumption. <http://www.csman.ac.uk/drummond/presentations/OWA.pdf>
5. Duftschmid G, Miksch S (2001) Knowledge-based verification of clinical guidelines by detection of anomalies. *Artif Intell Med* 22(1):23–41

6. Esposito M, Maisto D (2011) A verification method for clinical guidelines represented as fuzzy rules. In: International fuzzy system symposium, Ankara, Turkey. Accepted 17–18 Sept 2011
7. Jaulent M, Joyaux C, Colombet I, Gillois P, Degoulet P, Chatellier G (2001) Modeling uncertainty in computerized guidelines using fuzzy logic. In: Proceedings of the AMIA symposium, American Medical Informatics Association, p 284
8. Jin Y, Von Seelen W, Sendhoff B (1999) On generating FC3 fuzzy rule systems from data using evolution strategies. *IEEE Trans Syst Man Cybern Part B: Cybern* 29(6):829–845
9. Leape L (1990) Practice guidelines and standards: an overview. *Qual Rev Bull* 16(2):42
10. Liao T, Zhang, Z (1996) A review of similarity measures for fuzzy systems. In: Proceedings of the fifth IEEE international conference on fuzzy systems, vol 2. IEEE, pp 930–935
11. Niemi K, Geary S, Quinn B, Larrabee M, Brown K (2009) Implementation and evaluation of electronic clinical decision support for compliance with pneumonia and heart failure quality indicators. *Am J Health-Syst Pharm* 66(4):389
12. Rabe K, Hurd S, Anzueto A, Barnes P, Buist S, Calverley P, Fukuchi Y, Jenkins C, Rodriguez-Roisin R, van Weel C et al (2007) Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: gold executive summary. *Am J Respir Crit Care Med* 176(6):532
13. Reiter R (1987) On closed world data base. In: Readings in nonmonotonic reasoning, Morgan Kaufmann Publishers Inc., San Francisco
14. Saltman DC (1998) Guidelines for every person. *J Eval Clin Pract* 4(1):1–9
15. Scott-Wright A (2004) Managing revisions of rules and guidelines used in clinical information systems: exploring a hierarchical knowledge representation model, Massachusetts Institute of Technology, Massachusetts
16. Setnes M, Babuska R, Kaymak U, van Nauta Lemke H (1998) Similarity measures in fuzzy rule base simplification. *IEEE Trans Syst Man Cybern Part B: Cyberne* 28(3):376–386
17. Shekelle PG, Ortiz E, Rhodes S, Morton SC, Eccles MF, Grimshaw JM, Woolf SH (2001) Validity of the agency for healthcare and quality clinical practice guidelines: how quickly do guidelines become outdated? *J Am Med Assoc* 286(12):1461–1467
18. Shiffman R (1997) Representation of clinical practice guidelines in conventional and augmented decision tables. *J Am Med Inf Assoc* 4(5):382
19. Shortliffe E, Cimino J (2006) Biomedical informatics: computer applications in health care and biomedicine. Springer, New York
20. Viaene S, Wets G, Vanthienen J (2000) A synthesis of fuzzy rule-based system verification. *Fuzzy Sets Syst* 113(2):253–265
21. Woolcock A, Read J (1966) Lung volumes in exacerbations of asthma. *Am J Med* 41(2):259–273
22. Woolf S (1992) Practice guidelines, a new reality in medicine: II. methods of developing guidelines. *Arch Intern Med* 152(5):946
23. Yager R, Larsen H (1991) On discovering potential inconsistencies in validating uncertain knowledge bases by reflecting on the input. *IEEE Trans Syst Man Cybern* 21(4):790–801
24. Yang S, Tsai J, Chen C (2003) Fuzzy rule base systems verification using high-level petri nets. *IEEE Trans Knowl Data Eng* 15:457–473
25. Zadeh L (1965) Fuzzy sets. *Inf Control* 8(3):338–353

Chapter 47

Communication Impact on Project Oriented Teaching in Technology Supported Education

Martin Misut and Katarina Pribilova

Abstract Article describes results of research conducted at Faculty of Education of Trnava University in Trnava. One of the stated goals was to find out what is the impact of communication among students and teachers, as well, on the effectiveness of project oriented education. Experiments were made with students during the computer science courses.

47.1 Introduction

With the arrival of new technologies the possibilities of communication in education will also be possibly changed. Communication is essential for teamwork within projects and can significantly affect learning processes and learning outcomes as well. New technologies open up additional possibilities for the learning communities' creation. But that requires acquisition of new communication skills for the teachers and students.

Today's working with technology takes place through range of media, from interactive websites to virtual meeting environments. NCREL and The Metiri Group [1] included technological literacy and effective communication in academic standards in one of their reports. Research conducted at the Massachusetts

Described research as well as the publication of this paper was supported by VEGA grant no. 1/0247/10.

M. Misut (✉) · K. Pribilova
Faculty of Education, Trnava University, Priemyselna 4,
PO Box 9 918 43, Trnava, Slovakia, Rome
e-mail: domenico.maisto@na.icar.cnr.it

Institute of Technology (MIT) also indicates that adaptation onto changes in technology requires not only knowledge how to use technology tools, but the ability to use these tools [2].

Macromedia and Certiport [3] carried out research studies in North America, Great Britain and Australia. They surveyed people's views on the requirements, goals and trends in digital communications, as well as the effective communication using digital formats such as video, audio, animation and pictures. Initial findings indicate following three results in the use of digital communications:

- Trends: people use the new forms of digital communications to become more productive, their communication to be more persuasive and effective.
- Job success: individuals, organizations and industry, all find value in digital communication technology.
- Certification: Certification is of great interest for digital communication skills

Communication plays significant role in education. Intensive research has been carried out in the field of communication in education mainly in connection with new technology opportunities in last decade e.g. [4–8].

Two different modes are known for Computer Mediated Communication (CMC)—asynchronous and synchronous. Asynchronous CMC occurs in different time and does not need parallel participation of disputants; synchronous CMC occurs in real time and disputants must be present. Asynchronous CMC, broad used in e-learning, deepen discussion, enables communication of students in different time, keeps ongoing discussion, and all students can contribute to discussion. Soles a Moller [9] referred that synchronous CMC is more suitable for extroverts, while asynchronous CMC for introverts.

Burnett [10] noted that synchronous online chat has been long ignored as a medium for productive group discussion between distance learning students and their tutors. However, synchronous CMC is increasingly recommended as an appropriate discussion format in higher education [11]. Shotsberger [12] evaluated professionally developed program aimed at helping teachers to implement mathematics professional standards. The program included weekly web-board synchronous chats, which required the participants read materials, brainstormed ideas for learning and included the realization of standards. All participants described the possibility of interaction with the teacher to obtain new information on the implementation of the program as very effective. Ohlund, Yu, Jannssch-Pennell a Digangi [13] presented the results of research in which respondents were divided into four groups according to the use of communication medium: (1) email, (2) chat, (3) email and chat, and (4) other communication tool. Respondents who use synchronous and asynchronous forms of online discussions were largely able to fulfil the required course activities. Research has shown that the combined synchronous and asynchronous online discussing maximize personal engagement in learning. Yout and Shapiro [14] reported a case study in which students higher ranked asynchronous communication than synchronous communication. G. M. Johnson [15] has done research which compared two WebCT communication tools, synchronous chat and asynchronous discussion. The

research found that about 40 % of respondents indicate that they prefer synchronous chat and 60 % of respondents preferred asynchronous discussion. Experienced users decided more often to use chat instead of asynchronous discussion forum. Research also showed that about 43 % of respondents identified that they learned best when used synchronous chat and about 57 % of respondents identified the asynchronous discussion. Although these studies confirm that digital communication has become one of the basic skills needed for career advancement, it involves more than just achieving computer literacy. Success in school and at work depends on knowing how to choose the right medium for the message, and then design and create various forms of communication.

Since some general knowledge exists about student preferences and impact of communication onto education process we would like to know communication preferences of our students as well as the impact of student's communication practices on project teamwork in learning environment at Trnava University.

47.2 The Research Realization and Organization

Within the complex research supported by VEGA grant we researched also how students use individual electronic communication mediums in the education process realized in the form of project-oriented education. The research has been realised in fall term 2009–2010 at the Faculty of Education at Trnava University in Trnava. Students were split into two clusters: face-to-face and blended learning cluster. The model, recommended by Turek, was used for planning and realisation of the project oriented education [16].

The research goal was set as follow: to find out how form and means of communication impact the project oriented education effectiveness within the computer science courses. Following hypotheses were tested:

- H1: Students prefer electronic communication medium no regard of the education form.
- H2: There is statistically significant difference between project assessments of students working with following mediums: 1. Group—chat; 2. Group—e-mail; 3. Group—discussion forum; 4. Group—face-to-face discussion.
- H3: Duration of the communication medium usage has influence on the project assessments.

The research sample consisted of 171 students in the third year of undergraduate full time study in the fields of Teaching. Students were classified into four working groups according to the preferred communication medium. The first group used as the preferred medium synchronous chat, the second group used the asynchronous e-mail, third group asynchronous discussion forum, and the last group preferably used synchronous communication face-to-face.

Each working group formed team of three students. Students were grouped according to their study achievements. At the initial seminar, students were

Table 47.1 The proportion between the usages of communication medium

| | Face-to-face | Electronic means |
|------------------|--------------|------------------|
| Percentage ratio | 29.5 | 70.5 |

informed in detail about the implementation of the educational process throughout the semester. Each group had to solve the project, which consisted of the following tasks:

1. To prepare the relevant teaching plans for the selected subject.
2. To propose activities, using ICT, for different phases of the learning process within the subject.
3. To elaborate a short description of any means of ICT, its benefits, and how students use it in one lesson.
4. To prepare an interactive presentation about the project and present it.

Student projects were evaluated using the following criteria: originality, quality of proposed activities, using ICT, an interactive presentation (readability, clarity, interactivity, structure, presentation, visual aids use).

Students that studied face-to-face, attended seminars regularly (every week according to the schedule) – throughout the term and worked on their projects in classroom as well as outside. This was the reason, why they needed to keep a communication log, where they recorded each personal meeting, and other communication acts made during the work on projects.

Three other groups attended face-to-face seminars only three times during the semester. During these workshops, each group consulted finished work with the teacher. The objective of personal meeting was to check the results obtained for any given project. The project result presentations were made at the last face-to-face seminar. Students also attended five prescheduled two-hour meetings depending on the preferred communication medium (chat, e-mail, and discussion forum). These students needed to keep a communication log, where they record the time of communication, communication subject, communication participants, and communication medium, as well.

Research has proved some of our expectations through verifying validity of stated hypotheses.

47.3 Interpretation of Research Results

In Hypothesis H1, we assumed that students prefer electronic communication medium no regard of the education form. Table 47.1 shows the distribution of student preferences in whole student set. The preferences of the students groups are shown in detail in Table 47.2.

Table 47.2 The correlation between the usages of communication medium depending on the form of education

| Education form | Group | Face-to-face | Electronic means |
|------------------|---------|--------------|------------------|
| Blended learning | 1.group | 29.04 | 70.96 |
| | 2.group | 26.69 | 73.31 |
| | 3.group | 27.46 | 72.54 |
| Face-to-face | 4.group | 38.27 | 61.73 |

Table 47.3 Percentage share of communication means by groups

| Groups | Face-to-face | Chat | Phone | E-mail | Discuss forum |
|----------------------|--------------|------|-------|--------|---------------|
| % Ratio of all means | 29.5 | 27.8 | 1.6 | 31.1 | 10 |

We found that electronic communications significantly outweigh a face-to-face communication and even students who were enrolled for face-to-face form of education have a higher percentage of use of electronic communication means (61.73 %) compared with a face-to-face communication (38.27 %). Based on these results we proved the validity of the hypothesis H1. The electronic communication means, which are largely based on impersonal and yet interactive communications are becoming increasingly popular among students and provide them with new opportunities in education. Whether the independence of electronic communication, where teachers and learners are not necessarily in the same place and at the same time during communication or gain greater opportunities to obtain information necessary for the development tasks within the project.

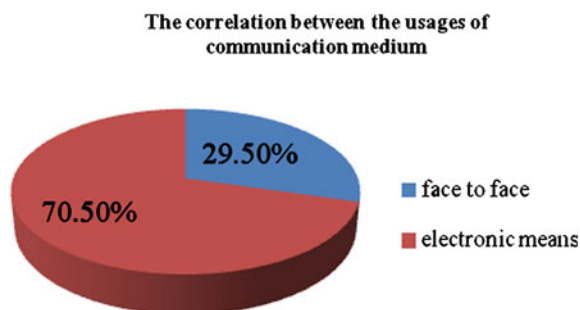
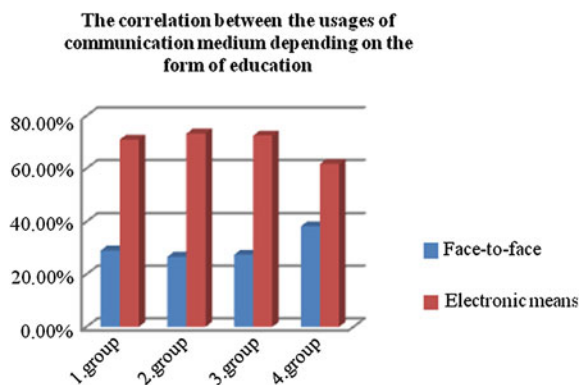
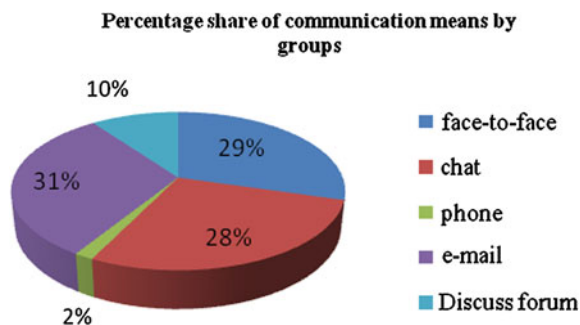
Facts, that electronic communication is currently developing, are confirmed by further results (Table 47.3) (Figs. 47.1, 47.2, 47.3).

The e-mail, asynchronous communication mean, was most often used by students in projects (31.1 % of all communication means). The second most common used communication mean was a face-to-face discussion represented with 29.5 % portion followed by synchronous chat with 27.8 %. One of possible explanation is that students have more time to write messages with asynchronous means because the sender does not wait for an immediate response. This increases the ability to process information and may have a positive impact on the quality of the information.

Hypothesis H2 stated that there is statistically significant difference between project assessments of students working in groups using different communication means as preferred selection: 1. group—chat; 2. group—e-mail; 3. group—discussion forum; 4. group—face-to-face discussion. Research was oriented onto the influence of preferred means of communication, type of communication medium and length of communication on the results. The goal was to find out whether the results of each group are comparable and which one of communication means leads to improved results for project results. The method of analysis of variance ANOVA $F(3.64) = 5.263$, $p = 0.0026$ confirmed that between the project results

Table 47.4 Influence of communication means on the results

| Groups | Project results Average | Project results SE |
|----------------------------------|----------------------------|-----------------------|
| 1. group—chat | 73.64188 | 1.316393 |
| 2. group—e-mail | 68.71679 | 1.723564 |
| 3. group—discussion forum | 77.50272 | 1.520040 |
| 4. group—face-to-face discussion | 71.43025 | 1.861662 |

Fig. 47.1 The proportions between the usages of communication medium**Fig. 47.2** The correlation between the usages of communication medium depending on the form of education**Fig. 47.3** Percentage share of communication means by groups

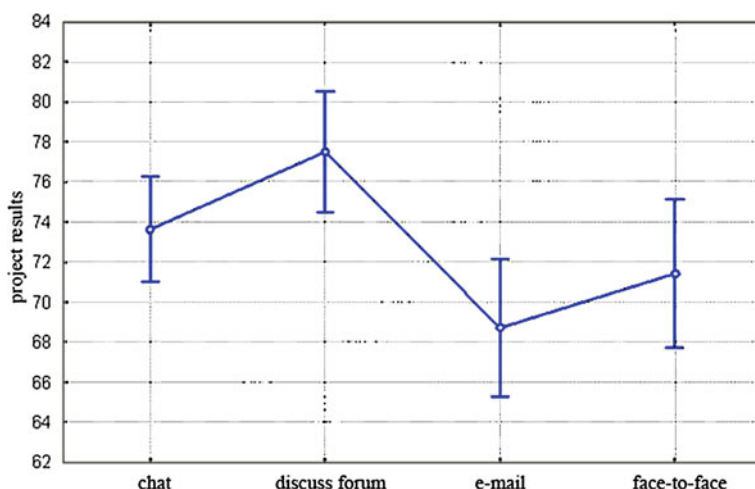


Fig. 47.4 Influence of communication means on the results

Table 47.5 Tukey post hoc test—comparison of peer groups

| Groups | Chat (<i>p</i>) | E-mail (<i>p</i>) | Discussion forum (<i>p</i>) | Face-to-face discussion |
|-------------------------|-------------------|---------------------|-------------------------------|-------------------------|
| Chat | — | 0.115797 | 0.230079 | 0.766985 |
| E-mail | 0.115797 | — | 0.001782 | 0.709253 |
| Discussion forum | 0.230079 | 0.001782 | — | 0.065365 |
| Face-to-face discussion | 0.766985 | 0.709253 | 0.065365 | — |

are a statistically significant difference, so hypothesis H2 is valid (Table 47.4; Fig. 47.4).

Students achieved the best results with usage of discussion forum as preferred communication mean. Students preferably used e-mail achieved the worst results. Comparing results of different groups among themselves, statistically significant difference exists only between the second (e-mail) and the third group (discussion forum) ($p = 0.0017$) (Table 47.5).

Results of described research showed, that inclusion of a discussion forum as a means of communication in the project oriented education can improve it taking into account the limits bordered the research.

Hypothesis H3 stated that the length of interpersonal communication using communication means among the members of project team has an impact on the results of the project. Multiple linear regression ($R^2 = 0.15$; $F(4.64) = 2.734$; $p = 0.036$) showed that only the length of a face-to-face communication had statistically significant ($p = 0.017$) effect on the results of the project (Table 47.6).

It is possible to conclude that the longer lasted face-to-face communication in solving students' project, the better were results of the projects (Fig. 47.5). However, the validity of hypothesis H3 is not confirmed because other variables

Table 47.6 Results of multiple linear regressions with the results of the evaluation project as a dependent variable

| | Beta | SE beta | B | SE B | <i>t</i> (63) | <i>p</i> |
|------------------------|-----------|----------|---------|---------|---------------|----------|
| Group | −0.294247 | 0.132528 | −1.8592 | 0.83738 | −2.22027 | 0.030006 |
| face-to-face (in min) | 0.315913 | 0.128568 | 0.0112 | 0.00454 | 2.45717 | 0.016769 |
| Discuss forum (in min) | 0.134761 | 0.116962 | 0.1663 | 0.14434 | 1.15217 | 0.253604 |
| Chat (in min) | 0.130709 | 0.126681 | 0.0033 | 0.00317 | 1.03179 | 0.306114 |

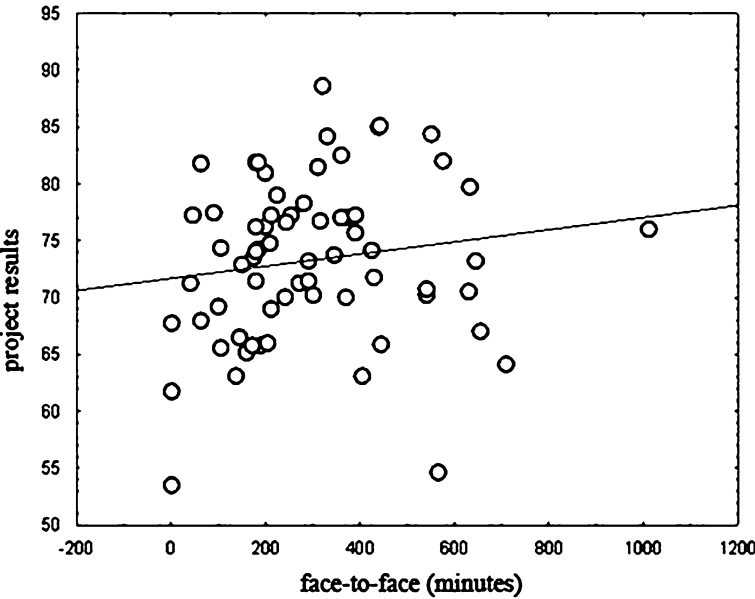


Fig. 47.5 Multiple linear regression

(time spent on the discussion forum and chat), no significant affected the evaluation of the project.

47.4 Conclusions

The research results showed that preferred communication medium and communication duration have an impact on the project results. It also showed that the discussion forum is a suitable communication medium for the project teamwork. Based on these findings, it is suitable to include this communication medium in the project oriented education.

Electronic communication expands everyday in the whole society, mainly in industry and business. Education cannot ignore this trend and moreover, needs to

prepare students for life. The success of graduates in the job market is dependent on their knowledge and skills. Research described in this article has ambition to develop educational process by improved understanding of the electronic communication use in project oriented education.

References

1. North Central Regional Educational Laboratory and the Metiri Group (2004) enGauge 21st century skills: literacy in the digital age. North Central Regional Educational Laboratory (NCREL), Oak Brook
2. Mitchel RESNICK, Natalie RUSK, and Stina COOKE (2003) The computer clubhouse: technological fluency in the inner city. ArchNet Digital Library, Blacksburg
3. Kirsti AHO (2005) Digital communication research: summary of results. Macromedia, Inc.
4. Lim HL (2009) E-communication patterns in collaborative learning networks. In: Proceedings of the 8th European conference on E-learning, 2009, pp 332–338, ISBN: 978-1-906638-52-8
5. Greener S (2009) Talking online: reflecting on online communication tools. *Campus-Wide Information Systems*, 26(3), 178–190, ISSN: 1065-0741
6. Mathiasen H, Schrum L (2008) Web 2.0 and social software: challenges and complexity of communication in education. In: Holzinger A (ed) HCI and usability for education and work. Springer, Berlin, pp 97–112, ISBN: 978-3-540-89349-3
7. Zounek J (2009) E-learning—jedna z podob učení v moderní společnosti. Masaryková Univerzita, Brno. ISBN 978-80-210-5123-2
8. Limniou M, Smith M (2010) Teachers' and students' perspectives on teaching and learning through virtual learning environments. *Eur J Eng Edu*, 35(6):645–653
9. Soles C, Moller L (2002) Myers Briggs type preferences in distance learning education. *IJET*, 2 <http://smi.curtin.edu.au/ijet/v2n2/soles/index.html>
10. Burnett C (2003) Learning to chat: tutor participation in synchronous online chat. *Teach High Educ* 8:247–261
11. Chen Y, Chen Nian-Shing, Tsai Chin-Chung (2009) The use of online synchronous discussion for web-based professional development for teachers. *J Comput Educ* 53(4):1155–1166
12. Shotsberger PG (2000) The human touch: synchronous communication in web-based learning. *Educ Tech*, 53–55
13. Ohlund B, Yu CH, Jannssch-Pennell A, Digangi SA (2000) Impact of asynchronous and synchronous internet-based communication on collaboration and performance among K-12 teachers. *J Educ Comput Res* 23:405–420
14. Shapira P, Youtie J (2001) Teaching with internet and multimedia technologies: insights from an online seminar on industrial modernization. *J Plan Educ Res* 21:71–83
15. Johnson GM (2006) College student psycho-educational functioning and satisfaction with online study groups. *Educ Psychol*, 26(5):677–688
16. Turek I (2008) *Didaktika 2008*. Iura Edition, spol. s. r. o., Bratislava ISBN: 978-80-8078-198-9

Chapter 48

A Failure Modes and Effects Analysis of Mobile Health Monitoring Systems

Marcello Cinque, Antonio Coronato and Alessandro Testa

Abstract Many solutions are emerging for the remote and continuous monitoring of unpredictable health problems, such as cardiac diseases. These are designed to be minimally invasive for health monitoring and based on smart and mobile technologies conformable to the human body, helping to improve considerably the autonomy and the quality of life of patients. Clearly, the correct functioning of these systems is very critical for the safety of patients, hence their practical application calls for stringent dependability requirements which need to be assessed against potential failure modes since the inception of the system, in its design phase. Despite the criticality of the problem, there is still little knowledge about the typical failures that may affect the correct functioning of these systems. Without such knowledge, it becomes difficult to devise effective countermeasures to failure events. To fill this gap, this paper proposes a *Failure Mode and Effect Analysis (FMEA)* for a typical mobile health monitoring system. Based on past results and extensive studies, the analysis allowed to identify the main failures, their consequences, and possible causes, affecting the functional components of modern health monitoring systems.

M. Cinque (✉) · A. Testa
Dipartimento di Informatica e Sistemistica, Università di Napoli Federico II,
Via Claudio 21, 80125 Napoli, Italy
e-mail: macinque@unina.it, testa@unina.it

A. Testa
e-mail: alessandro.testa@na.icar.cnr.it

A. Coronato · A. Testa
ICAR-CNR, Via P. Castellino 111, 80125 Napoli, Italy
e-mail: antonio.coronato@na.icar.cnr.it

48.1 Introduction

Today, more and more elderly people are challenged by acute and chronic illnesses and injuries. It has been estimated that eight out of ten elderly people are living with the health challenges of one or more chronic diseases. Hence, the development of systems for the remote monitoring of health conditions has earned lots of attention in the academia and the industry during the last years. This is also justified by the ever increasing healthcare costs and the increasing aging of the world population [1]. The remote monitoring of a patients' health status, while they are out of the hospital in their personal environment, helps to greatly reduce hospitalization costs, while offering the possibility of slowly progressing their chronic diseases and ensure continued recovery after being discharged from an acute care setting.

To this purpose, cabled measurement equipment is already used to guarantee reliable and robust control of vital signs. However such systems complicate patient autonomy and mobility. Hence, wireless technologies and mobile devices are starting to be applied to build more comfortable and patient-friendly health monitoring systems [2].

Despite these advantages, the use of wireless channels and the adoption of commodity hardware/software platforms, such as smartphones, pose new challenges on the correct functioning of health monitoring systems. Wireless channels can be affected by packet loss, due to shadowing and absence of signal coverage. Commodity platforms are not immune from failures, which could affect both the hardware and the control software. Finally, cheap and wireless-enabled medical devices can exhibit wrong readings and temporary disconnections from the so-called Body Area Network (BAN [3]). These issues may induce the medical staff to take wrong decisions or to administer wrong dosages of medicine. In turn, these decision can happen to be fatal for the patient.

For these reasons, the problem of failure detection and management in health monitoring systems is starting to be addressed in the literature, especially for mobile systems. However, several studies are based on simplistic failure assumptions or on basic fault-tolerance schemes (such as, sensor redundancy) which are not assured to cover all possible failure scenarios. For instance, sensor replication is ineffective against smartphone failures.

We claim the importance of analyzing the possible failure modes in the early stages of the development process, e.g., during the design, in order to come-up with architectural solutions able to face a large set of critical failures. This is a well established practice in the electronic field, where Failure Modes and Effect Analysis (FMEA) results are available for single medical devices. Nevertheless, this is not enough to characterize a whole mobile health monitoring system.

To fill the gap in the knowledge about the possible threats that may affect the correct functioning of health monitoring systems, this paper reports the results of a FMEA conducted to identify the failure modes of the main components composing such systems. The analysis takes advantage of our past experience and detailed field studies on the dependability of mobile devices, wireless communication

technologies, such as Bluetooth, and wireless sensor networks (WSNs), and builds on such results to propose a comprehensive characterization of the problems that may affect modern health monitoring systems. The resulting FMEA table is meant to be a guidance tool to direct future research efforts towards the realization of more dependable health monitoring systems.

The rest of the paper is structured in the following paragraphs. The related work is presented in [Sect. 48.2](#). [Section 48.3](#) introduces the FMEA methodology used to analyze the possible failures. [Section 48.4](#) describes the typical architecture of a mobile health monitoring system. In [Sect. 48.5](#) we discuss about the results on the realized FMEA. Finally, [Sect. 48.6](#) reports our concluding remarks.

48.2 Related Work

The recent research is progressively recognizing the need of novel solutions to build dependable health monitoring systems. These solutions mainly focus on two key issues: node failures and wireless network interference.

Regarding node failures, both WSNs and BANs may suffer from intentional or unintentional node removal or unresponsive nodes. While in WSNs, this issue can be resolved with new path discovery or redundant paths, in BANs this may cause the loss of important physiological data being monitored by the failing sensor. A combination of node redundancy and multi-sensor data fusion was one of the solutions proposed to face these issues [4, 5]. The introduction of redundant sensors measuring the same physiological sign (i.e., two or more oximeter sensors) avoids the loss of any vital data if a node becomes compromised or faulty. In addition, they can serve to facilitate multiple paths when the routing becomes an issue.

Interference is a major concern with all wireless devices, and has the potential to cause significant delays and data loss. Sensors can interfere with each other in the BAN as well as being subject to environmental noise. This is partially due to the lack of harmonious regulations and standards, as demonstrated in [6, 7]. A solution would be to eliminate the wireless aspect of the intra-BAN network [8, 9]. BAN systems such as MITHrill [8], SMART [8], and MobiHealth [10] all employ a wired connectivity between sensors and the aggregator. However, these solutions strongly limit the usability of the system, especially for elderly people, and makes it hard to interconnect all the sensors to commodity mobile devices, such as patients' smartphones, hence requiring ad-hoc aggregating devices which increase the overall cost of the system.

Hence, current solutions address only partial issues (e.g., node failures only) or they completely avoid to face other problems (e.g., using cables instead of wireless channels). However, several unexplored issues can limit the adoption of mobile health monitoring systems, such as smart phone failures, cellular network connectivity, and many others. Each failure mode in turn needs proper countermeasures to be handled at runtime. Thus, a more comprehensive view of the failure modes of these systems is needed.

48.3 FMEA Fundamentals

To properly evaluate a process or product for strengths, weaknesses, potential problem areas or failure modes, and to prevent problems before they occur, a **Failure Modes and Effects Analysis (FMEA)** can be conducted. FMEA is a team-based, systematic and proactive approach for identifying the ways that a process or design can fail, why it might fail, and how it can be made safer [11]. The purpose of performing an FMEA, as described in US MIL STD 1629 [12], is to identify where and when possible system failures could occur and to prevent those problems before they happen. If a particular failure could not be prevented, then the goal would be to prevent the issue from affecting health care organizations in the accreditation process.

An FMEA provides a systematic method of resolving the questions: *how can a process or product fail? What will be the effect on the rest of the system if such failure occurs? What action is necessary to prevent the failure?*. It represents a procedure for analysis of potential failure modes within a system for classification by the severity and likelihood of the failures. To realize a FMEA, the system is divided in components/functions that are divided in subcomponents/subfunctions; it considers a table in which the rows are composed by the subcomponents/subfunctions and the columns represent respectively the failure modes, the possible causes and the possible effects.

The FMEA team determines, by failure mode analysis, the effect of each failure and identifies single failure points that are critical. It may also rank each failure according to the criticality of a failure effect and its probability of occurring.

There are a number of reasons why this analysis technique is very advantageous. Here are just a few:

- FMEA provides a basis for identifying root failure causes and developing effective corrective actions;
- The FMEA identifies reliability and safety critical components;
- It facilitates investigation of design alternatives at all phases of design;
- It is used to provide other maintainability, safety, testability, and logistics analyses.

Since FMEA is effectively dependent on the members of the team which examines the failures, it is limited by their experience of previous failures. If a failure mode cannot be identified, then external help is needed from consultants who are aware of the many different types of product failure. FMEA is thus part of a larger system of quality control, where documentation is vital to implementation. In our case, we based the analysis both on our previous studies on different system components (such as WSNs, smart phones, and short range communication technologies) and on FMEA results available on some subcomponents, such as medical devices.

48.4 Mobile Health Monitoring Systems

To perform the FMEA, it is necessary to analyze in detail the typical architecture of current mobile health monitoring systems, in order to identify the individual components that compose them.

Recently, several health monitoring systems have been proposed in the market. We focus on three popular implementations, such as, the *MedApps System*, the *Nicolet Ambulatory Monitor System* and a system used by the *Center for Technology and Aging*.

The MedApps System [13] uses cellular, wireless and wired technologies with cloud-based computing to provide a healthcare connectivity platform that delivers flexible and scalable remote distribution. MedApps is designed to work with multiple external and internal devices. Patient data is collected, analyzed and forwarded, via cell phone to servers, providing a more robust picture of the patients' health.

The Nicolet Ambulatory Monitor System [14] is ideal for patients of all ages and combines a flexible, high quality diagnostic unit. Nicolet is a flexible, robust system specifically designed to handle the requirements of long-term monitoring. This system continuously monitors acutely ill patients at risk for brain damage and secondary injury, and it diagnoses patients' cerebral function (premature neonates to older adults).

In [15] authors discuss two areas of opportunity for remote patient monitoring: (i) *Chronic Disease Management and Post-Acute Care Management* and (ii) *Patient Safety*. In alignment with the mission of the *Center for Technology and Aging*, the solution proposed in [15] focuses on technology-enabled innovations, such as wireless connectivity, predominantly aimed at improving the health of older adults and promoting independent living in community-based, home, and long-term care settings.

Considering the underlying architectures of these existent systems, we can note that a mobile health monitoring system is usually composed by a number of sensors (medical devices), a gateway device (a handheld device) and a medical station; typical communication means are bluetooth (within the Body Area Network—Intra BAN communication), WiFi and cellular (external to the BAN—Extra BAN communication). Vital data are sensed by sensors (i.e., oximeter, blood pump, electrocardiogram—ECG, etc.) and transmitted to a mobile device over a bluetooth network. Next, data are sent to a remote station deployed, for an example, in a hospital by means of either a WiFi or a cellular connection (the medical center location).

In this typical network, we can observe that possible failures can occur in medical devices, in the bluetooth communication, in the mobile device, during the WiFi/cellular communication and finally in the local monitoring station of the caregiver. Figure 48.1 depicts the components of a generic mobile health monitoring system.

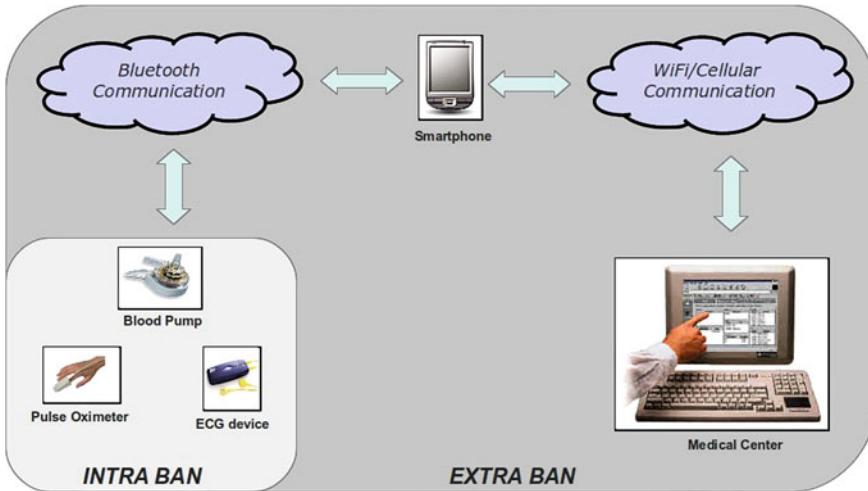


Fig. 48.1 A mobile health monitoring architecture

48.5 FMEA Results

In this section we present the results of the FMEA we performed on mobile health monitoring systems. The most frequent failure occurrences have been derived from past experiences on real architectures and from the existing literature, trying to relate failure occurrences with potential causes (faults).

The results related to the generic health monitoring system architecture presented in Sect. 48.4 are summarized in Table 48.1. With respect to the general architecture, we neglect the Medical Center location, since we assume it to be more reliable and under the direct control of the medical staff, who can suddenly intervene in case of failures (e.g., they can connect to the system using a different machine). Hence, we prefer to focus on the components which have to be used by patients, who might not be technology experts and who need to rely on a monitoring system able to work despite the occurrence of accidental failures.

In particular, we focus on technology-related failures, such as, failures due to hardware faults, software faults, or communication problems. Failures due to physical damage of nodes (e.g., physical crashes due to accidents or very adverse weather conditions), malicious activities (e.g., manual, and unexpected, node withdrawal or substitution), and security threats are excluded from the analysis.

For each component/function (sub-component/sub-function if it is present) of the system, failure modes, potential effects and possible causes are reported.

We identified four components/functions: the node (i.e., the sensor used to monitor the patient), the Intra BAN communication, the Extra BAN communication, and the gateway (i.e., the smartphone of the patient). Eight sub-components/sub-functions have been identified for the *node* component: the sensor board, the power supply unit, the CPU, and the OS (such as [16, 17] which are used in

Table 48.1 Failure Mode and Effect Analysis of a mobile health monitoring system

| Component | Sub-component | Potential failure mode | Potential effects of failure | Potential causes of failure |
|---------------------------|------------------------|-----------------------------------|--|--|
| Node (the medical sensor) | Sensor board | Stuck at zero | The device is out-of-order; it does not deliver any output to inputs | Sensing hardware |
| | | Null reading | The device delivers null output values | Sensing hardware |
| | | Out of scale reading | The device delivers no meaningful values | Sensing hardware |
| | Power supply | Stuck at zero | The device is out-of-order; it does not deliver any output to inputs | Natural energy exhaustion |
| | | Reset | The node resets itself to its initial conditions | Anomalous current request that cannot be supplied by batteries |
| | CPU | Stuck at zero | The device is out-of-order; it does not deliver any output to inputs | Micro-controller |
| | OS | Software hang | The device is powered on, but not able to deliver any output | Operating system's corrupted state |
| | ECG device adhesive | Incorrect reading | Wrong data values, irritation or rash of skin | Skin contact |
| | ECG device electrolyte | Incorrect reading | Wrong data values, irritation or rash of skin | Skin contact |
| | Patient cable | Discontinuous readings | Noise, wrong data values | Defective wire |
| | Blood pump | External leak (negative pressure) | Air leak into system | Tubing set leak |
| | | External leak (positive pressure) | Blood loss | Tubing set leak |
| | | Flow too low | Too little blood, alarms | Controller failure, rotor failure |
| | | Flow too high | Too much blood flow, alarms | Controller failure |

(continued)

Table 48.1 (continued)

| Component | Sub-component | Potential failure mode | Potential effects of failure | Potential causes of failure |
|-------------------------|-----------------------------------|--|--|---|
| Intra BAN communication | Transport and routing | Packet loss | The radio packet is not delivered | Packet corruption Buffer overrun |
| | | Isolation | The node is not longer connected to the sink node | Failure of all forwarding nodes |
| | Bluetooth stack | Bluetooth stack failure | A bluetooth module (e.g., L2CAP, BNEP, etc.) fails | Bluetooth stack's corrupted state |
| | Bluetooth channel | Header corruption | Header delivered with errors | Packet corruption |
| | | Header length mismatch | Header length deviates from the specified one | Packet corruption |
| | | Payload corruption | Payload delivered with errors | Packet corruption |
| Extra BAN communication | Data delivery failures | The network is not able to deliver the required amount of measurements | The number of failed nodes is more than a given threshold | |
| | Cellular/WiFi network unavailable | Monitoring stopped | Area without cellular/WiFi signal | |
| Gateway | Device (the smartphone) | Freeze | The device's output becomes constant; the device does not respond to the users input | Systems corrupted state |
| | | Self-shutdown | The device shuts down itself; no service is delivered at the user interface | Natural energy exhaustion or self-reboot due to corrupted state |
| | | Unstable behavior | The device exhibits erratic behavior without any input inserted by the user | System/application corrupted state |
| | | Output failure | The device delivers an output sequence that deviates from the expected one | System/application corrupted state |

(continued)

Table 48.1 (continued)

| Component | Sub-component | Potential failure mode | Potential effects of failure | Potential causes of failure |
|-----------|-----------------------|------------------------|--|--|
| | | Input failure | User inputs have no effect on device behavior | System/ application corrupted state; natural energy exhaustion |
| | Bluetooth application | Inquiry/scan failure | The scan procedure terminates abnormally | A bluetooth module fails or device out of range |
| | | Discovery failure | The discover procedure terminates abnormally | A bluetooth module fails or device out of range |
| | | Connect failure | The device is unable to establish a connection | A bluetooth module fails or device out of range |
| | | Packet loss | Expected packets are not received | Packet corruption |
| | | Data mismatch | Packets are delivered with errors in the payload | Memoryless channel with uncorrelated errors |

medical devices) are the general components of a node, and their analysis is based on our previous study on sensor networks [18]. In addition, we considered the failures of some specific medical devices, such as the ECG sensor (divided in the ECG Device Adhesive, the ECG Device Electrolyte), the patient cable, and the Blood Pump. The failures of such devices have been identified starting from existing studies, such as [19, 20]. Clearly, other devices can be added to the analysis if used in a specific setting.

Three sub-components/sub-functions have been identified for the *Intra BAN* function: transport/routing, Bluetooth stack and Bluetooth channel. These are based on our previous studies on sensor networks and on the Bluetooth protocol [18, 21]. Finally, one component/function has been identified for the *Extra BAN*, starting from [18, 22], and two sub-components/sub-functions have been identified for the *Gateway* component: the device (i.e., the smartphone, starting from our previous experiences on smartphone failures [23]), and the Bluetooth application, which is responsible to gather the measurements from Bluetooth medical devices (in facts, the majority of wireless medical devices use Bluetooth as the communication technology). In [21] we noticed that several failures may affect Bluetooth applications, due to problems of the underlying Bluetooth modules.

In the following, we detail the analysis performed for each identified component/function.

48.5.1 Node (*Generic Medical Device Components*)

From the prospective the mission of the BAN, a node is failed when (i) it is no longer able to deliver its measurements to the gateway, and (ii) it is not longer able to provide meaningful measurements. This can be due the malfunction of one of the components of the node, as detailed in the following.

48.5.1.1 Sensor Board

We assume the sensor board can fail according to three failure modes: *stuck-at-zero*, *null reading* and *out-of-scale reading*. A stuck-at-zero of the sensor board produces the effect of a out-of-order device, which does not deliver any outputs to external inputs. Potential causes lay into faults of the sensing hardware (e.g., as can be observed in [24], the humidity sensor produces a short circuit, causing a high current drain which turns off the overall node). Null readings cause the sensor to deliver null output values, for a certain interval of time. This may be caused by temporary short circuits that also cause the node to drain excessive power from batteries, hence shortening the overall lifetime of the node [24]. Out-of-scale readings cause the sensor board to provide no meaningful outputs, for a certain interval of time.

48.5.1.2 Power Supply

The power supply component may exhibit stuck-at-zero as well as reset failure modes (i.e., the node shutdowns and restarts itself). The former is due to battery energy exhaustion. The latter can be caused by anomalous power requests that cannot be supplied by batteries, e.g., the residual charge is not sufficient to provide the required amount of power.

48.5.1.3 CPU

The micro-controller can be affected by temporary or permanent failures, which prevent it to work correctly, hence delivering constant outputs.

48.5.1.4 OS

Software defects (bugs) or single event upsets (bit flips) may corrupt the state of the embedded operating system, causing the whole device to hang.

48.5.2 Node (*Specific Medical Device Components*)

48.5.2.1 ECG

Possible hazards are incorrect readings due to short circuits or too much current. This can be due to skin contacts of the ECG adhesive or electrolyte, which in turn may cause irritation or rash of patient's skin. The device receives a shock and stops to function. In these cases, it is needed to choose an adhesive or an electrolyte with low likelihood of reaction.

48.5.2.2 Patient Cable

Often, the medical device (even if wireless) is equipped with cables to connect the device (e.g., the ECG) with vital signs sensors (e.g., to be put on a finger). Patient cable failure is probably the most common cause of unacceptable vital sign tracings. Patient cable failure can show up as artifacts, noisy tracings, failure to acquire signal or a long time to acquire a usable signal. On many of the older cables and most stress test cables individual patient leads are replaceable. On newer medical units, with molded cables, it is necessary to replace the whole cable.

48.5.2.3 Blood Pump

Blood pumps are very important devices because they control the blood flow in a patient. But some failures are fatal and so it need control if there is too much blood flow or if the flow is too low. Other failures studied for this kind of devices represent external leaks; it is necessary to monitor the pressure and to observe if it is positive or negative. In the case of negative pressure, it need launch immediately an air in blood alarm because it is very dangerous for the patient life.

48.5.3 Intra-BAN

At the Intra-BAN level, we have distinguished the failures due to the transport and routing sub-functions from the ones specific to Bluetooth. Details are provided in the following.

48.5.3.1 Transport and Routing

During the transport of packets, losses may occur, i.e., the packet is not delivered to its intended destination for instance due to corruption or routing buffer overrun.

In the case of multi-hop communication, managed by a routing algorithm, the communication can also be compromised by isolation failures: a node, which does not fail itself, can manifest an isolation failure when it is no longer connected to the gateway.

48.5.3.2 Bluetooth Stack

The Bluetooth software stack is corrupted due to faults into one of its modules, such as L2CAP, BNEP, RFCOMM, etc.

48.5.3.3 Bluetooth Channel

Three Bluetooth channel level (i.e., the Baseband level) failures have been identified: Baseband header corruption, length mismatch, i.e., a mismatch between the packet length reported into the Baseband header and the actual one, and Baseband payload corruption. These failures are due to packet corruption and can in turn cause wrong readings or packet loss at the higher levels.

48.5.4 Extra-BAN

Data delivery failures occur when the system is not able to deliver the required amount of measurements to the medical center. Also, the monitoring application can result completely stopped if the cellular (of WiFi) network is unavailable.

48.5.5 Gateway

48.5.5.1 Device (the smartphone)

An analysis of the main failure modes of smart phones, performed in [23], revealed that these device may exhibit several failures, due to both hardware issues and software defects. Specifically, five failures have been identified: freeze (the device is completely blocked, and only pulling-out the battery restores proper operation), self-shutdown (the device resets itself due to battery exhaustion or reaction to a system corrupted state), unstable behavior, output failure, and input failure (due to system or application corrupted states).

48.5.5.2 Bluetooth Application

The application governing the Bluetooth communication may exhibit a variety of failures according to the utilization phase where they occur, i.e., inquiry/scan and discovery phases, connection, and data transferring. Failures during the connection can occur either while the connection is set up or while the role of the device is switched from master to slave. Unexpectedly, failures during data transfer, such as packet loss and mismatches in the received data, are experienced, despite error control mechanisms performed by Baseband. Correlated errors (e.g., bursts) can occur due to the nature of the wireless media, affected by multi-path fading and electromagnetic interferences.

All of these analyzed failures cause abnormal vital sign readings, or even it can happen that a value is not received at the Medical Center location; in this case an inaccurate monitoring is provided, potentially resulting in a significant hazard to patients. Health monitoring systems must be aware of all the possible failures, in order to react to them or, at least, to detect them. For instance, in case of failure detection, a possible action can be to call to the patient's home or to call to an emergency contact to suddenly check the patient status and restore the normal operation of the system.

48.6 Conclusions and Future Work

In this paper we conducted a Failure Mode and Effect Analysis of mobile health monitoring systems. The analysis considered the main components and the main medical devices adopted in a typical health monitoring system, and it is based partially on our previous experience and studies on some of the components, i.e., Bluetooth, sensor networks, and smart-phones, and partially on the results already available for medical devices. Even if the analysis represents only a base for further studies, it reveals that several failure modes are usually neglected by current health monitoring solutions, where only node crashes are considered, hence exposing patients to potential health risks. Future efforts will be devoted to the definition of architectural solutions for mobile health monitoring systems, able to take into account the failure modes identified in the this paper, and capable to detect failures at runtime in order to propose proper countermeasures, toward the goal of building more dependable health monitoring systems in the future.

References

1. Hao Y, Foster R, (2008) Wireless body sensor networks for health-monitoring applications. *Physiol Meas* 29(11): R27–R56. <http://dx.doi.org/10.1088/0967-3334/29/11/R01>
2. Paksuniemi M, Sorvoja H, Alasaarela E, Myllyla R (2005) Wireless sensor and data transmission needs and technologies for patient monitoring in the operating room and intensive care unit. In: Engineering in medicine and biology society, 2005. The 27th annual international conference of the IEEE-EMBS 2005

3. O'Donovan T, O'Donoghue J, Sreenan C, Sammon D, O'Reilly P, O'Connor K (2009) A context aware wireless body area network (ban). In: Pervasive computing technologies for healthcare, 2009. Pervasive health 2009, 3rd international conference on, April 2009, pp 1–8
4. Baskiyar S (2002) A real-time fault tolerant intra-body network. In: Local computer networks, 2002. Proceedings of the LCN 2002. The 27th annual IEEE conference on, Nov 2002, pp 235–240
5. Curiaç D-I, Volosencu C, Pescaru D, Jurca L, Doboli A (2009) A view upon redundancy in wireless sensor networks. In: Proceedings of the 8th WSEAS international conference on signal processing, robotics and automation. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2009, pp 341–346. <http://dl.acm.org/citation.cfm?id=1558916.1558975>
6. Hanna S (2009) Regulations and standards for wireless medical applications. In: Proceedings of the 3rd international symposium on medical information and communication technology, Feb 2009
7. Monrose K, Spadotto E, Hawkins J (2009) Ict convergence, confluence and creativity: the application of emerging technologies for healthcare transformation. In: Proceedings of the 3rd international symposium on medical information and communication technology, Feb 2009
8. Chen M, Gonzalez S, Vasilakos A, Cao H, Leung VC (2011) Body area networks: a survey. *Mob Netw Appl* 16:171–193. <http://dx.doi.org/10.1007/s11036-010-0260-8>
9. Hanson M, Powell H, Barth A, Ringgenberg K, Calhoun B, Aylor J, Lach J (2009) Body area sensor networks: challenges and opportunities. *Computer* 42(1):58–65
10. Mobihealth. <http://www.mobihealth.com/home>
11. Latino RJ, Flood A (2004) Optimizing fmea and rca efforts in health care. *J Healthc Risk Manag* 24(3):21–28. <http://dx.doi.org/10.1002/jhrm.5600240305>
12. Us mil std 1629 (1980) Procedure for performing a failure mode, effect and criticality analysis, method 102, Nov 1980
13. Institute of medicine committee on the future of emergency care in the United States health system (2006) Report brief. National Academy of Science, Washington, DC
14. Carefusion, nicoleet. <http://www.carefusion.com/medical-products/neurology/neuro-diagnostic-monitoring/eeg/nicolet-ambulatory-monitor.aspx>
15. Center for Technology and Aging (2009) Technologies for remote patient monitoring in older adults, Dec 2009
16. Qnx. <http://www.qnx.com/solutions/industries/medical/>
17. Threadx. <http://www.qnx.com/solutions/industries/medical/>
18. Cinque M, Cotroneo D, Martinio CD, Russo S (2007) Modeling and assessing the dependability of wireless sensor networks. In: IEEE symposium on reliable distributed systems, vol. 0, pp 33–44
19. Technical Manual Headquarters No. 5-698-4, Department of the army, Washington, DC, 29 Sept 2006
20. Toltec international inc. <http://www.toltec.biz/index.htm>
21. Carrozza G, Cinque M (2009) Modeling and analyzing the dependability of short range wireless technologies via field failure data analysis. *J Softw* 4(7):707–716. <http://ojs.academypublisher.com/index.php/js/article/view/0407707716>
22. Institute for healthcare improvement. <http://app.ihl.org/Workspace/tools/fmea/>
23. Cinque M, Cotroneo D, Kalbarczyk Z, Iyer R (2007) How do mobile phones fail? a failure data analysis of symbian os smart phones. In: Dependable systems and networks, 2007. DSN '07, 37th annual IEEE/IFIP international conference on, June 2007, pp 585–594
24. Szewczyk R, Polastre J, Mainwaring A, Culler D (2004) Lessons from a sensor network expedition. *EWSN* 4:307–322

Chapter 49

SemFus: Semantic Fusion Framework Based on JDL

Havva Alizadeh Noughabi, Mohsen Kahani and Behshid Behkamal

Abstract Data fusion techniques combine data from multiple sources and gather related information to achieve more specific inferences than could be achieved by using a single source. The most widely-used method for categorizing data fusion-related functions is the JDL model, but it suffers from semantics and syntax issues. In order to achieve semantic interoperability in a heterogeneous information system, the meaning of the information that is interchanged has to be understood across the systems. Semantic conflicts occur whenever two contexts do not use the same interpretation of the information. Using semantic technologies for the extraction of implicit knowledge is a new approach to overcome this problem. In this paper a semantic fusion framework (SemFus) is proposed based on JDL which can overcome the semantic problems in heterogeneous systems.

49.1 Introduction

In recent years, significant attention has been focused on multi sensor data fusion for both military and non-military applications such as robotics, wireless sensor networks, environmental monitoring and medical applications.

Many definitions of data fusion have been provided through the years, most of them derived from military and remote sensing fields. In 1991, the data fusion working group of the Joint Directors of Laboratories (JDL) defined data fusion as a “multilevel, multifaceted process dealing with the automatic detection, association, correlation, estimation, and combination of data and information from multiple sources” [1]. Another definition of data fusion is “the combination of data

H. A. Noughabi (✉) · M. Kahani · B. Behkamal
Ferdowsi University of Mashhad, Mashhad, Iran
e-mail: alessandro.testa@na.icar.cnr.it

from multiple sensors, and related information provided by associated databases, to achieve improved accuracy and more specific inferences than could be achieved by the use of a single sensor alone” [2].

There are different areas of information fusion applications including geospatial information systems, wireless sensor networks, intelligent transport systems, business intelligence, business performance management, loyalty cards and bio-informatics. In most of these areas the process of data/information fusion is nearly the same and suffers from semantic and syntax issues.

With the growth of Semantic Web (SW) technology, various research areas have been adopted SW technologies. So, in this paper a semantic fusion framework (SemFus) is presented based on JDL which can overcome the semantic problems in heterogeneous systems. In the following, we explain more about semantic heterogeneity and our motivation. In the next section, some related works are discussed in three main groups. The base JDL model is presented in Sect. 49.4. Then SemFus architecture is presented and an experiment of applying proposed model is discussed in Sect. 49.5. Finally, Sect. 49.6 concludes the paper and discusses future directions.

49.2 Motivation

In order to achieve semantic interoperability in a heterogeneous information system, the meaning of the information that is interchanged has to be understood across the systems. Semantic conflicts occur whenever two contexts do not use the same interpretation of the information. Goh identifies three main causes for semantic heterogeneity [3]:

Confounding conflicts occur when information items seem to have the same meaning, but differ in reality, e.g. owing to different temporal contexts.

Scaling conflicts occur when different reference systems are used to measure a value. Examples are different currencies.

Naming conflicts occur when naming schemes of information differ significantly. A frequent phenomenon is the presence of homonyms and synonyms.

Using semantic technologies for the extraction of implicit knowledge is a possible approach to overcome the problem of semantic heterogeneity.

Our proposals are made extending the model to include semantic in both data and levels of JDL model. The main contributions of this paper can be summarized as follows:

- Using ontology to describe the sources (Semantic refinement of all objects).
- Providing a semantic contextual description of the relationship among objects and observed events (Semantic Situation and threat refinement).
- Proposing a semantic reasoner to infer logical consequences from a set of asserted facts in Rule DB.
- Using RDFizer to convert preprocessed data into standard Resource Description Format (RDF) and stores it in the RDFstore.

49.3 Related Work

To investigate the state of the art in the field of information fusion, some related works are comparatively studied and classified in three main groups: information fusion models, JDL-based works and semantic fusion works.

49.3.1 Information Fusion Models

All of the architectures and models proposed to design information fusion systems can be classified as information-based, activity-based and role-based [4].

Information-based models are centered on the abstraction of the data generated during fusion, like JDL and DFD [5] model. Activity-Based Models are specified based on the activities that must be performed by an information fusion system. In such models, the activities and their correct sequence of execution are explicitly specified, like Boyd Control Loop [6], Intelligence Cycle [7] and Omnibus Model [8]. Role-based models represent a change of focus on how information fusion systems are modeled and designed. In such models, information fusion systems are specified based on the fusion roles and the relationships among them. The two members of this generation are the Object Oriented Model and the Frankel-Bedworth architecture [4].

49.3.2 JDL-Based Works

In [9] the role of the human in a fusion model is demonstrated and level 5 is added to JDL model as user refinement. Another extension to the JDL model is a hybrid hierarchical structure proposed as ProFusion2 (PF2) [10]. They have aimed to answer this question “how the JDL model can be applied in multi sensor automotive safety systems, since new sensors are integrated on-board, while new functions support the driver, intervene and control the vehicle”.

In [11] researchers a refinements to the existing definitions of the various levels of JDL model is proposed and interaction of levels is discussed.

The main idea of [12] is to apply JDL model in the bioinformatics domain. They have made a mapping of the JDL data fusion model to bioinformatics and investigated its applicability to this domain by applying the model in a stem cell differentiation study.

Another work has revised and expanded JDL model to facilitate the cost-effective development, acquisition, integration and operation of multi-sensor/multi-source systems [13].

In [14] the aim of the authors is summarized as “regarding improvements in the understanding of internal processing within a fusion node and extending the model

to include remarks on issues related to quality control, reliability, and consistency in data fusion processing, assertions about the need for co-processing of abductive/inductive and deductive inference processes, remarks about the need for and exploitation of an onto logically-based approach to data fusion process design, and extensions to account for the case of Distributed Data Fusion (DDF)".

Another revision of JDL model is based on partitioning of data fusion functions [15]. Partitioning is designed to capture the significant differences in the types of input data, models, outputs, and inference appropriate to broad classes of data fusion problems. In general, the recommended partitioning is based on different aspects of a situation for which the characterization is of interest to a system user. The last model, presented in [16], has served a novel reference model for Modeling and Simulation by comparing the levels distinguished in the JDL-U model with activities and phases in simulation projects.

49.3.3 Semantic Fusion Works

There are some works in semantic fusion. Most of them use ontology for semantic identification of resources. In [17] discussed about role of ontologies in data integration. In paper [18] proposed the use of probabilistic ontologies within a service-oriented architecture as a means to enable semantic interoperability in net-centric fusion systems; while in [19] researchers develop use cases in which ontologies are used both for the fusion process itself and for the development of fusion systems. Kokar et al. [20] provides a description of the classes and the properties in the ontology, and illustrates the formalization with some simple examples; and finally [21] presented an ontology fusion for agents that they designed it and sketched an approach which aims a developer to reduce difficulties in integration and adaptation of software entities into a heterogeneous distributed system. Knowledge is formally expressed in form of ontology and axiomatic logical model.

Another application of ontology is semantic integration of heterogeneous information sources. In [22] researchers explain ontology engineering methods and tools used to develop ontologies for information integration. It provides an approach for adaptive, context-aware information retrieval and reviews the use on ontologies for the integration of heterogeneous information sources. In Boury-Brisset [23] ontological engineering is used for situation and threat refinement and Smart et al. [24] suggested an approach featuring domain ontologies, reasoning capabilities, semantic queries and semantic integration techniques provides the basis for an integrated framework for improving situation awareness in military coalition contexts; and finally Gagnon [25] proposed an ontology-based information integration with a local to global ontology mapping as an approach to the integration of heterogeneous data sources.

The other works done in semantic fusion, focus on different areas. The main focus of [26] is to provide a conceptual framework for formally capturing various

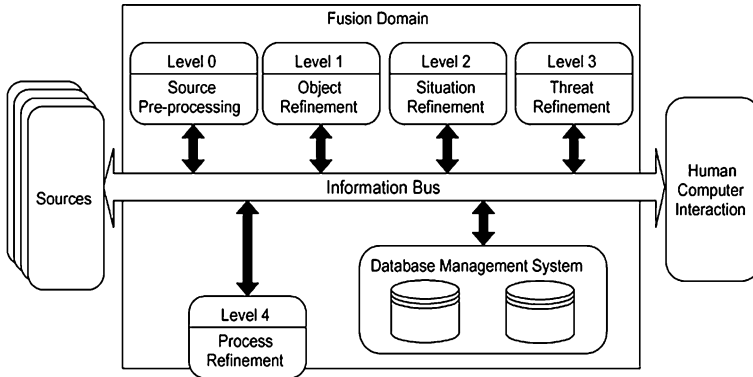


Fig. 49.1 The JDL model

sorts of complex relation-types, which can serve as a means for a more thorough decomposition of objects, attributes/properties, events, processes, and relations, necessary for higher level fusion processing. Another research describes a Situation Awareness Assistant (SAWA) that facilitates the development of user-defined domain knowledge in the form of formal ontologies and rule sets and then permits the application of the domain knowledge to the monitoring of relevant relations as they occur in evolving situations [27]. In [28] a three-layer flexible architecture is presented for sensor data fusion by exploiting the semantic web advances. This architecture contains three layers: the Data layer, the Processing layer and the Semantic layer; And finally [29] focus on reusability of fusion process and designed the basic concept of SAW core ontology.

49.4 An Overview of JDL Model

JDL is a popular model in the fusion research community. The model is composed of five processing levels, an associated database, and an information bus connecting all components. Its structure is depicted in Fig. 49.1 and its components are described as follow [2, 30]:

Sources. Sources are responsible for providing the input information, and can be sensors, a priori knowledge, e.g. reference and geographical information, databases, or human input.

Database Management System. This system supports the maintenance of the data used and provided by the information fusion system.

Human Computer Interaction. HCI is a mechanism that allows human input, such as commands and queries, and the notification of fusion results through alarms, displays, graphics, and sounds.

Level 0, Source Preprocessing. This level attempts to do estimation and prediction of signal and various preprocessing of data, e.g. normalization of signal

measurements, handling of missing values in the data set, handling of incomplete data sets, filtering out low quality measurements.

Level 1, Object refinement. This level attempts to do estimation and prediction of entity states. In fact this level transforms the data into a consistent structure and identifies objects.

Level 2, Situation refinement. Situation refinement tries to provide a contextual description of the relationship among objects and observed events. It uses a priori knowledge and environmental information to identify a situation.

Level 3, Threat refinement. Threat refinement evaluates the current situation projecting it into the future to identify possible threats, vulnerabilities, and opportunities for operations.

Level 4, Process refinement. This is responsible for monitoring the system performance and allocating the sources according to the specified goals and processing to support mission objectives.

49.5 SemFus: Semantic Fusion Framework

Our proposed model is based on JDL, so the main structure of levels is correspond to JDL levels. On the other hand, semantic technologies are used both in data and levels to obtain a semantic model for information fusion. So, at first refined definitions of semantic levels is presented. Then the way of applying ontology for semantic representation and interaction is investigated. For better understanding of the SemFus, the position of proposed model is illustrated in fusion tree and finally the steps of implementing will be presented. The structure of SemFus framework is depicted in Fig. 49.2.

49.5.1 Levels of SemFus

The JDL distinction among fusion “levels” (depicted in Fig. 49.1) provide an often useful distinction among data fusion processes that relate to the refinement of “objects”, “situations”, “threats” and “processes.” So, we have tried to refine definitions of the JDL levels.

Preprocessing (level 0) is defined as estimation and prediction of signal/object observable states on the basis of pixel/signal level data association and characterization. In level 1, Objects are described in RDF format and stored in RDFstore. For this purpose, RDFizer extracts data from Object or situation refinement level and convert data into RDF format and stores it in the RDFstore. A RDFstore is a system for storing and managing RDF data. Each resource is described by using concepts defined in the ontology and identified by an URI.

In the second level, situation identified semantically based on definitions of entities and their relationship described as RDF in level 1.

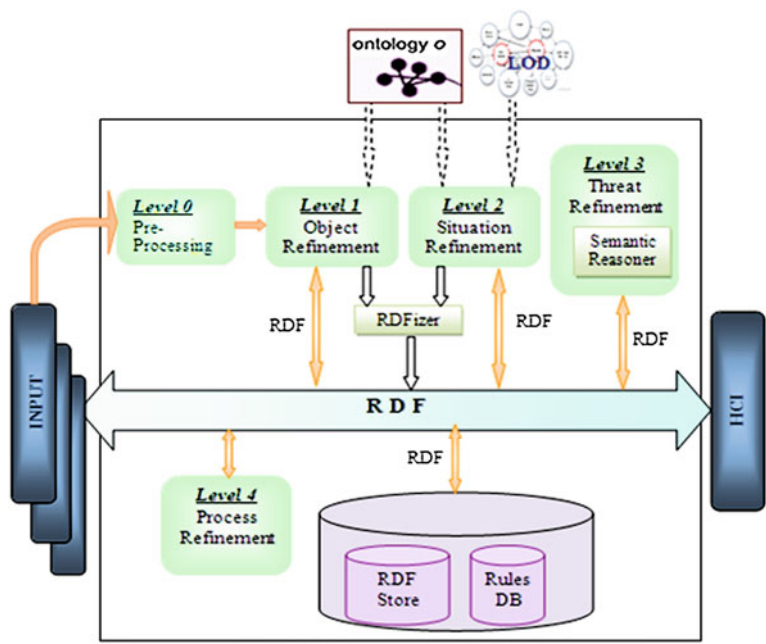


Fig. 49.2 SemFus framework

In addition to RDFstore, some other useful resources can be used for situation refinement including datasets of the linked datasets. Linked data describes a method of publishing structured data so that it can be interlinked and become more useful. It builds upon standard Web technologies such as HTTP and URIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried [31].

After situation refinement, the main task is assessment of situation to predict threats and opportunities for operations. By using a semantic reasoner, logical consequences from a set of asserted facts is inferred. Finally, Level 4 is responsible for monitoring the system performance and allocating the sources according to the specified goals and processing to support mission objectives.

49.5.2 SemFus Ontology

As it mentioned before, ontologies can be used in an integration task to describe the semantics of the information sources and to make the contents explicit. With respect to the integration of data sources, they can be used for the identification and association of semantically corresponding information concepts. But there are different ways of how to employ the ontologies. In general, three different

directions can be identified: single ontology approaches, multiple ontologies approaches and hybrid approaches [22]. Single ontology approaches use a global ontology providing a shared vocabulary for the specification of the semantics. Single ontology approaches can be applied to integration problems where all information sources to be integrated provide nearly the same view on a domain. But if one information source has a different view on a domain, finding the minimal ontology commitment becomes a difficult task. In multiple ontology approaches, each information source is described by its own ontology. This ontology architecture can simplify the change, but in reality the lack of a common vocabulary makes it extremely difficult to compare different source ontologies. To overcome the drawbacks of the single or multiple ontology approaches, hybrid approaches were developed. Similar to multiple ontology approaches the semantics of each source is described by its own ontology. But in order to make the source ontologies comparable to each other they are built upon one global shared vocabulary.

In SemFus, input data come from difference sensors and each of the information sources has a different view on a domain. Also, using a common vocabulary makes it easier to compare different source ontologies. So, hybrid approach is used to develop ontology as follows. Firstly, each source schema can be explicitly represented by a local ontology. All ontologies use the same representation language and are therefore syntactically homogeneous. Then, the global ontology provides a conceptual view over the schematically heterogeneous source schemas. The global ontology provides a high-level view of the sources. Therefore, a query can be formulated without specific knowledge of the different data sources. The query is then rewritten into queries over the sources, based on the semantic mappings between the global and local ontologies.

Also, a hybrid peer-to-peer system uses the global ontology as a mediator for query rewriting across peers and a common thesaurus or vocabulary, which can be formalized as ontology, can be used to facilitate the automation of the mapping process.

49.5.3 Comparison of SemFus and JDL

JDL and SemFUS provide a systemic view of information fusion. SemFUS can overcome the semantic problems by using of SW technologies. Also, a common thesaurus or vocabulary, which can be formalized as ontology, can be used to facilitate the automation of the mapping process. A brief feature comparison of SemFus and JDL is summarized in Table 49.1.

49.5.4 The Position of SemFus

For better conceptualization of SemFus, we have classified works done in data fusion area and structured as a tree. Figure 49.3 is depicted proposed fusion tree and illustrated the position of SemFus in compare with other works.

Table 49.1 Comparison of JDL and SemFus

| Characteristics | JDL | SemFus |
|-------------------------------------|-----|--------|
| Systemic view | + | + |
| Semantic conflicts | + | – |
| Standard format for description | – | + |
| Metadata representation | – | + |
| Global conceptualization | – | + |
| Support mapping | – | + |
| Semantic reasoning | – | + |
| Ontology based situation awareness | – | + |
| Connected to linked open data (LOD) | – | + |

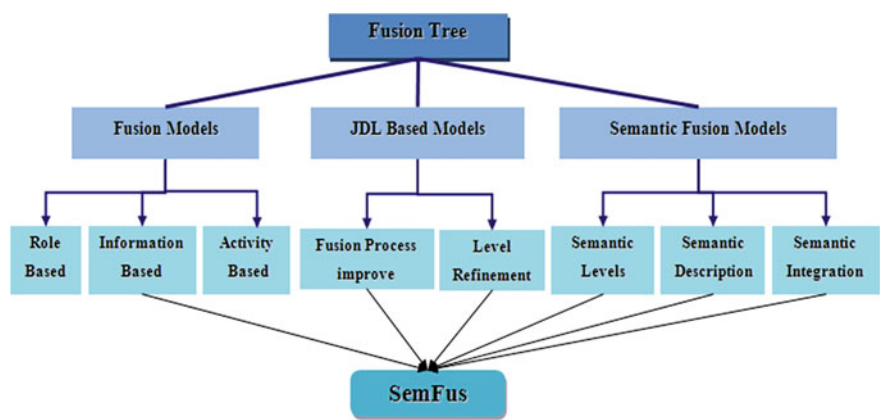


Fig. 49.3 Position of SemFus

49.5.5 Implementation of SemFus

To implement SemFUS framework, the steps are defined and summarized as follows:

- Providing a dataset (military, robotics, ...)
- Defining related ontologies
- Ontology-based object detection
- Ontology-based Situation awareness
- Semantic threat refinement
- Updating rule base using simulator and data mining tasks
- Result analysis and evaluation

These steps can be used in all domains including wireless sensor networks, environmental monitoring, smart home, medical applications and military environment. Here an experiment of applying proposed framework in a scenario is presented.

```
String queryString=
    "select ?obj" + "where { ?obj rdf:type scenario-ont:object }";
query= QueryFactory.create(queryString);
QueryExecution qe= QueryExecutionFactory.create(query,ontologymodel);
ResultSet results= qe.execSelect();
```

Fig. 49.4 A sample SPARQL in Jena

Fig. 49.5 An example of rules stored in rule database

```
[AlarmRule:
    (?s rdf:type scenario-ont:object)
    (?s scenario-ont:situationStatus warning)
    ->
    (alarm scenario-ont:becauseOf ?s)
```

At first must provide a sample dataset. At the second step, the ontology is developed for describing the properties of the objects of scenario. From the viewpoints of value type, the object properties can be classified into static and dynamic. The static values such as height and width can be set in the design phase and the dynamic values such as speed and direction are filled by dataset.

Then needed ontologies are implemented by Jena framework.

Afterward all data come from dataset are stored with time stamp in RDF store.

To perform levels of model like situation refinement some needed information are extracted from linked open datasets and RDFstore. We can retrieve needed data from RDFStore by SPARQL query. Figure 49.4 depicts a sample SPARQL on RDFStore in Jena.

Finally, semantic reasoner identifies treats of the environment using rules stored in rule database. An example of rules is shown in Fig. 49.5.

After all these steps, the rule database should be updated by new rules using data mining. To simplify the process of implementation, this step is not implemented in this experiment and it will be done in the future works.

49.6 Conclusion and Future Works

To achieve semantic interoperability in a heterogeneous information system, the meaning of the information that is interchanged has to be understood across the systems. In this paper a semantic fusion framework, SemFus, is proposed based on JDL to enrich the fusion process and overcome the semantic problems in heterogeneous systems. For this purpose, by discussing some related models in this area, previous works are classified in three main groups: fusion models, JDL-based models and semantic fusion. Then, SemFus framework is proposed and its position in the fusion tree is illustrated.

Finally, to show the functionality of SemFus practically, an experiment of a scenario is discussed.

In the future works, we are going to implement SemFus for other scenarios in different applications and evaluate it by measuring timeliness, accuracy, throughput, confidence, cost and so on.

References

1. White FE (1991) Data fusion lexicon. Data fusion subpanel of the joint directors of laboratories technical panel for C³, Code 4202. NOSC, San Diego
2. Hall DL, Llinas J (1997) An introduction to multi-sensor data fusion. *Proc IEEE* 85:6–23
3. Cheng HG (1997) Representing and reasoning about semantic conflicts in heterogeneous information sources. Doctoral dissertation, Sloan School of Management, MIT, Cambridge
4. Nakamura EF, Loureiro AA, Frery AC (2007) Information fusion for wireless sensor networks: methods, models, and classifications. *ACM Comput Surv* 39:3
5. Dasarathy BV (1997) Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proc IEEE* 85:24–38
6. Boyd JR (1987) A discourse on winning and losing. Unpublished set of briefing slides available at Air University Library, Maxwell AFB, Alabama
7. Shulsky AN, Schmitt GJ (2002) Silent warfare: understanding the world of intelligence, 3rd edn. Brassey's Inc., New York
8. Bedworth MD, O'brien JC (1999) The omnibus model: a new model for data fusion. In: *Proceedings of the 2nd international conference on information fusion, FUSION'99, ISIF, Sunnyvale*, pp 437–444
9. Blasch E, Plano S (2002) JDL level 5 fusion model: user refinement issues and applications in group tracking. *SPIE*, vol 4729. *Aerosense*, pp 270–279
10. Polychronopoulos A, Amditis A, Scheunert U, Tatschke T (2006) Revisiting JDL model for automotive safety applications: the PF2 functional model. In: *The 9th international conference on information fusion, Florence, 2006*
11. John JS (2007) Where's level 2/3 fusion—a look back over the past 10 years. In: *10th international conference on information fusion, 2007*
12. Synnergren J, Gamalielsson J, Olsson B (2007) Mapping of the JDL data fusion model to bioinformatics. *IEEE Explor* 44:1506–1511
13. Steinberg AN, Bowman CL, White FE (1999) Revisions to the JDL data fusion model. Sensor fusion: architectures, algorithms, and applications. In: *Proceedings of the SPIE*, vol 3719
14. Llinas J, Bowman CL, Rogova G, Steinberg AN, Waltz E, White FE (2004) Revisiting the JDL data fusion model II. In: *Proceedings of the 7th international conference on information fusion, Stockholm, 2004*
15. Steinberg AN, Bowman CL, White FE (2004) Rethinking the JDL data fusion model. In: *NSSDF conference proceedings, 2004*
16. DeVin LJ, Holm M, Ng AHC (2010) The information fusion JDL-U model as a reference model for virtual manufacturing. *J Robot Comp Integr* 26(6):629–638
17. Cruz IF, Xiao H (2005) The role of ontologies in data integration. *J Eng Intell Sys* 13(4): 245–252
18. Laskey KB, da Costa PCG, Wright EJ, Laskey KJ (2007) Probabilistic ontology for net-centric fusion. In: *10th international conference on information fusion, Quebec, 2007*
19. Kokar MM, Matheus CJ, Baclawski K, Letkowski JA, Hinman M, Salerno J (2004) Use cases for ontologies in information fusion. In: *Proceedings of the seventh international conference on information fusion*, pp 415–421

20. Kokar M, Matheus CJ, Baclawski K (2009) Ontology-based situation awareness. *Inform Fusion* 10:83–98
21. Kazakov M, Abdulrab H, Babkin E (2002) Ontology fusion approach for integration in heterogeneous distributed systems. *Engineering context-aware object-oriented systems and environments, ECOOSE*
22. Wache H, Voegele T, Visser U, Stuckenschmidt H, Schuster G, Neumann H, Huebner S (2001) Ontology-based integration of information—a survey of existing approaches. In: *Proceedings of IJCAI workshop on ontologies and information sharing*, pp 108–117
23. Boury-Brisset A-C (2003) Ontology-based approach for information fusion. In: *Proceedings of the 6th international conference on information fusion*, Cairns
24. Smart PR, Bahrami A, Braines D, McRae-Spencer D, Yuan J, Shadbolt NR (2007) Semantic technologies and enhanced situation awareness. In: *Paper presented at the 1st annual conference of the international technology alliance, ACITA, Maryland*
25. Gagnon M (2007) Ontology-based integration of data sources. In: *10th international conference on information fusion*, Quebec
26. Little EG, Rogova GL (2009) Designing ontologies for higher level fusion. *Inform Fusion* 10:70–82
27. Matheus C, Kokar M, Baclawski K, Letkowski J, Call C, Hinman M, Salerno J, Boulware D (2005) SAWA: an assistant for higher-level fusion and situation awareness. In: *Paper presented at the SPIE conference on multisensor, multisource information fusion*, Orlando
28. Zafeiropoulos A, Konstantinou N, Arkoulis S, Spanos D, Mitrou N (2008) A semantic-based architecture for sensor data fusion. In: *2nd international conference on mobile ubiquitous computing, systems, services and technologies*, 2008
29. Simperl E (2009) Reusing ontologies on the semantic web, a feasibility study. *Data Knowl Eng* 68(10):905–925
30. Hall DL (2001) *Handbook of multisensor data fusion*. CRC Press, Boca Raton
31. Bizer H, Berners-Lee T (2009) Linked data—the story so far. *Special issue on linked data. Int J Semantic Web Inform Sys* 53:1–22

Chapter 50

Development of GUI Based Test and Measurement Facilities for Studying Properties of MOS Devices in Clean Room Environment

Shaibal Saha and Supratic Chakraborty

Abstract This article describes a Graphics User Interface (GUI), meant for setting up an integrated Test and Measurement (T&M) facility, to study different electrical properties of Metal Oxide Semiconductor (MOS) devices is designed using Matlab 7.5.0 (R2007b). While developing the GUI, a probe station, connected to a Keithley Switch Matrix, and two other instruments namely, a Keithley Semiconductor Parameter Analyzer (SPA) and an Agilent manufactured Inductance-(L) Capacitance-(C) Resistance-(R) (LCR) Bridge, also connected to the switch matrix are considered. All the instruments are controlled over General Purpose Interface Bus (GPIB) protocol through a controller PC.

50.1 Introduction

Studies of different electrical parameters of Metal Oxide Semiconductor (MOS)-based devices, fabricated in a class-100 (Max 100 particles/ft³, size $\geq 0.5 \mu\text{m}$) clean room with a minimum feature size of $\sim 10 \mu\text{m}$ or less, requires a prober along with other characterization instruments namely, Semiconductor Parameter Analyzer (SPA), Inductance-(L) Capacitance-(C) Resistance-(R) (LCR) Meter. For better consistency of the result, it is necessary to place the device under test (DUT) inside the clean room and other instruments may be placed in an area of

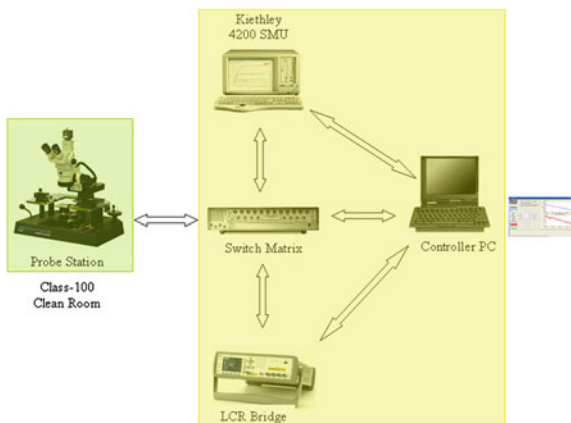
S. Saha (✉)

Applied Nuclear Physics Division, Saha Institute of Nuclear Physics,
Kolkata 700064, INDIA
e-mail: shaibal.saha@saha.ac.in

S. Chakraborty

Saha Institute of Nuclear Physics, Kolkata 700064, INDIA

Fig. 50.1 Instruments setup plan environment



lower cleanliness. A Switch Matrix is also employed for automatic switching between the above two measuring instruments. Usually test and measurement (T&M) instruments and their inbuilt Graphics User Interfaces (GUI) are not normally customized for a particular application relating to specialized frontier of research. On the other hand, researchers also require very powerful mathematical tools for processing the collected data either online or off line. So, an integrated test setup and remote control over a group of instruments is essential. Therefore, a software platform having three-fold facility is needed to fulfill all the three criteria (i.e. GUI development, General Purpose Interface Bus (GPIB) connectivity and powerful mathematical tool). Matlab 7.5.0 (R2007b) [1] is one of such unified software platform to support all the needs on a single controller PC. Figure 50.1 depicts the objective of this work in assigning a controller personal computer (PC) and connecting Signatone Probe Station with Agilent E4980A LCR Meter [2] and Keithley 4200 Source Monitor Unit (SMU), the SPA [3], through Keithley Switch Matrix [4]. In our customized set up, voltage vs. current (V–I) character of the DUT can be looked into with the help of Keithley 4200 SMU and capacitance versus voltage (C–V) character can be achieved with the help of Agilent E4980A LCR Meter using their respective inbuilt dc and ac voltage and frequency sources. The switch matrix is utilized particularly when a high-field stress measurement followed by a C–V measurement is required with a minimum loss of time. Therefore, after necessary switch connectivity between the passive Probe Station and the above two active T&M instruments, T&M on the DUT can be performed through controller PC. And surely GPIB communication delays (Data transfer rate >1 Mbytes/s, Short command ≈ 9 ms, Short query ≈ 18 ms, all the time depend on device) must not add to any detrimental effect on the data acquisition speed of those instruments. Moreover, each of the instruments has a fast buffer memory connected to dedicated high speed BUS. Upon necessary GPIB commands, the instruments start buffering the measured data at the highest transfer rate up to which the instruments can support by default.

Now, final part of the task falls on the GUI designer to introduce the user for taking over sole control and running the above mentioned experiment through a customized GUI on a computer monitor. The role of Human Computer Interaction (HCI) comes into picture now that includes physical, technical, physiological and psychological aspects to address through GUI design optimization. Designing & developing a GUI is an art from the designer's point of view and simultaneously a serious task of evolving a technical way out to involve users psychologically while using the GUI for experiments. On the aesthetic issue, "Look & Feel" attribute is the main point of consideration. This "Look & Feel" comprises of some aesthetics related elements viz. balance, equilibrium, symmetry, sequence, rhythm, as well as order and complexity [5, 6]. So, addressing aesthetics in a GUI is guided by some technical aspects of the elements of the GUI either in discrete or correlated manner. But the number of elements in a GUI depends on the very purpose of the GUI, its complexity and customization inputs by the users. Therefore, it is really very difficult to define a border line between aesthetics and technicality of any GUI, based on human computer interaction (HCI) and sometimes, both the problems overlap. As a GUI designer, it has been considered the optimal visual load to the user when populating the GUI with necessary elements according to runtime stages of the GUI [7, 8]. On the other hand, inattentional blindness of the user is also taken into consideration not only to reduce the possibilities of human error but also to take the advantages of it in using the GUI [8, 9]. Size, color, resolution, contrast, relative location of the elements, human field of visual perception, visual acuity, normal distance of human eyes from the computer monitor, eyeball movement pattern, human response time between seeing and mouse clicking etc. and above all attention of the user to a particular research problem also play an important role in designing the GUI [10–15]. It is always the point of foremost importance that additional technical features of the GUI must help creating an ambience to ease the way to reach the desired goal set in the GUI. Cares have been taken to avert issues of mind diversion to the users. During this design, the golden rule of measurement is considered. It states that in an ideal parametric measurement system, parameter must never feel that it is being measured. So, the GUI design target is to create purposefully such an ambience on the computer monitor which will perpetuate the user to concentrate on the research problem and not hindrance due to overload for unnecessary technicalities in the GUI. Following are the points to be considered to design a balanced GUI: (i) GUI should be 'look and feel' in nature, (ii) Easy to learn its operations by the users, (iii) Self guided to meet the user's expectation for subsequent stages, (iv) Metaphor based, (v) Fast and (vi) Algorithmically correct.

50.2 GUI Design

To reach the target designer should address the following points: (a) Purpose of the GUI, (b) User's Knowledge Base, (c) Mode of Output Presentation, (d) Size, Shape of the GUI, (e) Number of GUI Elements, (f) Inter Elemental Distance,

Orientation and Size, (g) Use of Text and Typography, (h) Runtime Priority of the GUI Elements, (i) Grouping of the GUI Elements, (j) Use of Color, (k) Elemental Boundaries, (l) Familiarity of the Metaphors, (m) Friendliness or Simplicity, (n) Transportability and (o) Usability.

50.2.1 Purpose of the GUI

The GUI will be used as a tool in scientific research. So, there should be certain amount of flexibility in data manipulation within the latitude of targeted research problem. In this particular GUI for Agilent E4980A LCR Meter, users enjoy the option to vary the inputs for spot or sweep measurements up to such limit that commensurate with the experimental conditions on the DUT and supported by the instrument. Provisions have been made available to the GUI to take multiple shots of data acquisition over wider sweep window with limited 201 (provision in the instrument) data points to have more significant results. Otherwise, cipher data will result at the far end of the data acquisition time window due to fast recombination process of the electron–hole pair in the DUT. Moreover, data *SAVE* button has been placed next to data *DISCARD* button on the GUI keeping in mind that in the research grade fabrication laboratory, the performance of a device is not as good as that in the production industry and many data sets are to be discarded.

50.2.2 User's Knowledge Base

This basis has been taken into account to symbolize the icons either with icon type, color or scientific textual message. Such as Radio buttons meant for selection, Push buttons for execution. English language has been chosen to combat with geographical barrier at this knowledge base [13].

50.2.3 Mode of Output Presentation

Output of the spot measurement result comprises of only two decimal numbers and thereby presentation in numeric form carries the full information to the user [16]. But, in case of sweep results in the form of either voltage or frequency, the set of output result comprises of $[201 \times 3]$ matrix. So, two line graph presentation (parameter values plotted against corresponding voltage or frequency) is accordingly designed as one of the best mode of information transfer.

50.2.4 Size and Shape of the GUI

Human enjoys almost 180° forward-facing horizontal field of view (FOV) for single eye and 120° for binocular vision. Remaining 30° of each peripheral vision on both sides has only monocular effect due to non overlap of images from both the eyes and it is of 100° in vertical direction. Considering 24 inches forward-facing horizontal viewing distance from the computer monitor, linear FOV (L) can be calculated using (50.1) [11, 14],

$$L = \frac{2\pi d}{360^\circ} \theta \quad (50.1)$$

(Where, d is the distance of monitor from eyes in feet)

Result from (50.1) shows that linear FOV of human vision forms a rectangular landscape mode. Hence, the shape of the GUI has been made rectangular landscape from the ergonomic point of view. Even, size of GUI for each instrument has been kept smaller than the full size of a 17 inch (standard) computer monitor to facilitate the user to open multiple windows simultaneously upon requirement.

50.2.5 Number of GUI Elements

To populate the GUI applet, the number of elements depends on experimental complexity, users demand, balance and adoption of graphics techniques by the designer. Active population of the GUI elements is directly related to the information processing capacity of the human brain [17]. It has been managed by grouping of similar elements. Lesser visual load also causes laser attention or more precisely partial attention thereby lack of seriousness on the subject [7–10]. So, optimum visual load is required in populating the GUI with elements, to stop mind diversion and thereby less inattentional error. To accomplish this, technique of runtime appearance/disappearance and activation/inactivation of some elements has been programmed. Inactive input fields could be wiped out to reduce visual overload but presence of some inactive fields maintain balance in elemental distribution on the GUI applet and help understanding the gross functioning of the GUI too.

50.2.6 Inter Elemental Distance, Orientation and Size

These three items are put into Fitts' law to calculate index of difficulty ($ID = \log_2 2A/W$) in one dimensional mouse pointer movement [20]. According to Fitts' law, the movement time (MT) required to select a target at a distance A (amplitude) and width of W in the movement direction is

Fig. 50.2 Agilent E4980A GUI flowchart

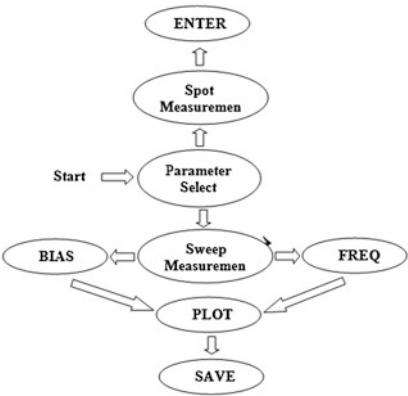
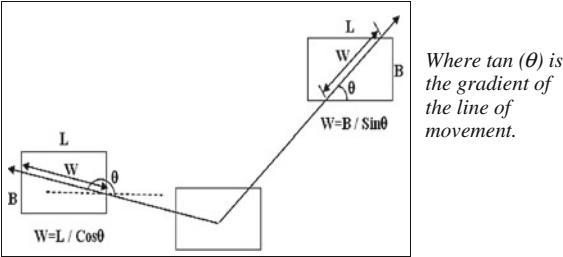


Fig. 50.3 Dependence of accuracy on target orientation



$$MT = a + b \cdot \log_2 \left(\frac{2A}{W} \right) \tag{50.2}$$

In (50.2), *a* and *b* are constants and values of which are determined through linear regression. *W* corresponds to the accuracy of the movement to reach the target point within overshoot and undershoot limits. There is only case of *Drop-down-menu* in selecting “*Parameter Select*” [21]. Except that, ID for all other movements has been calculated according to GUI flow chart for Agilent E980A LCR Meter and is shown in Fig. 50.2. To calculate the magnitude of *W*, in targeted mouse pointer movements, the projection of either length (*L*) or breadth (*B*) of the rectangular icons on the line of movement has been considered according to the relative position of the targeted GUI element which is evident from Fig. 50.3. In these cases, *W* is the function of *L*, *B*, and θ .

Therefore, in a particular relative position, orientation of the rectangular target elements plays a vital role in determining the ID of the movement. In case of square shaped target, the effect neutralizes partially and circular shaped target is the best fit for fixed *W*, where *W* is the diameter of the circle. It is always felt comfortable to see a text oriented horizontally (exception in some East Asian script) rather than vertically [14]. So, most of the elements are designed in rectangular landscape mode. In addition to the Fitts’ law, another important point of consideration that guided the arrangement of elements on the GUI is the golden

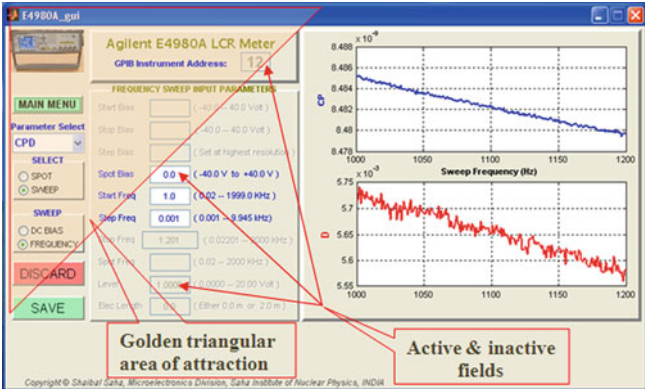
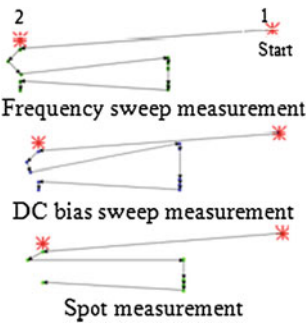


Fig. 50.4 Agilent E4980A LCR meter GUI

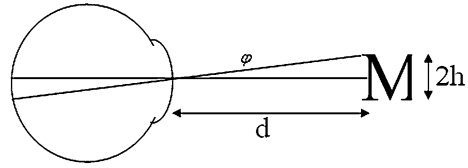
Fig. 50.5 Showing mouse pointing routes on the GUI along with caution locations marked with *



triangular position of attraction on any view [23]. So, there is a tradeoff between Fitts' law and golden location of attraction as shown in Fig. 50.4. For example, according to Fig. 50.5, *GPB-address* input field, an important icon, has been placed at a distant place from the remaining routes of the mouse pointer making ID greater compared to ID for next targeted movement. Even that, it is placed at top position on the GUI applet which is a location of visual attraction. To select another instrument, a push button named *MAIN MENU* has been provided. Obviously, it is also placed at an important but fairly distant location from *ENTER-PLOT-SAVE* or *DISCARD* button pairs to avert any inattentive cursor click on it.

Figure 50.5 shows, every occasion mouse pointer starts from 1st caution location (marked with *). Then it passes close to 2nd caution location (*MAIN MENU*) for reasons, if user thinks another GUI to attempt now according to Fig. 50.4. Otherwise, if the user thinks to continue with current GUI, then onwards for every subsequent pointer movement, both the caution locations go farther making ID greater. Once GPIB address is fixed, from 2nd iteration, these two caution icons will be inactive and mouse pointer can be started from next GUI element, taking advantage of Fitts' law [17–20].

Fig. 50.6 Human visual acuity



50.2.7 Use of Text and Typography

Text and typography for this GUI has been set considering the human visual acuity [11, 14]. Visible size of the element depends on human visual acuity or “Snellen” acuity (After the name Dutch ophthalmologist Hermann Snellen, 1834–1908). Visual acuity is an indication of the clarity or clearness of one’s vision. It means, a person with 6/6 visual acuity (normal), can see those objects just clearly from a distance of 6 m (~ 19.685 ft), subtended an angle 5 min ($5'$) of arc on the optic lens, as shown in Fig. 50.6, results in $2h = 0.88$ mm for $d = 24$ inch where $2h$ is the height of the object and d is the distance of the object from the eye.

Therefore, if the height of any textual letter (object) falls below 0.88 mm in the GUI, it will go beyond the character identification capability of eyes from a distance of 24 inch, subject to the luminosity of that letter and pixel per inch, sufficient for visual perception. Moreover, color contrast and wave length of light also play vital role in human visual perception at limiting visual acuity. In the Agilent E4980A GUI, it is set well above the limit, keeping in mind the unambiguous clarity from a distance of 45–70 cm from the computer monitor. Moreover, all the fields are barred by the software not to accept any illogical input and in any such situation an alert text message (in red color with bigger font size for legible resolution) is displayed. So, font type, sizes and color of the input parameters have also been considered for clarity.

50.2.8 Runtime Priority of the Elements

The golden triangular location of attraction on a visual display unit or in any page lays at the upper left hand corner which I already shown in Fig. 50.4 [23]. The Agilent E4980A GUI has three operational mouse pointer routes as shown in Fig. 50.5. According to selection, GUI elements are clicked one after another en-route final destination. Considering this, elements are placed starting from top-left corner and associated data input fields are placed side by side and other necessary selection fields are placed in downward direction according to priority. Only those fields are either made available or active that are in need to run the instrument for selecting a particular mode through mutually exclusive radio-buttons. While a user follows one flowchart, the elements of other

flowcharts, which are not being used, made either inactive or removed to keep the GUI aesthetically balanced and making it free of error prone due to inattentional blindness [8–10]. This approach minimizes the visual load as well.

50.2.9 Grouping of the Elements

The GUI elements, closely associated, are placed on a sub panel within the GUI applet to make a group. So, grouping of elements enable users to find out the functionally closely related elements easily and thus making the GUI simple. It also makes a virtual sense of lesser number of elements on the GUI (optimum visual load). The technique of need based dynamic is also employed to make the number of GUI element and thus keeping it within the limit of information processing capacity of the human brain at any time [17].

50.2.10 Use of Color

Use of color, adding a dimension to the GUI facilitates, a designer also considers in many ways to suite the GUI for the users. Suitable color soothes the vision, enhances contrast sensitivity of individual GUI elements, defines elemental boundary more prominently on the GUI applet. Red color is used to alert user by convention while blue enhances textual resolution and images at lower visual acuity for shorter wavelength. Though cones-cell (retinal cells) efficiency for color vision to all red, green and blue are same, luminous efficiency of human eye indicates a peak value at 560 nm [12, 14, 15]. On the other hand, when drawing attention is the main issue and not character recognition, same brilliant red color is used to indicate the run time state of the GUI. Here, the shape of red/green indicator is immaterial than its appearance.

50.2.11 Elemental Boundaries

Distinct elemental boundary defines the number of element unambiguously [5, 6]. More similar elements, placed on a panel, seem to be a single element if 3D effect is added with proper shade to the panel. This effect virtually minimizes element number and enables the user's information processing better with optimal visual load. This also minimizes chances of error due to inattentional mouse clicking on the GUI elements [19]. These advantages are also used in designing the GUI by adding suitable color, 3D effect, optimum separation between the elements etc.

50.2.12 Familiarity of the Metaphors

Familiar iconic-elements have been used to put an end to the language and geographic boundary limits on the GUI [13].

50.2.13 Friendliness or Simplicity

These two, complementary in nature, attribute to make the GUI look and feel. It also adds ‘acceptance’ attribute to the GUI. This along with familiarity of the metaphors made the GUI operation self explanatory. It also helps in cutting down the training time on actual instruments. In this GUI, metaphors or nomenclatures have been chosen in most indicative way to perceive the ongoing back-end processes by the user with pre occupation. Necessary additional delays have been introduced in different stages of the software to set the appearance and disappearance time of the GUI elements for synchronizing with either processing time or the response time of the users. It makes a sense of reality on the users’ mind even for a virtual instrumentation. Flowcharts of the GUI have been designed considering the mind set of user for most expected next step towards the targeted result [22]. This helps the user in keeping full attention in research experiment and never requires memorizing successive steps on the GUI. As a GUI designer, adequate self restraints have been imposed in using unnecessary techniques causing the attention diversion of the user from his/her research problems.

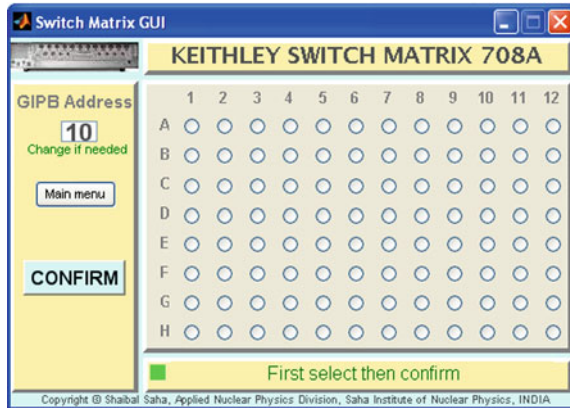
50.2.14 Transportability

All the above GUIs have been put under one single GUI named *Selection GUI* and a standalone Matlab executable file has been created on Windows XP platform. Now, user enjoys the option of opening all or any GUIs one after one in different windows by selecting the desired one through *Selection GUI*. To do this, every time user has to route through *Selection GUI*. This enables user to use any one of the GUIs independently in other T&M setup without being interfered by other GUIs kept in GUI repository. This facilitates designer to add many different instrument controlling GUI to the repository just making it versatile without losing identity of individual GUI and universal portability on Windows XP platform.

50.2.15 Usability

Usability is one of the most important aspects of the GUI design. As it is a supportive work for the main scientific experimentation, the result of the

Fig. 50.7 GUI for Keithley switch matrix



experiment depends on the correctness of the algorithm running behind the GUI. And *Familiarity*, *Friendliness or Simplicity* and *Transportability* add to the value on the *usability* for the GUI. Thus, usability overcomes the geographical, language and cultural barrier in using the GUI.

50.2.16 GUI for Keithley 708A Switch Matrix

Considering Fig. 50.1, the Keithley switch matrix is required for physical connectivity between passive Signatone Probe Station and either Agilent E4980A LCR Meter or Keithley 4200 SMU. A GUI has designed to control the switch matrix over GPIB protocol from the controller PC, as shown in Fig 50.7.

All the above GUI design considerations, mentioned in Agilent E4980A LCR Meter GUI design, have also been utilized in designing the GUI for Keithley 708A Switch Matrix. A special care has been taken to control the switching over GPIB command while instrument connectivity is changed from one to another. GPIB ensures first disconnecting the already connected instrument by passing initializing command every occasion to the Switch Matrix and then connect command for other instrument. Thus avoids fatal consequence for instrument and DUT.

50.2.17 Main Selection GUI

Figure 50.8 shows the initial entry point of all GUIs for instrument control. It has a drop down menu to select the particular instrument to control & use through GPIB and respective GUI. This opening GUI serves the very purpose of maintaining visual load to the user by facilitating accommodation of all individual instrument GUIs along with their elements and filtering as well with selectivity [7–10, 17].

Fig. 50.8 Main selection GUI



50.3 Safety Features of the GUI

Following safety features have been embedded deliberately in the programming. In using this GUI, obviously researchers should need some latitude in manipulating input data within the range of the instruments to address the research problem. So, researchers may come across inattentional error in data input. Moreover, sometimes instruments are also prone to hang over junk input data. These should be averted.

50.3.1 GPIB Address Input Field

For each instrument GUI, input to this field is made changeable. By default, it assumes the previous value. User enjoys the option to change it once, prior to first run of the instrument and thereafter, the field becomes inactive during the entire session and inhibits inattentional catastrophe. But, this facilitates the user to use the GUI in another experimental setup assigning different GPIB address.

50.3.2 Setup Selection Fields

Appearance and disappearance of these fields are programmed dynamically according to mode of selection. This takes care of optimal visual load and eliminates possible inattentional error.

50.3.3 Data Input Fields

Appearance and disappearance of these fields too are programmed dynamically according to selection mode. Initially, it always appears in spot measurement mode by default. Moreover, input fields are protected from any illogical or out of range data or beyond resolution data insertion. In such cases, the GUI program takes appropriate measures of its own and interactively alerts the user for such change. This also eliminates the possibilities of inattentional error.

50.4 Conclusion

In this GUI, necessary provisions are made available which can help a researcher to perform an accurate measurement and data acquisition on DUT in quick succession. Though the basic platform is Matlab 7.5.0 (R2007b), creating standalone Matlab executable file facilitates others to use the GUI without Matlab support. Above all, novel concept of multiple-GUI repository opens the provision of inclusion of a number of instrument control GUI under the main selection GUI considering the limitation of GPIB connectivity of instruments.

Acknowledgments Authors wish to acknowledge Subhajit Karmakar for their continuous inspiration and extending help in learning Matlab 7.5.0 (R2007b) and GPIB protocol.

References

1. Ref. manual of Matlab 7.5.0 (R2007b)
2. <http://cp.literature.agilent.com/litweb/pdf/5989-4435EN.pdf>
3. www.keithley.com/data?asset=4527—United States
4. www.keithley.com/data?asset=424—United States
5. Jansen BJ (1998) The graphical user interface: an introduction. SIGCHI Bull 30(2):22–26
6. Mishra U (2007) Inventions on GUI aesthetics. Available at SSRN: <http://ssrn.com/abstract=1264690>
7. Lavie N, Hirst A, de Fockert JW, Viding E (2004) Load theory of selective attention and cognitive control. J Exp Psychol Gen 133(3):339–354
8. de Fockert J, Bremner A (2011) Release of inattention blindness by high working memory load: Elucidating the relationship between working memory and selective attention. Cognition 121:400–408
9. Moore CM (2001) Inattention blindness: “perception or memory and what does it matter? Psyche 7(2)
10. Chun MM, Wolfe JM (2000) Visual attention, Chapter 9. In: Goldstein EB (ed) Blackwell handbook of perception, Oxford, Blackwell, UK
11. Proulx MJ (2010) Size matters: large objects capture attention in visual search. PLoS One 5(12):e15293
12. Irwin DE, Colcombe AM, Kramer AF, Hans S (2000) Attentional and oculomotor capture by onset, luminance and color singletons. Vision Res 40(10–12):1443–1454
13. Kitajima M, Polson PG (1995) A comprehension based model of correct performance and errors in skilled, display-based, human-computer interaction. Int J Hum-Comp Stud 43(1):65–99
14. Montgomery M Anatomy, physiology & pathology of the human eye. http://www.tedmontgomery.com/the_eye/index.html
15. Mullen KT (1985) The contrast sensitivity of human color vision to red-green and blue-yellow chromatic gratings. J Physiol 359:381–400
16. Artail HA (2003) Data and presentation techniques for fast, simple and automatic plotting. Comp Stand Interf 25(2):195–210
17. Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63:81–97
18. Green M (1992) Visual search, visual streams and visual architectures. Percept Psychophys 50:388–403

19. Bressan P, Pizzighello S (2008) The attentional cost of inattention blindness. *Cognition* 106:370–383
20. Fitts PM (1954) The information of the human motor system in controlling the amplitude of movement. *J Exp Psychol* 47(6):381–391
21. MacKenzie IS, Buxton W (1992) Extending Fitts' law to two-dimensional tasks. In: *Proceedings of the ACM conference on human factors in computing systems—CHI '92*, pp 219–226
22. Summerfield C, Egner T (2009) Expectation (and attention) in visual cognition. *Trends Cogn Sci* 13(9):403–409 (Epub 2009 Aug 27)
23. [Eyetoools 2008] Eyetoools, Eyetoools research and reports: Eyetoools, enquire, and did-it uncover search's golden triangle (2008) http://www.eyetoools.com/inpage/research_google_eyetracking_heatmap.htm (retrieved 2009-06-15)

Chapter 51

Prediction of Failure Risk Through Logical Decision Trees in Web Service Compositions

Byron Portilla-Rosero, Jaime A. Guzmán and Giner Alor-Hernández

Abstract In a service composition, the Quality of Services can be useful to identify those hidden data for a traditional composition; they can be a decisive factor for determining the behavior of future compositions since they allow evaluating risks resulting from reasons totally dependent on both the service environment and/or the composition system. Importance of this data is reflected on the way they are obtained, estimated, and applied to a composition. This paper has specifically studied the following three characteristics: availability, reactivity of services in periods of time, and management of beliefs to determine influence of services composition and to determine failure risk in such a composition through machine learning.

51.1 Introduction

Web services composition is a complex process which includes a careful analysis of requirements, semantics, and behavior of existing services, service verification, adaptation, contracting, and its performance. This process due to its complexity is a

B. Portilla-Rosero (✉) · J. A. Guzmán
School of Systems, Universidad Nacional de Colombia, Medellín, Colombia
e-mail: beportillar@unalmed.edu.co

J. A. Guzmán
e-mail: jaguzman@unalmed.edu.co

G. Alor-Hernández
Division of Research and Postgraduate Studies, Instituto Tecnológico
de Orizaba, Orizaba, Mexico
e-mail: galor@itorizaba.edu.mx

context where there are failure risks. An example of this is the moment when services used within this process result in unexpected failures caused by the service environment, such as: being available at the moment it is required or complying with its objective in an already time limit; as well as the way how the composer conducts instantiation processes in function of service data. For this purpose, one of the main components to obtain new information on both services and composer is the QoS; the QoS acquire additional information of those services which are not generally obtained with traditional composition processes and such information allows conducting a multi-dimensional analysis of the service behavior. This information allows making a statistical follow-up of the service behavior with respect to the environment in which it has been executed and of the composition environment where its participation is required. Importance of using these criteria is having an additional evaluation aspect which allows making a selection of more accurate Web services due to the evaluation of characteristics not previously taken into account during the composition process. In this way, modeling of compositions is possible and being conscious of the future failure risks which can occur during the composition is also possible. This evaluation allows identifying more appropriate services in order to meet requirements demanded for a composition plan and identifying alternative services for such requirements as well.

One of the tools used for describing behaviors is the machine learning from information resulting from the execution processes; behavior of services in function of QoS can be predicted. This allows making a permanent follow-up of service behavior and determining its influence within a composition.

For this specific case, logical decision trees have been used; these trees allow making an interpretation analysis, just as described in logical representations easily analyzed by the machine and by a human expert, in such a way that there can be a feedback during the composition processes in a way closer to reality.

This paper has been organized as follows: [Sect. 51.2](#) shows QoS evaluated to predict the risk factor in each service and the way how it has been estimated, and to analyze risks; [Sects. 51.3](#) and [51.4](#) show the model for predicting risk factors in the service composition; [Sect. 51.5](#) shows results obtained; [Sect. 51.6](#) shows related works; and [Sect. 51.7](#) shows both conclusions and future works.

51.2 QoS Representation and Failure Risk

In a Web dynamic environment there are many services to conduct similar functions and they show not very remarkable differences at first sight. However, when employing a set of non-functional characteristics of services such as the QoS, there is a possibility to widen this difference in order to determine which the best service is with respect to others to comply with required conditions.

51.2.1 QoS Attributes

This work encompasses three $c_k \in C$ criteria to determine the risk factor in service compositions, availability, reactivity, management of beliefs. Although the first two criteria have been used in other researches [1–3], this work includes a temporality characteristic in order to be more accurate with the analysis conducted to services.

Availability $A(s_i)$: Given a s_i service, availability is a frequency of the number of Web service successes when responding to a petition and the number of times this service has been requested for responding to such a petition in a p composition plan, bearing in mind a specific period of the day, estimated in d days and h hours, as in

$$A(s_i) = Frec(s_{success}) \in P | execution(s_i) \rightarrow d, h. \quad (51.1)$$

Reactivity $R(s_i)$: Given a s_i service, service reactivity is used to determine how appropriate using a Web service is under execution time restrictions ($t < T$), bearing in mind time periods classified by d days and h hours in which service behavior is seen affected and the p composition plan for which it was required, as in

$$R(s_i) = \{Frec(s_{success}) \in P\} t \leq T | execution(s_i) \rightarrow d, ht \in T. \quad (51.2)$$

Management of beliefs $MB(s_i)$: Given a s_i service, management of beliefs is used to determine the range of more feasible instances to be used by the composer in a service instantiation process. Instances are then defined as service inputs and outputs I_α, O_c . This factor takes those instances which have been the result of a successful execution of the service I_R, O_R as in

$$MB(s_i) = \{I_C, O_C \equiv I_R, O_R\} \in P | Frec(O_R) \cong 100 \%. \quad (51.3)$$

51.2.2 Risk and Analysis of Failure Probability

Risk is defined as one event or a set of events focused on the wrong assumption of the composition mechanisms during a service instantiation, service availability, and service reactivity, bearing in mind execution time restrictions which may alter a service behavior, resulting in a negative impact which prevents meeting of objectives, and measured through association of each event values.

A failure is a quantifiable value product of a Web or composition system service inconsistency, as a result of a risk evaluation.

A failure risk, including availability and reactivity of services $risk(s)$ is estimated, as in

$$risk(s) = \omega_1 * (1 - A(s)) + \omega_2 * (1 - R(s)) \quad (51.4)$$

Where $\omega_1 + \omega_2 = 1$ represents the impact for each criterion estimated through reciprocal of the ranks (RR) [4], which allows allocation of weights to a set of criteria based on their importance.

This importance has generally been given by the final user, which entails a high degree of subjectivity in the final result when estimating the risk. This work, however, started by identifying the most representative criterion in the composition; therefore, the entropy was applied as a multi-criteria decision analysis technique with the purpose of resolving subjectivity when allocating impacts.

Identification of failure risks using management of beliefs is evaluated in those composers based on assumptions to reach a composition plan. In this case, matching between instances and instantiated services is intended to be decreased in order to reduce the error factor in a composer's assumption when comparing environment with real world.

Finally, the composition plan risk $Risk(\rho)$ is estimated, as in

$$Risk(p) = Min \left(\prod_{i=1}^n risk(s_i) \right) \quad (51.5)$$

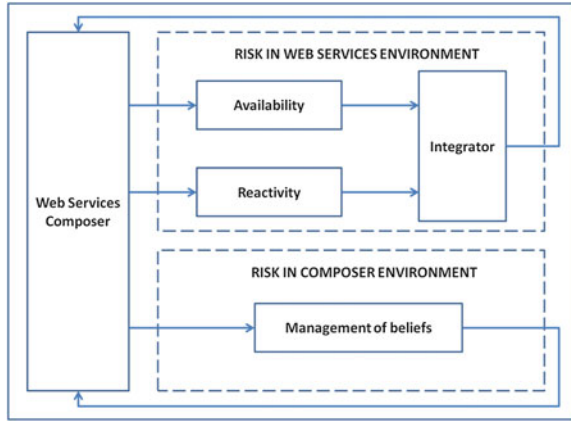
51.3 Inference of Values Associated to Failure Risk

The advantage of a permanent monitoring during service execution allows acquiring service behavior under multiple execution environments in an automatic way much closer to reality. This fact is very useful when we need to determine which services are to be used from a set of services and the most appropriate ones to better resolve a problem in specific cases. Therefore, a model based on logical decision trees is determined in order to automatically obtain service behavior and infer risk to which the composition is subject to with respect to selected services.

This model Fig. 51.1 includes two possible failure risks: Risks generated by the service environment, which is measured by availability and reactivity of the same; and composition environment risk as described by management of beliefs. These three criteria are treated on an independent basis, bearing in mind the following three specific steps:

51.3.1 Acquisition and Translation of the Execution Information

This includes service information collection from observations generated during the execution of such services. For each $c_k \in C$ criterion selected, a set of observations $O_j = \{o_1, o_2, o_3, \dots, o_n\}$ is generated, where each observation of the

Fig. 51.1 Learning model

set stores the $s \in S$ service which has been executed and the set of data specifically rated for such a service.

With respect to availability, the tuple representing observations is $(s_i, d, h, goal, cl)$ where s_i represents service executed; d is the day s_i when service was executed; h represents the time s_i when service was executed; $goal$ is the objective to be reached through the composition plan; cl determines whether the service is ready to be used; that is, if the service is not available at the time it has been called to execution, this should be deemed as a failure; otherwise, the service will be classified as successful.

The tuple representing service reactivity is $(s_i, te, d, h, goal, cl)$ where, s_i is the service to be executed; te is the time restriction for a service execution assigned by the customer; d is the day s_i when service was executed; h represents the time s_i when service was executed; $goal$ is the objective to be reached through the composition plan; cl is the classification assigned to a service when executed. This is associated to the time a service takes to be executed. Classification is successful if the service is executed during a restrictive period of time issued by the customer; otherwise, its execution should be deemed as a failure.

Finally, observations for management of beliefs are represented through the tuple $(s_i, inputs, outputs, cl)$ where s_i represents the service executed; $inputs$ is the input of the executed service; $outputs$ represents outputs of the executed service in function of inputs; cl represents the kind of relationship between supposed outputs by the composition system and outputs issued by the service after its execution. Kinds of observations are: success when there is a matching between supposed outputs by the composer and real acquired outputs of the real world, and failure when there is not such a matching.

Therefore, each time service s_i is executed, the list of observations O_j is increased $O_j = \{o_1, o_2, o_3, \dots, o_n\}$ thus obtaining the information about the behavior of the service evaluated under a specific execution environment, as in

$$\forall s_i \in S | \exists c_k \rightarrow O_j \{o_1, o_2, o_3, \dots, o_n\} i = 1, 2, \dots, n, k = 1, 2, 3 \quad (51.6)$$

Fig. 51.2 Learning algorithm

```

1. procedure buildtree( $T, \varepsilon, Q$ )
2.   if  $E$  is homogeneous enough then
3.      $K := \text{most\_frequent\_class } \varepsilon$ 
4.      $T := \text{leaf}(k)$ 
5.   else
6.      $Q_b := \text{best element of } p(Q), \text{ according to some heuristic}$ 
7.      $T.\text{test} := C' \text{ where } Q_b = \leftarrow Q, C$ 
8.      $\varepsilon_1 := \{E \in \varepsilon \mid |E \cup \beta| = Q_b\}$ 
9.      $\varepsilon_2 := \{E \in \varepsilon \mid |E \cup \beta| \neq Q_b\}$ 
10.    buildtree( $T.\text{left}, E_1, Q'$ )
11.    buildtree( $T.\text{right}, E_2, Q'$ )

```

Later, each observation is interpreted for each evaluated $c_k \in C$ criterion and is translated to a learning representation L language in which the knowledge base kb and the target function τ are identified; that is, what is desired to be learnt from the knowledge base kb . With respect to the target function, learning of the execution behavior of services through its cl classes is considered, defining whether or not the execution was successful. On the other hand, the knowledge base is generated from remaining data of observations for each c_k criterion.

51.3.2 Generation of the Control Knowledge

Generation of the control knowledge deals with the fact that a logical decision tree T is constructed for each s_i service; where each node contains necessary conditions under which each c_k criterion is applied and leaf nodes of the T contain prediction values for each c_k criterion. Probability estimation of rules is obtained from the frequency of learning examples.

Construction of the tree is conducted through the algorithms for induction of logical decision trees [5], as shown in Fig. 51.2.

After applying the learning algorithm, a set of logical rules $R\{r_1, r_2, r_3, \dots, r_n\}$ representing each s_i service behavior is obtained, bearing in mind the c_k evaluation criterion, as follows:

Availability: $\forall r_j \exists d \cup h \cup \text{goal}$

Reactivity: $\forall r_j \exists d \cup h \cup \text{te} \cup \text{goal}$

Management of beliefs: $\forall r_j \exists \text{inputs} \cup \text{outputs} \cup \text{goal}$

Finally, the set of rules generated becomes the control knowledge representing the behavior of services making part of a composition.

51.4 Integration of the Control Knowledge

With respect to availability and reactivity criteria, predicted data are associated to the service as a probabilistic value representing the service risk. These criteria share the fact that their value is dependent on day and hour when service has been executed, and are intended to improve composition while minimizing failure risks.

Fig. 51.3 Criterion-availability learning error



Estimation of the management of beliefs is made by the composer who is suggested with the predicted relations in order to control the instantiation process and improve the composer’s beliefs with respect to the real world.

51.5 Experimentation and Evaluation

The experiment encompassed the evaluation of the learning model through a controlled environment which simulates the behavior of ten Web services and correctness of the model was defined through the mean squared error. This behavior has been defined under a normal distribution where the mean is the real behavior of the service.

Each service shows a specific behavior; hence, a study case where a Web service is taken is shown and a set of 141 observations is taken, to identify the learning error decrease from the number of observations. Figure 51.3 show results of Service 1 evaluation with respect to the availability criteria and Fig. 51.4 show results of Service 1 evaluation with respect to the reactivity criteria.

Figures 51.3 and 51.4 show that the learning model is intended to have an approach to the real service behavior in order to determine which is the set of services decreasing the failure risk in a composition plan, as shown in Fig. 51.5.

Figure 51.5 shows the evaluation of the learning model which assesses the number of services that are used in composition plan. The assessment identified the decline of failed services when using the learning model and therefore shows that the use of the model aids in reducing the failure risk in the composition.

Evaluation of the management of beliefs allowed seeing how a composition system based on beliefs can improve the instantiation process and decrease the composer’s failure risk when defining the instances showing a better behavior according to a set of inputs and their relationship to the real world, as shown in Fig. 51.6.

The learning model error related to composition plan is showed in Fig. 51.7, while increase the number of observations by each service, the control knowledge allows to describe better the real behavior of services in the world, so the error in composition plan decreases because our model uses the services that is less likely to fail i.e. the composer does not select those services with high risk of fail.

Fig. 51.4 Criterion—reactivity learning error

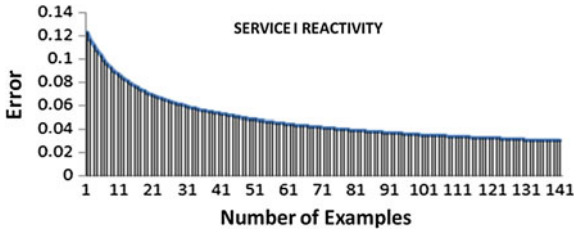


Fig. 51.5 Selection of services in a composition plan

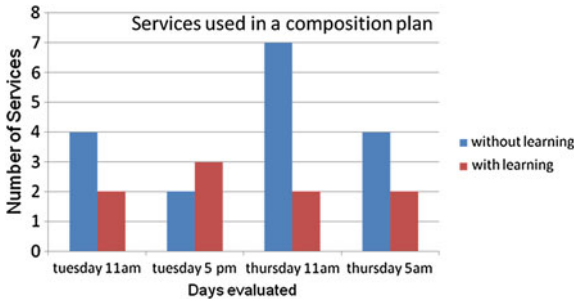


Fig. 51.6 Criterion—management of beliefs learning error

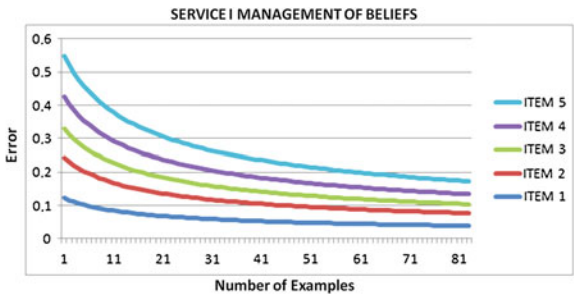
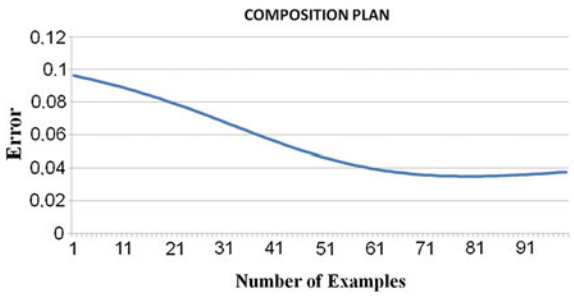


Fig. 51.7 Composition plan—learning error



51.6 Related Works

One of the most important researches is [6, 7] which establishes an approach for selecting services; this approach is intended to minimize impact of atomic web service failures through QoS. Risk is measured through the service failure probability and the result of its impact. Failure probability is specifically known because it uses a loss function in which the service execution cost is evaluated, as well as the message transfer time. Execution variations are determining factors for reaching a risk analysis. In this work, the execution determines the risk impact, but variation of executions during periods of time represented in hours and days is not taken into account, so the system is able to make a wrong decision when a service is requested in a specific period in time.

Reference [8] establishes two types of risks during the execution: the first risk verifies that the service execution is successful; if so, the user qualifies that service with a compensation for its good performance; the second risk measures the contrary; that is, when the service is not executed. In this work, allocation of quality values by the user can become a very subjective factor and the real behavior of services would not be taken into account.

Reference [3] is focused on the QoS management in order to improve the service composition; the work describes a model which allows predicting quality of services involved in the workflows. Management of risks in this work is focused on the service reliability in order to analyze software errors and the relationship among the times service was not executed. In this work, periodical data of execution and service analysis are not considered as additional information to determine the times services require to be executed and to achieve a better specification of the service behavior.

Reference [9] tackles the analysis of failures through the use of fault taxonomy [10] which includes three important failures: Physical failures, development failures, and interaction failures among services. This analysis is intended to omit those services deemed as a failure producer and those services are not part of the composition process. However, monitoring does not analyze service temporality; that is, to bear in mind service behavior in periods of time (days and hours) with the purpose of obtaining more information for conducting a more detailed failure analysis.

Risk management analyzed in [11] shows an approach to minimize losses associated to computational costs and communication costs, caused by failures in service execution during the composition. This failure is calculated through the impact of the failure risk. For this purpose, dependency of failures among services (that is, if a service is dependent on another one, the dependent service should be canceled or compensated) and compensation dependency (that is, if a service is dependent on another one, it should be canceled) are analyzed. This work evaluates the failure risk present without evaluating the moment in which or the circumstances under which the failure occurs; this would provide more accuracy when assigning value to estimate service failure risks.

Finally, in [12] establishes that risk is a service value associated by a user to the composition transactional properties in such a way that their addition can determine the service to be chosen during the composition process. However, selection of services would be including a subjective decision variable because the real behavior of the service would not be analyzed to make a decision, but a user's specific value.

51.7 Conclusions and Future Works

Through the analysis of failure risks in service compositions, verifying susceptibility in which Web services are involved is possible, and the way their behavior directly impacts the service composition. This work has shown a model based on logical decision trees which allow acquiring information related to QoS in an automatic manner and with more accuracy; QoS which evaluate the failure risk of both services and composer from the execution of services, thus identifying and determining the service behavior under a composition objective. This allows predicting both service and composer behaviors for acting in future compositions and leading the composer for a better service selection.

The importance of validating time periods associated to service execution was proven, since the failure risk factor is minimized in the composition, making emphasis on the most appropriate services to be used.

Bearing in mind the valuation of risks as future work, our intention is to use criteria which can measure the failure risk using additional values which make no part of the criterion but which can affect it, in order to improve expressivity. In relation to this work, temporality was used as an additional calculation value; however, it is important to use measures such as semantic impacts or composition objective meeting relations. In this way, values obtained through criteria tend to be more objective.

Acknowledgments This paper is supported by the project “programa de fortalecimiento del grupo de investigación Sistemas Inteligentes Web—SINTELWEB” quipu code 20201009532.

References

1. Ran S (2003) A model for Web services discovery with QoS. *ACM SIGecom Exch* 4:1–10
2. Zeng L et al (2004) QoS-aware middleware for Web services composition. *IEEE Trans Softw Eng* 30(5):311–327
3. Cardoso J, Sheth A, Miller J, Arnold J, Kochut K (2004) Quality of service for workflows and Web service processes. *J Web Semant* 5(3):319–338
4. Saeid M, Azim A, Ghani A, Selamat H (2011) Rank-order weighting of Web attributes for Website evaluation. *Int Arab J Inf Technol* 38:30–38

5. Blockeel H, De Raedt L (1998) Top-down induction of first order logical decision trees. *Artif Intell* 101(1–2):285–297
6. Kokash N, D’Andrea V (2007) Evaluating quality of Web services: a risk-driven approach. *Bus Inf Syst* 8(6):180–194
7. Kokash, N (2007) Risk management for service-oriented systems. Doctoral Consortium, Proceedings of the international conference on web engineering (ICWE), vol. 4607. LNCS, Springer, Como, pp 563–568
8. El Haddad J, Manouvrier M, Ramirez G, Rukoz M (2008) QoS-driven selection of Web services for transactional composition. *IEEE Int Conf Web Serv* 653–660
9. Chan M, Bishop J (2009) The design of a self-healing composition cycle for Web services. Seams, 2009 ICSE workshop on software engineering for adaptive and self-managing systems, pp 20–27
10. Chan M, Bishop J, Steyn J, Baresi L, Guinea S (2009) A fault taxonomy for Web service composition. In: *Service-oriented computing—ICSOC 2007 workshops*. Springer, Berlin, pp 363–337
11. Liu H, Zhang W, Ren K, Liu C, Zhang Z (2009) A risk-driven selection approach for transactional Web service composition. *Eighth international conference on grid and cooperative computing*, pp 391–397
12. Cardinale Y, El Haddad J, Manouvrier M, Rukoz M (2010) Web service selection for transactional composition. *International conference on computational science, ICCS 2010*, pp 2683–2692

Chapter 52

SEC-TEEN: A Secure Routing Protocol for Enhanced Efficiency in Wireless Sensor Networks

Alkore Alshalabi Ibrahim, Abu Khalil Tamer
and Abuzneid Abdelshakour

Abstract When Wireless sensor Networks are composed of a huge number of sensor nodes with limited energy resources, designing an energy efficient protocol will become a critical key issue that needs to be dealt with in order to expand the life span of the entire network. This paper proposes adding security features to the TEEN (Threshold sensitive Energy Efficient sensor Network) protocol by adopting the authentication part of the so-called SEC-LEACH (Secure Low Energy Adaptive Clustering Hierarchy) protocol. The paper then continues by showing how well this improvement goes by running experimental testing with multiple rounds. The new proposed SEC-TEEN improves the network security in and makes it more robust to external threats.

52.1 Introduction

Sensor networks are a future approach for a variety of applications, such as monitoring safety and security of buildings and spaces, measuring traffic flows, and tracking environmental pollutants and many other applications [1].

A. A. Ibrahim (✉) · A. K. Tamer · A. Abdelshakour
Department of Computer Science and Engineering, University of Bridgeport,
Bridgeport, CT 06604, USA
e-mail: ialkorea@bridgeport.edu

A. K. Tamer
e-mail: tabukhal@bridgeport.edu

A. Abdelshakour
e-mail: abuzneid@bridgeport.edu

Wireless sensors network (WSN) typically consists of a large number of wireless sensors that are able to communicate with each other using low power wireless data routing protocols.

A wireless sensor network (WSN) generally consists of a base station that can communicate with a number of wireless sensors via a radio link. Data are collected at the wireless sensor node, compressed, and transmitted to the head cluster or to the base station directly if required as in LEACH protocol. Many sensor networks have mission-critical tasks, so it is clear that security needs to be taken into account at design time [1]. The security part of any routing protocol is essential because it's important for most applications. WSNs lack physical protection and are usually deployed in open, unattended environments, which makes them vulnerable to attacks [2]. Cluster-based organization has been proposed for ad hoc networks in general and WSNs in particular [1]. Sensor networks are usually deployed with one or more base stations. A base station has resources like CU (Computation Unit), memory and life time energy. However, sensors are very primitive.

52.2 Hierarchical Protocol

Protocol constraints are more for WSN because of the design architecture of sensor nodes. This led researchers to design protocols to the nature of WSN. Haque et al. [3] describes some protocols used in SWN. One of the simplest routing protocols is the single-tier model. Single-tier model can cause the gateway or the base station to overload with the increase in sensors density and distant nodes will be out of energy quickly due to long range communication. Furthermore, WSN of thousand nodes will have to deal with frequent signal collisions and heavy packet losses [3].

For these reasons and more single-tier architecture is not scalable for larger set of sensors covering a wider area of interest since the sensors are typically not capable of long-distance communication. So, scalability is one of the major design attributes of sensor networks.

A hierarchical sensor network with multi hop model consisting of sensor nodes, cluster heads, and relay nodes. There are two types of sensor groups; a group of sensor nodes led by a cluster head, and the second type is a group of cluster heads with one cluster head as head of that group. Each sensor group collects data from a particular geographical area and sends the data to the nearest sensor nodes within the cluster. If the neighboring nodes are relay nodes, they forward those data using the appropriate routing path. Finally, the cluster head aggregates the data and forwards the aggregated data to its upper level cluster head [4].

52.2.1 LEACH Protocol

LEACH is Low-Energy Adaptive Clustering. Nodes organize themselves into local cluster with one node acting as the local base or cluster-head. LEACH uses distributed cluster formation. LEACH randomly selects few sensor nodes as cluster heads (CHs) and rotates this role to evenly distribute the energy load among the sensors in the network. The cluster head (CH) nodes compress data arriving from nodes that belong to the same cluster, after the data has been aggregated, aggregated packet sent to the base station in order to reduce the amount of information that must be transmitted to the base station.

The operation of LEACH consist of two phases: setup phase and steady State phase.

During the setup phase, the clusters are organized and CHs are selected. In the (steady state phase) the actual data transmit to the base station takes place. During the setup phase, a predetermined fraction of nodes, p , elect themselves as CHs as :

A sensor node chooses a random number, r , between 0 and 1. If this random number is less than a threshold value, $T(n)$, the node becomes a cluster-head for the current round.

The threshold value is calculated based on an equation that incorporates the desired percentage to become a cluster-head, the current round, and the set of nodes that have not been selected as a cluster-head in the last $(1/p)$ rounds, denoted by G . It is given by Eq. (52.1).

$$T(n) = \frac{p}{1 - p \left(r \bmod \left(\frac{1}{p} \right) \right)} \text{ if } n \in G \quad (52.1)$$

Where p is the desired percentage of cluster heads (e.g., $p = 0.05$, r = the current round, and G is the set of nodes that have not been selected as cluster-heads in the $1/p$ last rounds. Each elected CH broadcasts an advertisement message to the rest of the nodes in the network that they are the elected new cluster-heads. After the non-cluster head nodes receive this advertisement; they have to decide which cluster they want to join. This decision is based on the signal strength of the advertisement. The non cluster-head nodes inform the appropriate cluster-heads that they will be a member of the cluster. After receiving all the messages from the nodes that would like to be included in the cluster and based on the number of nodes in the cluster, the cluster-head node creates a TDMA schedule and assigns each node a time slot when it can transmit. This schedule is broadcasted to all the nodes in the cluster. During the steady state phase, the sensor nodes can begin sensing and transmitting data to the cluster-heads. The CH node, after receiving all the data, aggregates the data and sends it to the base-station. After a certain time,

which is determined by a priority; the network goes back into the setup phase again and enters another round of selecting new CH. Each cluster communicates using different CDMA codes to reduce interference from nodes belonging to other clusters [5, 6].

The nodes die randomly and dynamic clustering increases lifetime of the system. LEACH is completely distributed and requires no global knowledge of network [6].

52.2.2 TEEN Routing Protocol

TEEN is “Threshold sensitive Energy Efficient sensor Network protocol”.

TEEN is a reactive protocol proposed in [7] for time-critical applications. Nodes are arranged in hierarchical clustering scheme in which some nodes acts as 1st and 2nd level cluster heads. Like leach after forming the CH, it gets some attribute from the user. Once the attribute is received the CH broadcasts the attribute to its cluster nodes. Attribute are hard and soft thresholds for sensed attributes (values) to its cluster members. Hard threshold is the minimum possible value of an attribute to trigger a sensor node to switch on its transmitter and transmit to the cluster head. The sensor nodes start sensing and transmitting the sensed data when it exceeds hard threshold. Nodes transmit only when the sensed attribute is in the range of the target. By doing that we are reducing the number of transmissions significantly. The transmitted sensed value is stored in an internal variable called “Sensed Value” (SV). The cluster nodes again start sensing. Once a node senses value at or beyond the hard threshold, it transmits data only when the value of that attributes changes by an amount equal to or greater than the soft threshold. As a result from doing that, we are reducing the number of transmissions if there is little or no change in the value of sensed attribute. The energy is conserved since the sensor nodes in the cluster senses continuously but transmits only when the sensed value is above hard threshold. The soft threshold further reduces the transmission which could have been occurred when there is a little change (or) no change in sensed attribute. As the cluster heads need to perform extra computations it consumes more energy compared to other nodes. In order to evenly distribute the energy consumption each nodes in the cluster is given a chance to act as a cluster head for a fixed cluster period [8]. The attributes can also be changed during every cluster change time. The main drawback of this protocol is that the transmission from nodes to cluster head will not be there when the sensed value is not greater than Hard threshold, hence the cluster head will never come to know even when any one of the sensor node dies. Accurate and clear picture of the network can be obtained by fixing the soft threshold as smaller value even though it consumes more energy due to frequent transmissions. Each node in the cluster is assigned a transmission slot using TDMA schedule to transmit data to its cluster head. To avoid collisions during cluster heads communication with BS, CDMA schedule is used [8].

52.2.3 SEC-LEACH Routing Protocol

In SEC-LEACH, prior to network deployment, this protocol generates a large pool of S keys and their IDs. Each node is then assigned a ring of m keys drawn from the pool pseudorandomly, without replacement, as follows:

A pseudorandom function (PRF) used to generate its unique ID idX for each node, idX then used to seed a pseudorandom number generator (PRNG) of a large enough period to produce a sequence of m numbers. R_X , the set of key ids assigned to X , can then be obtained by mapping each number in the sequence to its correspondent value modulus. Also prior to deployment, for each node is assigned a pair wise key shared with the BS [2]. By applying these modifications, SEC-LEACH protocol works as follows:

During setup phase, (Step 1) when a self-elected CH broadcasts its ADV message, it includes the ID of the keys in its key ring and their nonce, non CH nodes now cluster around the closest CH with whom they share a key. (Step 2) The other sensors computes the set of CHs keys IDs and choose the nearest CH with whom they share a key, then they will send two things, the first thing is a join REQ message, protected by a MAC that produced by the share key, the second one is nonce that is broadcasted by CH to prevent reply attacks; as well as the ID of the key chosen to protect this link, so that the receiving CH knows which key to use to verify the MAC]. In third step the CHs send the time schedule (to finish setup phase) to the sensors chose to become their members [2, 9].

In the steady-state phase, (Step 4) node-to-CH communication is protected using the same key used to protect the join REQ message in Step 2. A value computed from the nonce and the reporting cycle is also included to prevent replay.

In fifth step the CHs can now decrypt the sensing reports they receive, perform data aggregation, protect them by symmetric key shared between the CH and the BS and send the aggregate result to the BS. A counter is included in the MAC value also, to provide freshness.

52.2.4 Security Goals

Since nodes choose a cluster-head based on received signal strength, attacker device with good signal can disable the entire network by using the HELLO flood attack to send a powerful advertisement to all nodes in the network. Due to the large signal strength of the advertisement, every node is likely to choose the attacker device as its CH. The attacker device can selectively forward data received, while the rest of the network is effectively disabled. The attacker device can use the same technique to mount a selective forwarding attack on the entire network using only a small number of nodes if the target number of cluster heads or the size of the network is sufficiently small. Simple countermeasures such as

refusing to use the same cluster-head in consecutive rounds or randomized selection of a cluster head (rather than strongest received signal strength) can easily be defeated by a Sybil attack [10].

52.3 Proposed Solution

52.3.1 *SEC-TEEN: A Secure Enhanced Efficiency Routing Protocol*

In our proposal, we used the security techniques given by Sec-LEACH to design our secure TEEN protocol. In other words, we propose to generate a large pool of S keys and their ids prior to network deployment. Each node is then assigned a ring of m keys drawn from the pool pseudo-randomly, without replacement.

The new TEEN algorithm can then be run with the following modifications: when a self-elected CH broadcasts its adv message, it includes the ids of the keys in its key ring; the remaining nodes now cluster around the closest CH with whom they share a key.

During the self-elected CH broadcasts its id and a nonce. Then each ordinary node in the network computes the set of key ids, chooses the closest CH with whom it shares a key, and sends it a join-req message, protected by a MAC. The MAC is generated, and includes the nonce from CH's broadcast to prevent replay attacks, as well as the id of the key chosen to protect this link (so that the receiving CH knows which key to use to verify the MAC). At the end of the setup phase, the CHs send time slot schedule to the nodes chosen to join their clusters [2]. In the steady-state phase, the cluster-head broadcasts to its members along with its adv message the following two values:

Hard Threshold (HT): This is a threshold value for the sensed attribute. It is the absolute value of the attribute beyond which, the node sensing this value must switch on its transmitter and report to its cluster head [7].

Soft Threshold (ST): This is a small change in the value of the sensed attribute which triggers the node to switch on its transmitter and transmit [7]. Nodes sense their environment continuously. The first time a parameter from the attribute set reaches its hard threshold value, the node switches on its transmitter and sends the sensed data (node-to-CH communication) using the same key used to protect the join-req message. A value computed from the nonce and the reporting cycle is also included to prevent replay. The sensed value is stored in an internal variable in the node, called the sensed value (SV) [2, 7]. The nodes will next transmit data in the current cluster period, only when both the following conditions are true:

1. The current value of the sensed attribute is greater than the hard threshold.
2. The current value of the sensed attribute differs from SV by an amount equal to or greater than the soft threshold.

Whenever a node transmits data, SV is set equal to the current value of the sensed attribute.

Thus, the hard threshold tries to reduce the number of transmissions by allowing the nodes to transmit only when the sensed attribute is in the range of interest. The soft threshold further reduces the number of transmissions by eliminating all the transmissions which might have otherwise occurred when there is little or no change in the sensed attribute once the hard threshold [7].

The CHs can now decrypt the sensing reports they receive, perform data aggregation, and send the aggregate result to the BS. The aggregate result is protected using the symmetric key shared between the CH and the BS [2].

52.4 Mathematical Model:

In this section, we discuss the performance of the protocols for a selected set of parameters. To evaluate the performance of our protocol, we have implemented it on the Matlab with the Sec-LEACH [9], as well as Sec-TEEN protocols. Our implementation is done on network size (100 sensors); 500 rounds were processed with the following initial values of main parameters [9]:

- The desired percentage of CHs (P) is set to 0.05.
- Each sensor starts with 0.5 J energy.
- The amplifier energy is assumed to be 100 pJ.
- The electronic energy is assumed to be 50 nJ.
- Each sensor data range is set to 30 m.
- Each node has a 2000-bit data packet to send to the BS.

52.5 Simulation Results

Sec-TEEN works with extra conditions for sending data at each node, such condition reduces the total transactions required in the network communications. This leads to highly reduced data overload compared with Sec-LEACH. This trend can be exhibited by analyzing the following graph;

Our simulation shows that for the network of size 100 nodes, both Sec-LEACH and Sec-TEEN have different volume of data being transmitted between nodes to cluster heads and vice versa.

Referring to Fig. 52.1, we find that both of the Sec-LEACH and Sec-TEEN produce different total data overload in the specified network. From this point, when it comes to our Sec-TEEN protocol, it remarkably involves fewer amounts of data sent between nodes in the system. It is noticed from Fig. 52.1 that after a

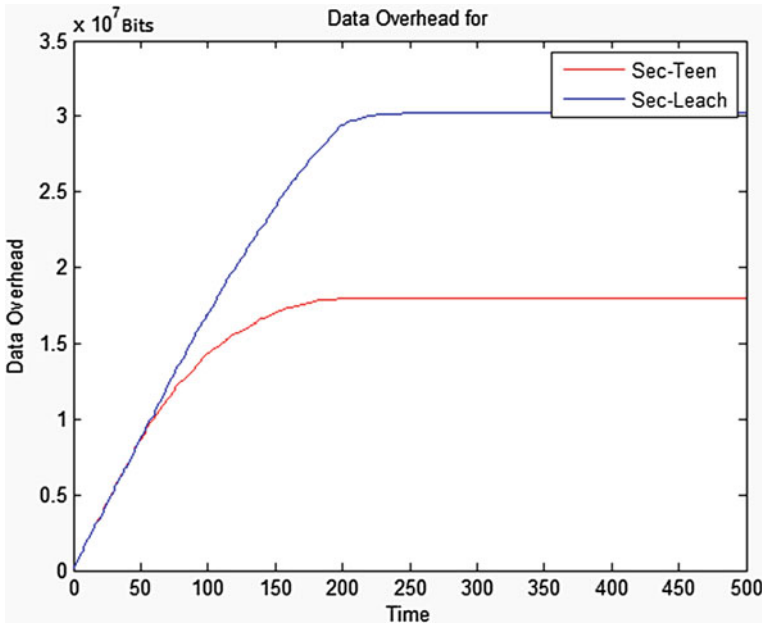


Fig. 52.1 Total data overload in Sec-LEACH and Sec-TEEN after 500 rounds for network size of 100 nodes

period of time both protocols will start having a steady phase in terms of data overhead and that obviously is due to the energy being completely consumed by all of the nodes. The highest data volume sent using the Sec-TEEN was 1.76×10^7 bits where it was 3.1×10^7 bits in the case of Sec-LEACH [9].

Referring to Fig. 52.2 below and after applying both algorithms on the network, we were able to achieve better results in terms of energy saving. When comparing the two protocols, we found that the Sec- TEEN performs much better than Sec-LEACH. Sec-TEEN has significant lower values of the energy consumed by the nodes than what is consumed by the Sec-LEACH. However, the two protocols have a remarkably considerable variation in such values in the case of energy saving.

Our experiment shows that the variation of energy consumption is very significant under the simulation network (i.e. 100 sensors). With a 0.5 J energy initially given to each node in both protocols, we can tell that nodes in the Sec-TEEN tend to consume their energy at a slower rate if compared to the Sec-leach nodes.

Consequently, as can be seen from the Fig. 52.2, the 100 nodes of the Sec-TEEN protocol consumed all of their energy at time 246 while they consumed all of their energy at time 224 for the Sec-LEACH. This variation comes from the

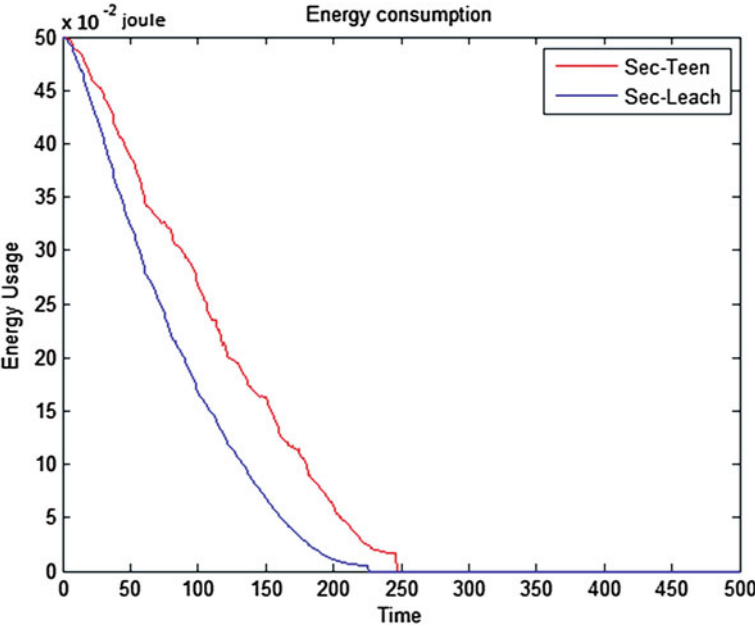


Fig. 52.2 Total energy consumption occurs in Sec-LEACH and Sec-TEEN after 500 rounds, for network size of 100 nodes

nature of how Sec-TEEN works. Using the hard as well as soft Threshold values impacted the performance of Sec-TEEN. Energy is saved when the sensed data do not belong to the range of desired values which saves the energy needed to send such data.

According to the energy saving analysis, we can easily figure out that the number of alive nodes that may appear in Sec-TEEN will be much higher than the number in Sec-LEACH. This can be depicted in Fig. 52.3, below; keeping in mind, the number of Alive Nodes depends on the energy consumption by the network.

52.6 Conclusion

In this paper we introduced Sec-TEEN which is a new improved version of the generally used technique namely LEACH. It performs better in terms of efficiency, energy dissipation and lifetime. The results show that wireless sensor networks can be made highly efficient with this protocol and also it is very flexible and can be

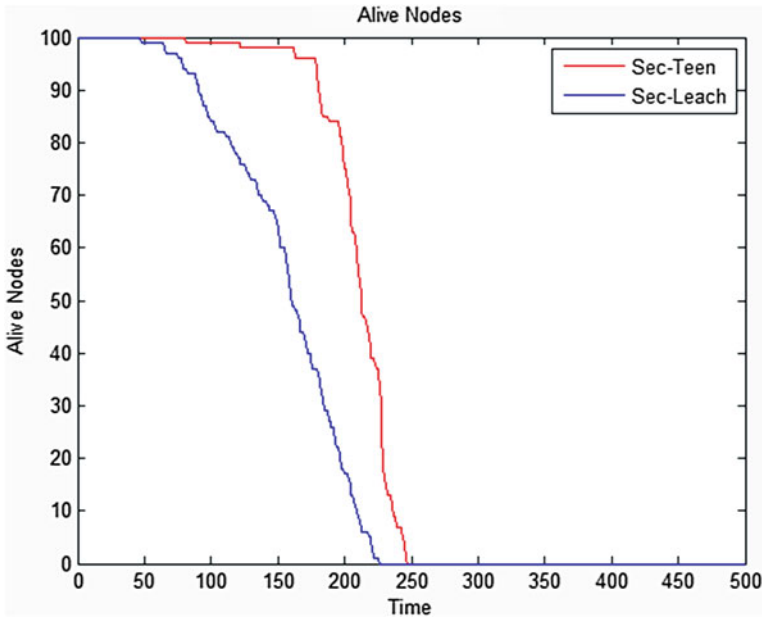


Fig. 52.3 Alive nodes occur in Sec-LEACH and Sec-TEEN after 500 rounds, for network size of 10 nodes

applied to other techniques that are related to the clustered hierarchy without affecting the main purpose of the original technique. Moreover the security level can be maintained using Sec-TEEN while providing better performance.

References

1. Shi E, Perrig A (2004) Designing secure sensor networks. *IEEE Wireless Commun* 11(6):38–43
2. Oliveira LB, Wong HC, Bern M, Dahab R, Loureiro AAF (2006) SecLEACH—A random key distribution solution for securing clustered sensor networks. In: Fifth IEEE international symposium on network computing and applications (NCA 2006), 24–26 July 2006, pp 145–154
3. Haque ME, Matsumoto N, Yoshida N (2009) Context-aware multilayer hierarchical protocol for wireless sensor network. In: Third international conference on sensor technologies and applications (SENSORCOMM '09), 18–23 July 2009, pp 277–283
4. Panja B, Madria SK, Bhargava B (2006) Energy and communication efficient group key management protocol for hierarchical sensor networks. In: IEEE international conference on sensor networks, ubiquitous, and trustworthy computing, 5–7 June 2006, p 8
5. Heinzelman WR, Chandrakasan A, Balakrishnan H (2000) Energy-efficient communication protocol for wireless microsensor networks. In: Proceedings of the 33rd annual Hawaii international conference on system sciences, p 10
6. Bilami A, Boubiche DE (2008) A hybrid energy aware routing algorithm for wireless sensor networks. In: IEEE symposium on computers and communications (ISCC 2008), 6–9 July 2008, pp 975–980

7. Manjeshwar A, Agrawal DP (2001) TEEN: a routing protocol for enhanced efficiency in wireless sensor networks. In: Proceedings 15th international parallel and distributed processing symposium, pp 2009–2015
8. Baghyalakshmi D, Ebenezer J, Satyamurthy SAV (2010) Low latency and energy efficient routing protocols for wireless sensor networks. In: International conference wireless communication and sensor computing (ICWCSC 2010)
9. Abuhelaleh Mohammed, Elleithy Khaled, Mismar Thabet (2009) Clustered Hierarchy in Sensor Networks: Performance and Security. *Int J Netw Secur Appl* 1(2):51–62
10. Karlof C, Wagner D (2003) Secure routing in wireless sensor networks: attacks and countermeasures. In: Proceedings of the first IEEE international workshop on sensor network protocols and applications

Chapter 53

An Integration of UML-B and Object-Z in Software Development Process

Mehrnaz Najafi and Hassan Haghighi

Abstract Visual and formal modeling notations can complement each other when developing software systems. Object-Z (OZ) is an object-oriented extension of the Z notation for writing formal specifications. Much work exists on translations between UML and OZ. However, UML is not a formal modeling language. This delays verification and validation of UML visual models until translation to OZ. On the other hand, UML-B is a UML-like formal modeling language that supports object-oriented modeling concepts. In this paper, we propose a formal mapping from UML-B models to OZ constructs in order to integrate these two object-oriented visual and non-visual formal notations. In this way, we assist the software development process by using UML-B as a visual modeling notation at early conceptual modeling stage and OZ at next stages when requirements are better understood. Also, an opportunity is provided to develop code from UML-B models using existing approaches for mapping OZ specifications to object-oriented programs. Finally, using UML-B instead of UML, we are able to verify visual models in the early conceptual modeling stage of the software development process without translating them into OZ specifications.

53.1 Introduction

Visual modeling provides an opportunity to develop specifications that are easy to understand [1]; hence, they can be used in early conceptual modeling stage of software development process. On the other hand, formal methods are used to

M. Najafi (✉) · H. Haghighi
Faculty of Electrical and Computer Engineering, Shahid Beheshti University,
Tehran, Iran
e-mail: M.Najafi@sbu.ac.ir

H. Haghighi
e-mail: h_haghighi@sbu.ac.ir

increase the reliability of software systems. However, most of the formal specification languages are difficult to deal with due to their mathematical basis; hence, this is not a good choice to use them in early stages of the software development process when requirements are not well understood. Please confirm the corresponding author is correctly identified and amend if necessary.

The most frequently adopted approach to overcome this deficiency is to define transformations between formal and visual models [1–11]. For instance, Meyer and Souquieres [2] proposed a systematic transformation from OMT diagrams to B specifications. McUmber and Cheng [3] presented a framework for formalizing class diagram and dynamic model using a metamodel based transformation between UML [12] and VHDL and also Promela/SPIN.

Also, many contributions [1, 4–9] have been done which propose informal/formal transformations between UML models and OZ constructs [13, 14]. However, the integration of UML and OZ suffers from the following drawbacks:

1. UML is not a formal modeling language, and also it is not supported by any verification and validation processes. This delays verification and validation of visual models until translation to OZ.
2. Informal transformations between UML and OZ are often described imprecisely at the model level [4].
3. It is difficult to verify informal transformations.

On the other hand, there are a number of approaches and tools [15, 16] in the literature which integrate UML-B [17, 18], a graphical formal modeling notation based on UML, and Event-B [19] which is a new variant of the B method. Although Event-B is a non-visual formal specification language, it is not object-oriented; hence, this formalism changes the structure of initial UML-B models.

In this paper, we rely on the approach described in [5] which gives a formal mapping from UML models to OZ specifications in order to integrate OZ and UML-B. More precisely, we propose a formal mapping from UML-B models to OZ constructs. We first give a formal definition of both UML-B and OZ meta-models in OZ. Then, we translate UML-B constructs to their OZ counterparts using mapping functions in OZ. In this way, we assist the software development process by using UML-B as a visual modeling notation in the early conceptual modeling stage and OZ in next stages.

To mention another contribution of our work, it is worth noting that OZ is a more powerful object oriented formal specification language in comparison to UML-B because it has powerful semantics and calculus, i.e., predicate calculus and set theory. Also, OZ strongly supports the notion of object, and its specification style corresponds to constructs of typical object-oriented programming languages very much; therefore, it is an acceptable decision to use UML-B models in the early stage of software development process and then interpret them into OZ when adding more details to specifications is needed.

Finally, using UML-B instead of UML, we are able to verify visual models in the early conceptual modeling stage of the software development process without

translating them into OZ specifications. This considerably reduces the total cost and time of the software development process.

Section 53.2 gives preliminaries of our work. In Sects. 53.3 and Sections 53.4, we present OZ metamodel and its formal description in OZ. Sections 53.5 and 53.6 give UML-B metamodel and its formal description in OZ. In Sect. 53.7, we give a formal mapping from UML-B models to OZ specifications. Section 53.8 includes a case study, and Sect. 53.9 concludes the paper.

53.2 Preliminaries

We should first review the work of Kim and Carrington [5] who proposed a formal mapping from UML models to OZ specifications at the metamodel level. They first proposed UML and OZ metamodels and their formal descriptions in OZ. Then, they described a formal mapping from UML models to OZ specifications using mapping functions in OZ.

As a brief overview of OZ, it is worth noting that class schema is the major new construct in OZ (in comparison to Z) which encapsulates a single state schema with relevant operations. The structure of the class schema in OZ can be found in [13]. The current version of UML-B has four kinds of diagrams: package, class, context and state machines diagrams. The class diagram is used to describe the behavioural part of a model [20]. Each class in a class diagram may have attributes, associations, events, state machines and invariants.

53.3 OZ Metamodel

Figure 53.1 shows part of OZ metamodel which mainly relates to class schema. In Fig. 53.1, `OZModelElement` is a top level metaclass from which all possible modeling elements in OZ can be drawn. Class schema is represented by `OZClass` which contains superclasses, a set of features (i.e. attributes and operations) and invariants. Another important metaclass is `OZSpecification`. In OZ, a specification is usually developed in a bottom-up approach. Once behaviors of individual objects and their interrelationships are modeled in terms of classes, the whole system is modeled by composing the developed individual classes from the system's point of view [5]. Syntactically, there is no denotation to distinguish the system class from other classes in OZ [5]. However, the intention of the system class differs from other classes. Each `OZSpecification` has a system class and a set of component classes. `OZFeature` is a metaclass from which all possible attributes and operations of classes can be drawn.

Figure 53.2 shows the structure of attributes and operations in class schemas that are represented by `OZAttribute` and `OZOperation` metaclasses, respectively. It is worth mentioning that we introduced metaclass “`OZType`” in OZ metamodel given in [1] in order to build these parts of OZ metamodel.

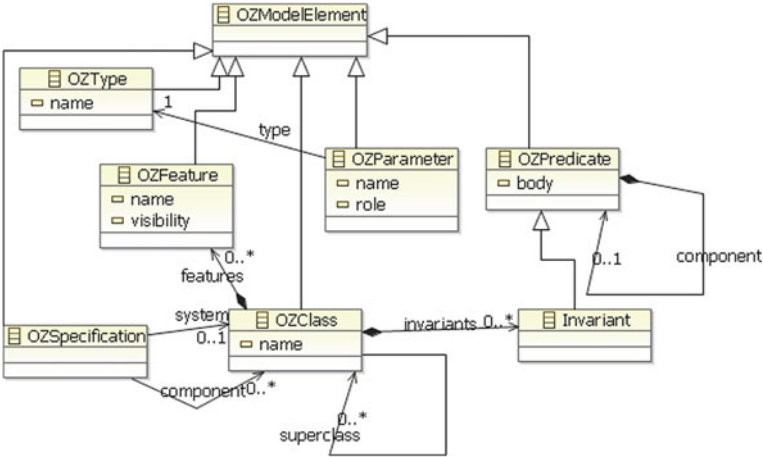
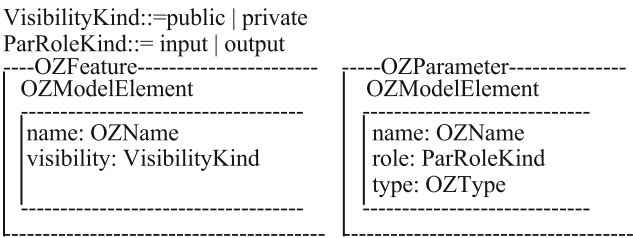


Fig. 53.1 Part of OZ metamodel (class schema)

53.4 Formal Description of Metamodel in OZ

We give a formal definition for the modeling constructs in the OZ metamodel using OZ classes. The OZ class OZFeature is a formal description of metaclass OZFeature. Each feature has a name and also visibility property determining its access type. The OZ class OZParameter models OZ parameters formally. Each parameter has a name, type and role.



The OZ class OZAttribute is a formal description of class attributes. This class inherits all of the features of OZFeature; see Fig. 53.2. OZOperation models class operations formally in the same way as OZAttribute

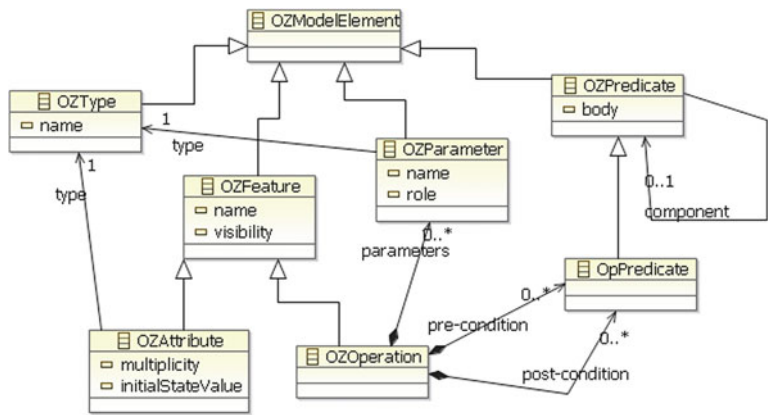
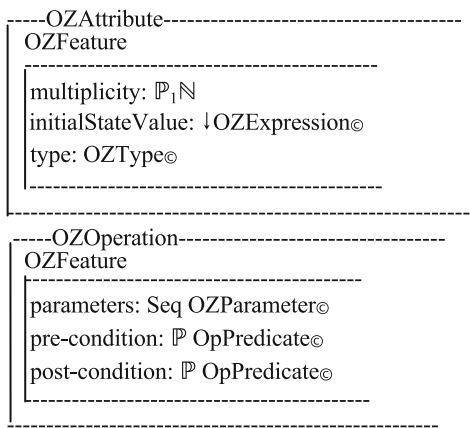
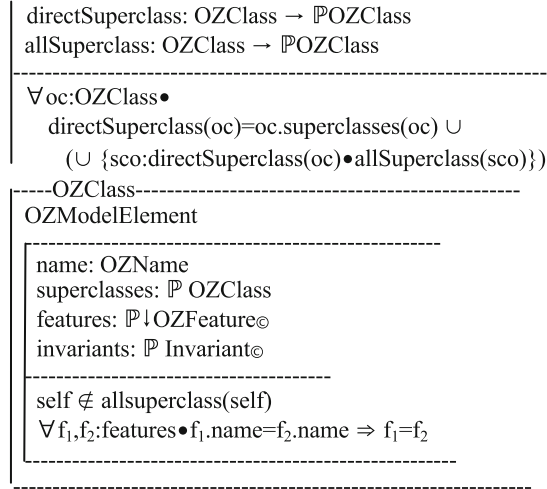


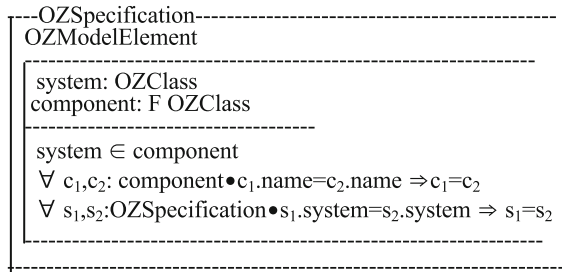
Fig. 53.2 Part of OZ metamodel (class schema operations and attributes)



To formalize OZ Class, we consider an OZ class which has name, features, superclasses and invariants. Attribute “name” maintains the name of the class. Each class has its own set of features (i.e. attributes and operations) and invariants. Any circular inheritance is not allowed [5]. Also, feature names should be unique within a class. These properties are specified as invariants of OZClass.



In class *OZSpecification*, an *OZ* specification is composed of a system class and a set of component classes. If two *OZ* specifications contain the same system class, they model the system from an identical viewpoint [5]. Thus, the two specifications can be considered as identical. Class names should be unique within the *OZ* specification in which they are used.



53.5 UML-B Metamodel

We present parts of UML-B metamodel which we are concerned in our mapping from UML-B to *OZ*. Figure 53.3 shows part of UML-B metamodel which mainly relates to class schema. We applied some changes in the UML-B metamodel given in [21] in order to build this part of metamodel. More precisely, we introduced association “invariants” and removed metaclass “stateMachineCollection” to propose the concept of UML-B class invariants. In Fig. 53.3, UML-B class is represented by *UML-BClass* which has a set of attributes, events and invariants.

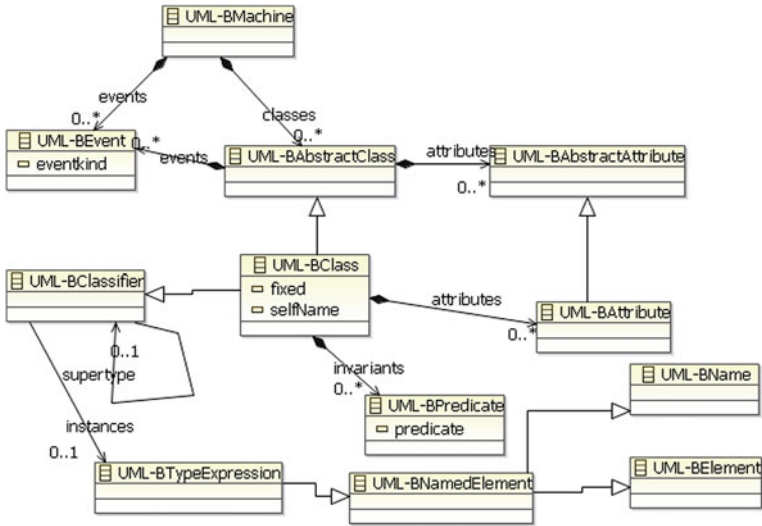


Fig. 53.3 Part of UML-B metamodel (class)

Also, it has two attributes “selfName”, which denotes self attaching to class name and “fixed”, which determines how to treat with class in translation to Event-B. Note that we do not propose any formal mapping from state machines to OZ constructs in this paper.

Figure 53.4 shows the structure of events in UML-B. UML-BEvent is a metaclass which represents UML-B events. It contains a set of event variables (i.e., event parameters), guards and actions and also an attribute eventKind which is ordinary, anticipated or convergent.

Figure 53.5 shows the structure of class attributes in UML-B. In this figure, attributes are represented by UML-BAttribute which has an initial value, a set of properties (i.e., injective, surjective, functional and total) and type.

53.6 Formal Description of UML-B Metamodel In OZ

The OZ classes OZUML-BProperty and OZUML-BAttribute are formal descriptions of their counterparts in UML-B metamodel, i.e., UML-BProperty and UML-BAttribute, respectively. We do not propose formal definition of UML-BVariableElement and UML-BAbstractAttribute because they can be modeled in the same way as UML-BProperty and UML-BAttribute.

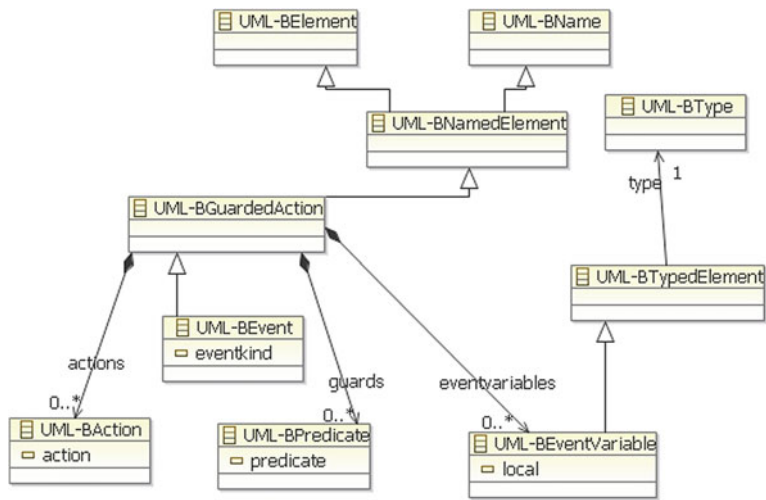


Fig. 53.4 Part of UML-B metamodel (event) [21]

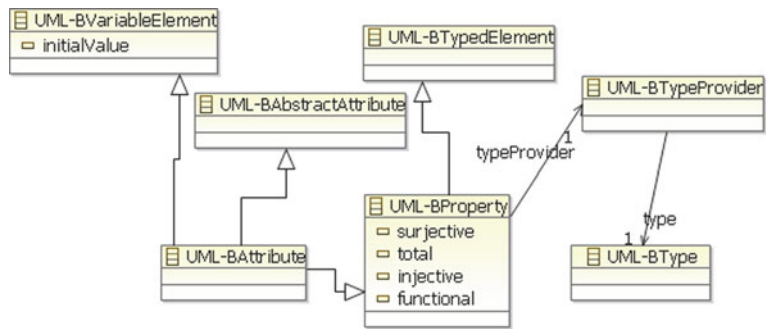
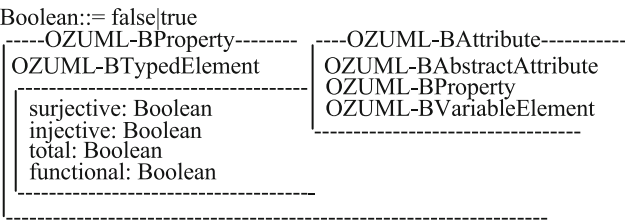
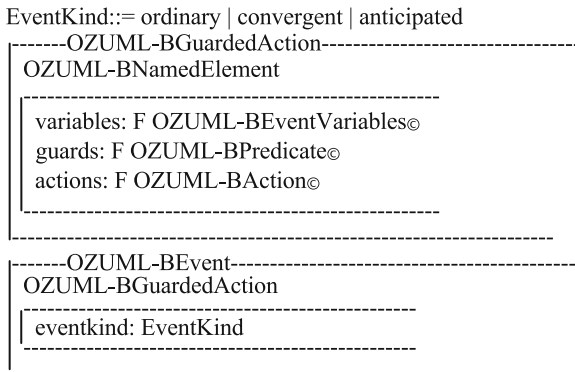


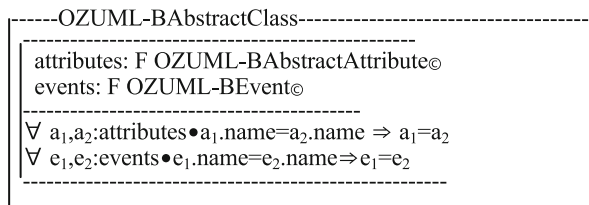
Fig. 53.5 Part of UML-B metamodel (attribute) [21]



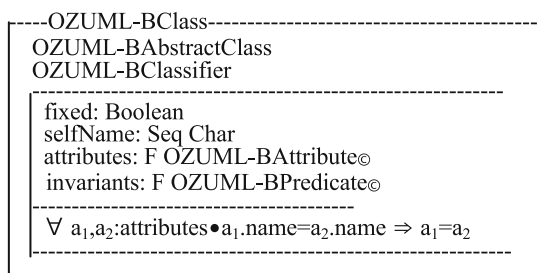
Formal description of events in UML-B is described using OZ class named **OZUML-BEvent** which inherits **OZUML-BGuardedAction** and has an attribute named “eventkind”. **OZUML-BGuardedAction** models guarded actions which have their own set of guards, actions and variables.



OZ Class OZUML-BAbstractClass is a formal definition of metaclass UML-BAbstractClass. Each abstract class has its own abstract attributes and events. Also, the invariants of OZUML-BAbstractClass state that abstract attributes and events must have unique names within an abstract class.

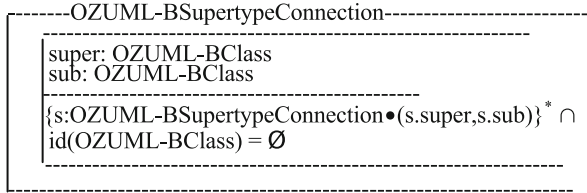


OZUML-BClass corresponds to class in UML-B. Each UML-B class has its own attributes, invariants and events.

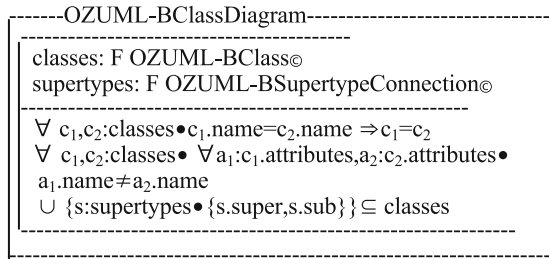


There are two types of connections in UML-B class diagram. One is association which is in the form of an attribute of at least one of the classes at the end of the association (we do not need to consider a new OZ class to formalize this concept because it is formalized implicitly using UML-BAttribute). Supertype is another type of connection which implies inheritance relationship between classes.

OZUML-BSupertypeConnection is a formal description of Supertype connection. Each Supertype connection has two variables “sub” and “super” declared to represent the superclass and subclass involved in a Supertype connection, respectively. The invariant prohibits any circular inheritance [5].



A UML-B class diagram is a collection of classes and their connections. Classes and their attributes should have unique names within a class diagram. Also, classes involved in the Supertype connection should be classes in the class diagram. We formalize this invariant in the same way as [5].



53.7 A Formal Mapping Between UML-B Constructs and OZ Constructs

In this section, we describe a formal mapping from UML-B models to their counterparts in OZ using our mapping functions in OZ.

53.7.1 Transformation Rule for UML-B Class

We consider an OZ mapping function, named `mapUML-BClassToOZ`, to map a UML-B class to at least one OZ class. This function takes a UML-B class and returns corresponding OZ class(es). The UML-B class name is used as OZ class name. The OZ class has all attributes of UML-B class as attributes in its state

schema and also in its visibility list. We use two functions convUPredToOPred , which maps UML-B predicates to OZ predicates, and convUTypeToOType , which maps UML-B types to OZ types.

Events of a UML-B class are considered as operations of an OZ class in the way that guards and actions of events are considered as preconditions and postconditions of operations, respectively. This is done using two functions convUPredToOpPred , which maps UML-B predicates to operation predicates in OZ, and $\text{convUActionToOpPred}$, which maps UML-B actions to operation predicates in OZ. Also, variables of events are considered as parameters of operations in OZ. Note that all of operations of OZ class are put in its visibility list. Finally, invariants of UML-B class are considered as invariants of OZ class.

```

| convUPredToOPred: OZUML-BPredicate  $\rightarrow$  OZPredicate
| convUTypeToOType: OZUML-BType  $\rightarrow$  OZType
| convUPredToOpPred:  $\mathbb{P}$ OZUML-BPredicate  $\rightarrow$   $\mathbb{P}$ OpPredicate
| convUActionToOpPred:  $\mathbb{P}$ OZUML-BAction  $\rightarrow$   $\mathbb{P}$ OpPredicate

|-----
| mapUML-BClassToOZ: OZUML-BClass  $\rightarrow$   $\mathbb{P}$ OZClass
|-----
|  $\forall$  uc:OZUML-BClass •
|   mapUML-BClassToOZ(uc)= {oc:OZClass |
|     oc.name  $\wedge$  <S,e,l,f>= uc.name  $\wedge$ 
|        $\forall$  ua:uc.attributes •  $\exists$  oa:oc.features | oa  $\in$  OZAttribute •
|         oa.name=ua.name  $\wedge$  oa.visibility=public  $\wedge$ 
|         oa.multiplicity= {x, y, z, r| ((x=0  $\Leftrightarrow$  ua.surjective=false)  $\vee$ 
|           (x=1  $\Leftrightarrow$  ua.surjective=true))  $\wedge$  ((y=0  $\Leftrightarrow$  ua.injective=false)
|              $\vee$  (y=1  $\Leftrightarrow$  ua.injective=true))  $\wedge$  ((z=0  $\Leftrightarrow$  ua.total=false)
|                $\vee$  (z=1  $\Leftrightarrow$  ua.total=true))  $\wedge$  ((r=0  $\Leftrightarrow$  ua.fucntional=false)  $\vee$ 
|                 (r=1  $\Leftrightarrow$  ua.functional=true))}  $\wedge$ 
|         oa.initialStateValue =convUPredToOPred(ua.initialValue)  $\wedge$ 
|         oa.type =convUTypeToOType(ua.type)  $\wedge$ 
|        $\forall$  ue:uc.events •  $\exists$  oo:oc.features|oo  $\in$  OZOperation •
|         oo.name=ue.name  $\wedge$  oo.visibility=public  $\wedge$ 
|         oo.pre-condition =convUPredToOOpPred(ue.guards)  $\wedge$ 
|         oo.post-condition =convUActionToOpPred(ue.actions)  $\wedge$ 
|          $\forall$  uv:ue.variables|uv.local=false •  $\exists$  op:ranooo.parameters •
|           op.name=uv.name  $\wedge$  op.role= input  $\wedge$ 
|           op.type= convUTypeToOType(uv.type)
|        $\wedge$  oc.invariants=convUPredToOInv(uc.invariants)

```

53.7.2 Transformation Rule for UML-B Supertype Connection

In order to propose formal mapping for Supertype connection, we use function $\text{mapUML-BSupertypeToOZ}$ which takes a Supertype connection and returns classes relating to this Supertype connection as superclass and subclass. Also, we consider a predicate in this function which states the resulting superclass must be one of the superclasses of the resulting subclass.

| |
|---|
| $\begin{aligned} &\text{mapUML-BSupertypeToOZ: OZUML-BSupertypeConnection} \rightarrow \\ &\mathbb{P}(\text{OZClass} \times \text{OZClass}) \end{aligned}$ <hr style="border-top: 1px dashed black;"/> $\begin{aligned} &\forall s: \text{OZUML-BSupertypeConnection} \bullet \\ &\quad \text{mapUML-BSupertypeToOZ}(s) = \{\text{supoc}, \text{suboc}: \text{OZClass} \mid \\ &\quad \text{supoc} \in \text{mapUML-BClassToOZ}(s.\text{super}) \wedge \\ &\quad \text{suboc} \in \text{mapUML-BClassToOZ}(s.\text{sub}) \wedge \\ &\quad \text{supoc} \in \text{suboc.superclasses} \end{aligned}$ |
|---|

53.7.3 Transformation Rule for UML-B Class Diagram

After proposing formal mapping for different parts of UML-B class diagram, we are now in a position to give formal mapping from a UML-B class diagram to an OZ specification using function $\text{mapUML-BClassDiagramToOZ}$. When a class is interpreted as a component within a class diagram, the class represents a set of existing objects of that class at a certain point in time [5]. We formalize this semantics in the same way as [5]: we instantiate corresponding OZ classes as sets within the OZ specification. For this purpose, we define a function powerC which maps OZ class to its power set form. When a class is inherited by other classes, the type of the set that represents the existing instances of the class is a PolyClass (i.e., polymorphism of class). Thus, we define function polyC which maps an OZ class to its corresponding PolyClass .

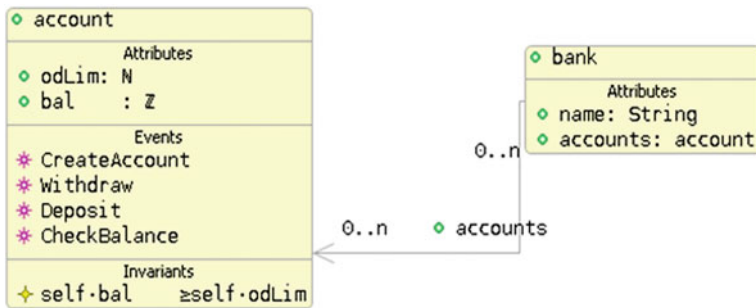


Fig. 53.6 Bank and account class diagram in UML-B [20]

```

| PClass:  $\mathbb{P}$ OZClass
| PolyClass:  $\downarrow$ OZClass
| polyC: OZClass  $\rightarrow$  PolyClass
| powerC: OZClass  $\rightarrow$  PClass
| mapUML-BClassDiagramToOZ: OZUML-BClassDiagram  $\rightarrow$ 
|   OZSpecification
| -----
|  $\forall u: \text{OZUML-BClassDiagram}; o: \text{OZSpecification} \bullet$ 
|   mapUML-BClassDiagram(u) = o  $\Leftrightarrow$ 
|     {c: u.classes • c.name} = {c: o.component • c.name}  $\wedge$ 
|     #u.classes = #o.component  $\wedge$ 
|      $\forall uc: u.classes \bullet \exists oc: o.component \bullet$ 
|       oc  $\in$  mapUML-BClassToOZ(uc)  $\wedge$ 
|        $\forall s: u.supertypes \bullet \exists osub, osup: o.component \bullet$ 
|         (osub, osup)  $\in$  mapUML-BSupertypeToOZ(s)  $\wedge$ 
|        $\forall oc: o.component \bullet$ 
|         {soc: o.component | oc  $\in$  soc.superclasses} =  $\emptyset \Rightarrow$ 
|            $\exists at: o.system.features | at \in \text{OZAttribute} \bullet at.name = oc.name \wedge$ 
|             at.type = powerC(oc)
|         {soc: o.component | oc  $\in$  soc.superclasses}  $\neq \emptyset \Rightarrow$ 
|            $\exists at: o.component.features | at \in \text{OZAttribute} \bullet$ 
|             at.name = oc.name  $\wedge$  at.type = (polyC; powerC)(oc)
  
```

Fig. 53.7 Properties of event Withdraw [20]

| Class Event : withdraw | | | | | | | | | |
|------------------------|-------------|--|--|------|------|-------|----|---|-------|
| Properties | Name: | withdraw | | | | | | | |
| Refines | | | | | | | | | |
| Parameters | | | | | | | | | |
| Witness | | | | | | | | | |
| Guards | | | | | | | | | |
| Actions | | | | | | | | | |
| Errors | | | | | | | | | |
| | Parameters: | <table><tr><th>Name</th><th>Type</th><th>Local</th></tr><tr><td>am</td><td>N</td><td>false</td></tr></table> | | Name | Type | Local | am | N | false |
| Name | Type | Local | | | | | | | |
| am | N | false | | | | | | | |
| | Guards: | self.bal ≥ am | | | | | | | |
| | Actions: | self.bal = self.bal - am | | | | | | | |

Fig. 53.8 Account class schema in OZ

| |
|---|
| -----account----- |
| !(odLim, balance, CreateAccount, Withdraw, Deposit, CheckBalance) |
| ----- |
| odLim: \mathbb{N} |
| balance: \mathbb{Z} |
| ----- |
| convUPredToOInv(self.balance >= self.odLim) |
| ----- |
| ---CreateAccount----- |
| ----- |
| ---Withdraw----- |
| am?: \mathbb{N} |
| ----- |
| convUPredToOOpred(self.bal >= am) |
| convUPredToOOpred(self.bal = self.bal - am) |
| ----- |
| ---Deposit----- |
| ----- |
| ---CheckBalance----- |
| ----- |

Fig. 53.9 Bank class schema in OZ

| |
|-------------------------------------|
| -----bank----- |
| !(name, accounts) |
| ----- |
| name: convUTypeToOType(String) |
| accounts: convUTypeToOType(account) |
| ----- |

53.8 Case Study

We map the following UML-B class diagram to an OZ specification using our mapping functions. Two classes `account` and `bank` are in this class diagram. Also, these classes are connected using an association, named `accounts`.

As an instance, we show parameters, guards and actions of “Withdraw” in class “`account`”:

The OZ classes “`account`” and “`bank`” are resulted from mapping of the UML-B class diagram in Fig. 53.6.

53.9 Conclusions and Future Work

One can benefit from our mapping in the software development process by using UML-B as a visual modeling notation in the early conceptual modeling stage and OZ in next stages when requirements are better understood. Also, we are now able to verify visual models without translating them to OZ specifications. Finally, after translating the UML-B model to OZ specification using our formal mapping, it is possible to use existing approaches, such as [22], (Fig. 53.7) to map the OZ specification into object-oriented code (Fig. 53.8).

In our future work, we are going to present formal mappings for the dynamic part of UML-B models, suggest formal mappings for refinement in UML-B, and prove the correctness of given mappings formally (Fig. 53.9).

References

1. Kim SK, Carrington D (2002) A formal metamodeling approach to a transformation between visual and formal modeling techniques. Technical Report No. 02-23, University of Queensland, Brisbane
2. Meyer E, Souquieres J (1999) A systematic approach to transform OMT diagrams to a B specification. In: Formal methods conference, vol 1, LNCS 1708, Springer, Verlag, pp 875–895
3. Mc UMBER W, Cheng B (2001) A general framework for formalizing UML with formal languages. In: IEEE Conference on Software Engineering, IEEE Computer Society Press, pp 433–442
4. Kim SK, Carrington D (2002) A formal metamodeling approach to a transformation between the UML state machine and Object-Z. In: ICFEM, LNCS 2459, Springer, Verlag, pp 548–560
5. Kim SK, Carrington D (2000) A formal mapping between UML models and Object-Z specifications. Technical Report No. 00-03, University of Queensland, Brisbane
6. Ehrler TD (2004) An informal mapping from UML models to Object-Z specifications. MSc thesis, University of London, Twickenham
7. Dupuy S, Ledru Y, Chabre-Peccoud M (1998) Translating the OMT dynamic model into Object-Z. In: ZUM’98- The Z formal specification notation, 12th international conference of Z users, LNCS. No. 1498, Springer-Verlag, pp. 347–366

8. Kim SK, Carrington D, Duke R (2001) A metamodel-based transformation between UML and Object-Z. In: *Proceedings of IEEE Symposia on Human-Centric Computing Languages and Environments*, IEEE Computer Society Press, pp 112–119
9. Roe D, Broda K, Russo A (2003) Mapping UML models incorporating OCL constraints into Object-Z. Technical Report, Imperial College, London
10. Wang E, Richter H, Chen B (1997) Formalizing and integrating the dynamic model with OMT. In: *Proceedings 19th international conference on software engineering*, pp 45–55
11. Younes AB, Ayed LJB (2008) From UML activity diagrams to Event-B for the specification and verification of workflow applications. In: *32nd annual IEEE international computer software and applications conference*, IEEE Computer Society Press, pp 643–648
12. Booch G, Jacobson I, Rumbaugh J (1998) *The unified modeling language—A reference manual*, Addison Wesley
13. Smith G (2000) *The object-Z specification language*. *Advances in formal methods*, Kluwer Academic Publishers, Dordrecht
14. Duke R, Rose G (2000) *Formal object-oriented specification using object-Z*, Macmillan, UK
15. Snook C, Butler M (2008) UML-B and Event-B: an integration of languages and tools. In: *The IASTED international conference on software engineering (SE2008)*
16. Snook C, Butler M (2004) U2B—a tool for translating UML-B models into B. In: *UML-B specification for proven embedded systems design*, April 2004
17. Snook C, Butler M (2006) UML-B: formal modelling and design aided by UML. *ACM T Softw Eng Meth* 15(1):92–122
18. Snook C, Butler M, Oliver I (2004) The UML-B profile for formal systems modeling in UML. In: *UML-B specification for proven embedded systems design*
19. Abrial JR (2010) *Modeling in Event-B: system and software engineering*, Cambridge University Press, New York
20. Said MY (2010) *Methodology of refinement and decomposition in UML-B*, PhD Thesis, University of Southampton
21. Joachim T (2010) *Bringing requirements engineering to formal methods: timing diagrams for Event-B and KAOS*, PhD Thesis, University of Southampton
22. Najafi M, Haghighi H (2011) An animation approach to develop C++ codes from Object-Z specifications. In: *International symposium on Computer Science and Software Engineering*, pp 9–16

Chapter 54

Algorithm for Dynamic Traffic Rerouting and Congestion Prevention in IP Networks

Martin Hrubý, Margaréta Kotočová and Michal Olšovský

Abstract With the expanding amount of data transferred over communication links it is necessary to improve the links and appropriate network devices to match the traffic requests. The most common way of increasing network throughput and performance in general is usually the replacement of the network devices and links. This way is reliable but usually expensive. Different way of improving network performance is the change of way how the traffic is handled and distributed over network. In this paper we propose an algorithm for dynamic traffic rerouting in IP networks based on statistical probability and load experienced on a network link. This algorithm accomplishes congestion prevention and even distribution of traffic on available network resources.

54.1 Introduction

Sub-optimal traffic distribution in computer networks is the issue we presently face when delivering bandwidth intensive services with QoS guarantees. Usually the decision on which paths to use to forward traffic is left on the interior routing

M. Hrubý (✉) · M. Kotočová · M. Olšovský
IEEE Conference Publishing, Institute of Computer Systems and Networks, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Ilkovičova 3, 842 16, Bratislava 4, Slovakia
e-mail: hruby@fiit.stuba.sk

M. Kotočová
e-mail: kotocova@fiit.stuba.sk

M. Olšovský
e-mail: olsovsky@fiit.stuba.sk

protocol [1, 2]. Our objective is to optimize the flow of traffic in the network by implementing a statistical rerouting algorithm which will penalize greedy flows and prioritize sensitive flows in a computationally efficient and effective manner. The objective of interior gateway routing protocols is to choose the best route toward a destination and forward traffic along that route. Typically traffic is only spread across multiple links when equal-cost multipath routing is in effect [3]. With policy-based routing, the network administrator can route certain statically defined traffic along different paths but this approach is hard to implement efficiently and cumbersome to maintain, not mentioning the disability of such a solution to react to ever changing demands of traffic flows. Traffic engineering tunnels may be deployed but this solution is complex and requires knowledge of future traffic load distribution to be done efficiently, not mentioning the disability to react dynamically to changing traffic requirements [4].

In this paper we propose an algorithm for dynamic rerouting of traffic flows in order to spread the bandwidth demands more evenly across all available resources and prevent the congestion of those links that are chosen as best paths by the routing protocol [5].

Current approaches to traffic engineering and traffic flow optimization utilize the traffic matrix which contains origin–destination pairs together with traffic demands imposed on the computer network [6]. The issue with traffic matrices is that, based on chosen origin–destination granularity, the table can become extremely large and keeping such a large matrix updated is computationally infeasible. In this paper we propose a dynamic traffic rerouting algorithm based on network load, which does not require the presence of a traffic matrix.

54.2 Concept

We assume a transit computer network (e.g., service provider backbone) which is IP-based and has a link-state interior gateway routing protocol configured (e.g., OSPF or IS-IS). Generally, flows will traverse the best path as chosen by the routing protocol, but additional resources may be available which are left under-utilized. This may, in times of heavy traffic cause congestion of the primary link.

In our approach a central computing server is present which has SNMP access to each network node (i.e., router) in the network topology. By listening to SNMP traps signaling congestion, the computing server can recalculate traffic routes and deliver updated configurations to selected routers in order to reroute the flow of traffic along a different route.

The objective is to reroute a big enough portion of the flows to alternate paths, thereby preventing the congestion from occurring and spreading the load imposed on available under-utilized resources. To achieve this, the computing server will calculate which prefixes need to be switched over an alternate path and then will deliver a configuration update to their current next-hop router which will cause them to be advertised with a worse metric.

Table 54.1 Data structure
for list of interface

| # | Name |
|---|------|
|---|------|

The algorithm we propose in this paper aims at rerouting “congestors”, in other words, a small amount of flows which are greedy and are responsible for a disproportionate amount of network traffic. Because no traffic matrix is used, it is impossible to determine which flows are responsible for the congestion and our algorithm proposes an iterative switching method to identify them.

54.3 Algorithm

Basic functionality of the algorithm can be divided into following two phases:

- 1. Learning phase
- 2. Production phase.

While designing our algorithm we adhered to the traffic engineering process model as proposed in [7, 8] and a similar approach might be found in [9]. Learning phase is responsible for creating a trustworthy model of the network communication. It means that it usually takes some time to get reliable statistical information of the communicating subnets in general. Once this model is created, the algorithm can enter production phase—we can easily and quickly react on upcoming congestion in the network. We will be able to redirect specific traffic, which is statistically responsible for the largest part of the upcoming congestion. Production phase will be acting as learning phase as well—it will update collected information about subnets to reflect actual situation to provide the most accurate data.

For fully operational process we will need to store specific information about prefixes and network interfaces for each node in the monitored network. For these purposes we have introduced new data structures—tables with structures as shown in Tables 54.1 and 54.2. Despite the fact that Table 54.2 has multiple rows, both tables represents one item in the list of stored IP prefixes and interfaces, called headers. Together with these structures we have introduced a parameter called parameter of alternative interface (λ). The λ parameter represents current availability of the interface based on the current link load and bandwidth. The λ can be calculated using formula (54.1).

$$\lambda = (1 - \text{link_load}) * \text{bandwidth} \tag{54.1}$$

Table 54.1 is used for storing all known prefixes and their details. We assume that every single prefix, which has at least one backup path, will be stored in this table. For each prefix following information will be stored:

Table 54.2 Data structure for list of prefixes

| # | Prefix | Mask | Int |
|-----------------|------------|-------------|----------|
| θ | χ | Ω | ω |
| λ_c | IP_c | λ_0 | IP_0 |
| λ_1 | IP_1 | λ_2 | IP_2 |
| ... | | | |
| λ_{n-1} | IP_{n-1} | λ_n | IP_n |

- *Prefix*—network prefix, more the one identical prefixes can occur but not with the same mask. Pair prefix with mask is unique. Stored in dotted-decimal notation.
- *Mask*—subnet mask of the IP prefix. Stored in dotted-decimal notation.
- *Int*—current outgoing interface for a specific IP prefix. Stored as integer used as a key to list of interfaces
- θ —switch count—variable used for counting prefixes switched between interfaces. By default every prefix has this values set to 8. This value is decremented by 1 together with each switch operation. Stored as integer.
- χ —“congestor” is a probability that a specific subnet is responsible for upcoming or existing congestion on the link. By default set to 50 %. Stored as integer.
- λ_c — λ of current outgoing interface. Stored as integer.
- IP_c —IP address of the next-hop which is used for specific subnet using its current outgoing interface. Stored in dotted-decimal notation.
- $\lambda_{0...n}$ — λ of outgoing interfaces 0 to n. Filled with valid values only if interface can be used as an alternative outgoing interface for a specific IP prefix, otherwise filled with -1. Stored as integer.
- $IP_{0...n}$ —IP addresses of the alternative next-hops which can be used for specific subnet. Filled with valid IP only if interface can be used as alternative outgoing interface for specific prefix and this IP is device behind that interface. Stored in dotted-decimal notation.
- Ω —estimation of bandwidth generated and occupied by specific IP prefix. Once group of flows is switched, we are able to determine the amount of data this group was producing. This amount is divided with number of switched subnets and result is put into Ω . By default set to 0. Stored in megabits.
- ω —percentual probability of Ω . As Ω is inaccurate, we need to know, how trustworthy this information is. Ω got by switching less subnets is more trustworthy and can be calculated using Eq. (54.2):

$$\omega[\%] = 100/\eta \quad (54.2)$$

where η stands for number of switched prefixes.

The Algorithm will be comprised in the following steps:

54.3.1 Data collection part

1. Retrieve all prefixes with current next-hop and all available backup next-hops
2. Calculate the λ parameter of all available interfaces
3. θ filled as 0, χ as 50 %, Ω as 0 Mb and ω as 0 %

54.3.2 Optimization part

4. Congestion detected (link loads > 80 %).
5. Select 20 % of prefixes for switch operation based on mask, χ , Ω and ω which are switch capable—their outgoing interface is congested and they have another known path.
6. Get next-hop IP addresses for chosen prefixes and calculate new metric for chosen prefixes.
7. Update next-hop routers with new metric.
8. Update stored variables for switched prefixes.
9. Get differentials for link loads of interfaces after switching prefixes and compare with previous link loads.
10. Retrieve current link loads
11. If link loads are less than 80 % switch operation was successful, χ of switched prefixes will be increased by 5 and optimization is finished. Otherwise calculated differentials are compared and in case they are small, another group of prefixes is selected and old prefixes have χ decreased by 5. Algorithm continues in step 5. Otherwise group of prefixes is decreased by one half and process continues in step 5.

For better understanding, the algorithm is described in pseudo code on the following lines:

begin

//Initial data input

PREFIXES_TO_SWITCH = w_random(ALL_PREFIXES, 20 %);

OPTIMIZING = TRUE;

ALL_LINK_LOADS = link_loads(THIS_ROUTER);

while (OPTIMIZING) **do**

begin

//get next-hop IP addresses for chosen

prefixes

```

NEXT_HOPS = next_hop_IP(PREFIXES_TO_SWITCH);
foreach (NEXT_HOP in NEXT_HOPS)
begin
foreach (PREFIX in PREFIXES_TO_SWITCH)
begin
  //calculate new metrics for chosen
  prefixes
  //and update next-hop routers with
  these new metrics
  //effectively switching them over an
  alternate path
  NEW_METRIC = calculate_metric(PREFIX);
  send_update(NEXT_HOP, PREFIX, NEW_METRIC);
end
end
  //update state variables for switched
  prefixes
  update(PREFIXES_TO_SWITCH);
end
  //get differentials for link loads after
  switching prefixes
  DIFFERENTIALS = ALL_LINK_LOADS
  link_loads(THIS_ROUTER);
  if (link_loads(THIS_ROUTER) < 80 %)
  then OPTIMIZING = FALSE;
  else
    if (DIFFERENTIALS = small)
    then
      begin
        switchback(PREFIXES_TO_SWITCH);
        PREFIXES_TO_SWITCH = w_random(ALL_PREFIXES -- PRE-
        FIXES_TO_SWITCH, 20 %);
      end
    else
      begin
        ALL_PREFIXES = w_random(PREFIXES_TO_SWITCH, 50 %);
      end
    end
    update(PREFIXES_TO_SWITCH);
  end

```

Pseudo code is using bellow variables:

- THIS_ROUTER—holds the ID of the current (congested) router
- ALL_PREFIXES—holds all prefixes eligible for switching

- **PREFIXES_TO_SWITCH**—holds prefixes chosen for switching in the next iteration
- **PREFIX**—holds a single network prefix
- **ALL_LINK_LOADS**—holds link loads for each link of a given router
- **OPTIMIZING**—boolean variable to indicate an optimization in progress
- **NEXT_HOPS**—holds a list of all next-hop IP addresses
- **NEXT_HOP**—holds a single next-hop IP address
- **NEW_METRIC**—holds the value of the new metric for a given prefix
- **DIFFERENTIALS**—holds increments (decrements) of link loads after switching

Pseudo code is using bellow functions:

- *w_random()*—weighted random function, returns a preferred array of prefixes which currently have next-hop IPs over a congested link and are suitable for switching over an alternate link (i.e., are congestors). Prefixes are sorted by χ , Ω , ω and finally by mask.
- *next_hop_IP()*—returns the (array of) next-hop IP address for a parameter prefix.
- *calculate_metric()*—returns new calculated metric for provided prefix in a way that prefix with this metric will be preferable.
- *send_update()*—sends configurations to update route advertisements of a “switched-to-be” prefix on a next-hop router.
- *link_loads()*—returns link loads for all links of a parameter router.
- *switchback()*—switches prefixes switched in a previous iteration back to their original next-hop IPs together with decreasing their χ .
- *update()*—update state variables for switched prefixes, especially χ , Ω and ω using following formulas:

$$\begin{aligned}\chi &= \chi + 5 \mid (!\text{OPTIMIZING AND } \theta > 0) \text{ OR} \\ &\quad \text{DIFFERENTIALS} = \text{small}) \\ &= \chi - 5 \mid \text{other}\end{aligned}\tag{54.3}$$

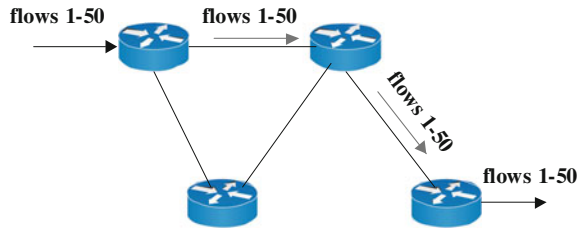
$$\begin{aligned}\Omega &= \text{DIFFERENTIALS}/\eta \mid \omega < 100/\eta \\ &= \Omega \mid \text{other}\end{aligned}\tag{54.4}$$

$$\begin{aligned}\omega &= 100/\eta \mid \omega < 100/\eta \\ &= \omega \mid \text{other}\end{aligned}\tag{54.5}$$

The key function of the *update()* function is to distinguish “congestor” prefixes from other prefixes so they will be easily recognized in next iterations (switch operations) and will decrease the convergence time.

Using proposed algorithm we did preliminary testing described in the next chapter.

Fig. 54.1 Common traffic flow



54.4 Preliminary Testing

Currently we have not deployed our algorithm into a real network environment, thus exact performance measurements are not yet available. However, in our present work we aimed our focus on implementing a mathematical model of our algorithm in Matlab. This enabled us to create simulations which show promising results and encourage us to deploy the algorithm into a real laboratory environment for further testing and fine-tuning. In this chapter we would like to summarize our preliminary results obtained via simulation of the algorithm on the below topology.

In our simulations we assumed an initial statistical distribution of the network load on available links, where the probability that an IP prefix is routed via a link is as depicted in the figure above. A set of 20 IP prefixes were used where each prefix consumed a random portion of network bandwidth ranging from 0 Mb/s to 20 Mb/s. Our simulations repeated the basic statistical distribution of network load until link A was utilized more than 80 % (*i.e. congested*) at which point our proposed algorithm was triggered (Figs. 54.1, 54.2, 54.3, 54.4, 54.5).

In our first experiment, link A was utilized at 81.9 %, link B at 9.8 % and link C at 0 %. Link A, approaching congestion, triggered an alarm and set off the optimization algorithm. Randomly selected IP prefixes were switched over to the least utilized link, link C. In the first iteration (as visible in time interval 2, in Fig. 54.6), 4 IP prefixes were randomly switched which resulted in offloading 14.3 % of the traffic onto link C. Link A's utilization was reduced to 67.7 % which is under the threshold and acceptable. No further iterations were performed, but traffic generation was simulated for an additional 4 time intervals to verify that the link will, statistically, not become congested again. Our first experiment was successful in the first iteration because our algorithm, by chance identified IP prefixes with mostly non-zero traffic.

For our second experiment, IP prefixes with mostly zero active flows were chosen in the first iteration and this therefore did not result in effective traffic offloading, as determined in time interval 2 in Fig. 54.7 were Link A is still congested (at over 83.1 %). Flows switched over to link B were replaced with new candidates in iteration 2 and this resulted in 8.6 % of traffic offloading from link A to link C as seen in time interval 3 below. Further 3 time intervals were simulated to verify that the link will not become congested again.

Fig. 54.2 Congestion occurred

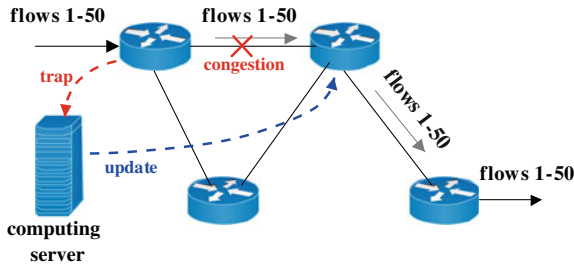


Fig. 54.3 Congestion suppressed—specific flows offloaded

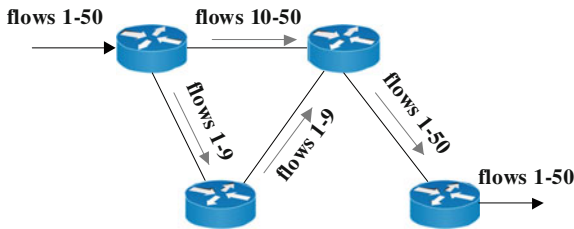
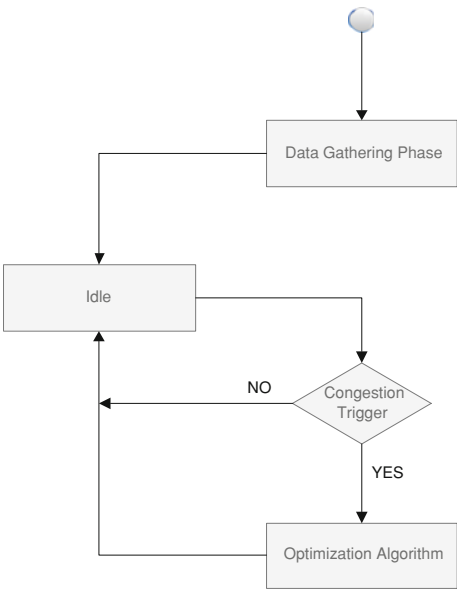


Fig. 54.4 General flow of the optimization process



For our third experiment, we increased the number of prefixes two-fold to 40. As seen in Fig. 54.8, in the first iteration our algorithm randomly chose 8 IP prefixes to be switched onto link B, which caused 33.5 % link utilization decrease on link A, preventing congestion from occurring.

Fig. 54.5 Simulation topology

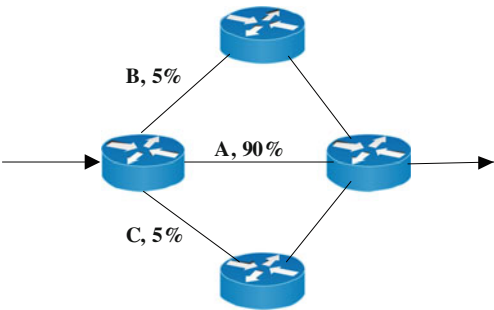


Fig. 54.6 Simulation 1 of the algorithm

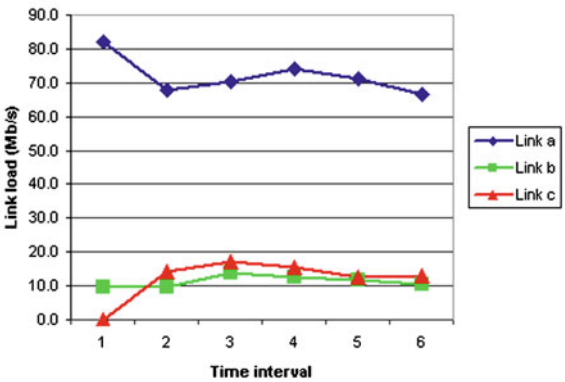
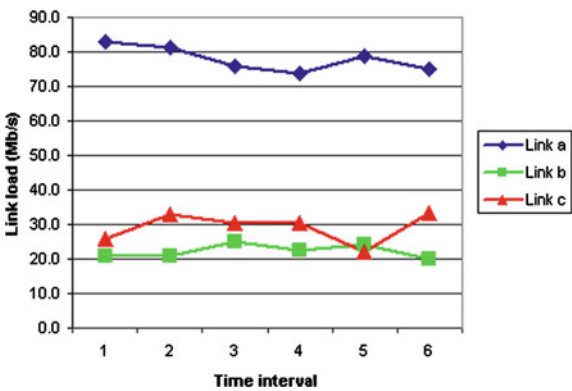
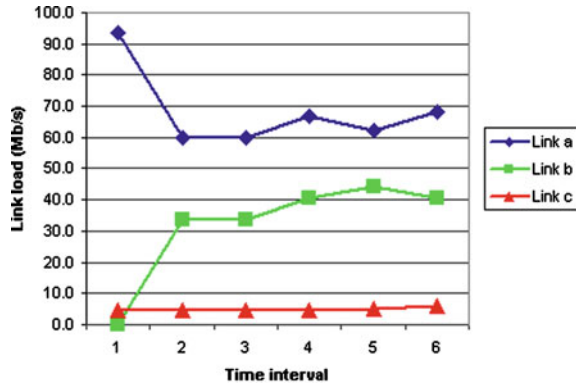


Fig. 54.7 Simulation 2 of the algorithm



In this experiment we have shown that the higher the number of IP prefixes, the bigger the probability of randomly selecting non-zero flows in the first iteration is. The experiment was simulated for an additional four time intervals to determine that link A will not become congested again.

Fig. 54.8 Simulation
3 of the algorithm



54.5 Conclusion

In this paper we have proposed an algorithm for statistical, iterative traffic off-loading in times of network congestion. Our preliminary results, simulated in a mathematical framework, have shown that the algorithm successfully prevents congestion of a network link in less than three iterations, switching a portion of network flows over an alternate path.

The algorithm functions in two ways. First, the congestion is prevented by switching a big enough portion of the network demand along another link. Second, the number of IP prefixes switched along the alternate path is minimized while maintaining status information about them.

We believe that this paper can contribute to the important research field in the communication network performance.

In our future work we will implement this algorithm into a real network environment and test its ability to predictably react to network congestion in computer networks during peak traffic demands and over long time periods.

Acknowledgments The support by Slovak Science Grant Agency (VEGA 1/0649/09 “Security and reliability in distributed computer systems and mobile computer networks”) is gratefully acknowledged.

References

1. He J, Bresler M, Chiang M, Rexford J (2007) Towards robust multi-layer traffic engineering: optimization of congestion control and routing. *IEEE J Sel Areas Commun* 25(5):868–880
2. Wang N, Ho K, Pavlou G, Howarth M (2008) An overview of routing optimization for internet traffic engineering. *IEEE Commun Surv Tutor* 10(1):36–56
3. Abrahamsson H, Bjorkman M (2009) Robust traffic engineering using l-balanced weight-settings in OSPF/IS-IS. In: *Broadband communications, networks, and systems*, pp 1–8
4. Wang X, Wan S, Li L (2009) Robust traffic engineering using multi-topology routing. In: *Global telecommunications conference*, pp 1–6

5. Casas P, Fillatre L, Vaton S (2008) Multi hour robust routing and fast load change detection for traffic engineering. In: IEEE international conference on communications, pp 5777–5782
6. Gunnar A, Johansson A, Telkamp T (2004) Traffic matrix estimation on a large IP backbone: a comparison on real data. In: Proceedings of the 4th ACM SIGCOMM conference on internet measurement, 25–27 October 2004, pp 149–160
7. Awduche DO, Jabbari B (2002) Internet traffic engineering using multi-protocol label switching (MPLS). *Computer networks. Int J Comput Telecommun Netw* 40(1):111–129
8. Zhang M, Liu B, Zhang B (2009) Multi-commodity flow traffic engineering with hybrid MPLS/OSPF routing. In: IEEE global telecommunications conference, pp 1–6
9. Fortz B, Rexford J, Thorup M (2002) Traffic engineering with traditional IP routing protocols. *IEEE Commun Mag* 40(10):118–124

Chapter 55

Fovea Window for Wavelet-Based Compression

J. C. Galan-Hernandez, V. Alarcon-Aquino, O. Starostenko
and J. M. Ramirez-Cortes

Abstract Wavelet foveated compression can be used in real-time video processing frameworks for reducing the communication overhead while keeping high visual quality. Such algorithm leads into high rate compression results due to the fact that the information loss is isolated outside a region of interest (ROI). The fovea compression can also be applied to other classic transforms such as the commonly used the discrete cosine transform (DCT). In this paper, a fovea window for wavelet-based compression is proposed. The proposed window allows isolate a fovea region over an image. A comparative analysis has been performed showing different error and compression rates between the proposed fovea window for wavelet-based and the DCT-based compression algorithms. Simulation results show that with foveated compression high ratio of compression can be achieved while keeping high quality over the designed ROI.

J. C. Galan-Hernandez · V. Alarcon-Aquino · O. Starostenko
Department of Computing, Electronics and Mechatronics, Universidad de las Americas
Puebla, Sta. Catarina Martir, Cholula, 72810 Puebla. C.P., Mexico
e-mail: juan.galanhz@udlap.mx

V. Alarcon-Aquino
e-mail: vicente.alarcon@udlap.mx

J. M. Ramirez-Cortes (✉)
Department of Electronics, Instituto Nacional de Astrofisica, Optica y Electronica,
Tonantzintla, Puebla, Mexico
e-mail: olsovsky@fiit.stuba.sk

55.1 Introduction

Video processing is an intensive task and even more when it is restricted into a time window such as in real-time frameworks [1]. Compression algorithms can help to reduce communication overhead between computing nodes by reducing the redundancy of the data transmitted. Wavelet-based video compression algorithms achieve high ratios of compression. However, such algorithms also cause loss of information on video frames. In applications where a region of interest (ROI) can be isolated, foveation can be used in order to constraint the information loss only on those areas outside of the ROI in order to increase the quality of the reconstructed image. Several applications where ROIs over video frames can be identified can benefit from fovea compression such as medical video processing framework searching for melanomas by perform a lossy compression over the video frames, leaving the ROI intact for later processing. previous works in [2–4] show different foveation methods assessing the final quality of the image against the original image using human vision system (HVS) criteria. Foveation also guarantee certain quality from the HVS perspective granting the output results with enough quality for a user to inspect the ROI without noticing the loss of quality unless a close inspection outside the ROI is performed. In the work reported in this paper, a comparative analysis between wavelet-based and the DCT-based foveated compression algorithms is carried out. The remainder of this paper is organized as follows. In Sect. 55.2 an overview of foveated compression is given. Section 55.3 describes the proposed approach. Section 55.4 presents simulation results, and Sect. 55.5 presents conclusions and future work.

55.2 Foveated Compression

Wavelet transforms involve representing a general function in terms of simple, fixed building blocks at different scales and positions. These building blocks are generated from a single fixed function called mother wavelet by translation and dilation operations [5].

55.2.1 Wavelets and the Discrete Wavelet Transform

The purpose of wavelet transforms is to represent a signal into the time–frequency domain. To perform this task two functions are required, namely, a wavelet and a scaling function. If a set of mother wavelets and scaling functions is orthonormal it is called an orthonormal bases and is defined as follows [6]:

$$\{ \varphi_{l_0,n} \}_{0 \leq n \leq 2^{l_0}} \cup \{ \psi_{j,n} \}_{j < l_0, 0 \leq n \leq 2^j} \quad (55.1)$$

where 2^{l_0} is the size of the signal. Each $\psi_{j,n}$ is a translated copy of ψ at scale j :

$$\psi_{j,n}(t) = \sqrt{2^{-j}}\psi(2^{-j}t - n) \quad (55.2)$$

and each $\varphi_{l_0,n}$ is a translated copy of the scaling function φ at scale l_0 :

$$\varphi_{l_0,n}(t) = \sqrt{2^{-j}}\varphi(2^{-j}t - n) \quad (55.3)$$

For images, a two dimension wavelet transform is needed. In two dimensions, the decomposition ladder is constructed using three mother wavelets, $\psi_{j,m,n}^d$, $\psi_{j,m,n}^v$ and $\psi_{j,m,n}^h$ defined as follows [6]:

$$\psi_{j,m,n}^d = \psi_{j,m}(x)\psi_{j,n}(y) \quad (55.4)$$

$$\psi_{j,m,n}^v = \psi_{j,m}(x)\varphi_{j,n}(y) \quad (55.5)$$

$$\psi_{j,m,n}^h = \varphi_{j,m}(x)\psi_{j,n}(y) \quad (55.6)$$

with the scaling function:

$$\Phi_{j,m,n} = \varphi_{j,m}(x)\varphi_{j,n}(y) \quad (55.7)$$

where $\psi_{j,m,n}^d$ are the diagonal coefficients, $\psi_{j,m,n}^v$ are the vertical coefficients and $\psi_{j,m,n}^h$ are the horizontal coefficients. With the wavelet base defined, the next step is using it to represent a signal. The sum over all time of the signal multiplied by scaled, shifted versions of the mother wavelet ψ is given by

$$a_j(n) = \int_{-\infty}^{\infty} f(j)\psi(n,j)dt \quad (55.8)$$

where f is the signal to be represented. However, for compression this transform is not suitable because it expands the signal into more coefficients than the samples of the signal itself. A better transform, suited for compression and many other applications, is called the discrete wavelet transform (DWT). In [7], the DWT is calculated through a simple algorithm that applies two filters, a low pass filter and a high pass filter. This algorithm is known as the fast wavelet transform. In wavelet analysis of a signal f , we often speak of approximations and details. The approximations are the low-frequency components of the signal, see (55.9); whereas the details are the high-frequency components of the signal, see (55.10).

$$a_j[n] = \langle f, \varphi_{j,n} \rangle \quad (55.9)$$

$$d_j[n] = \langle f, \psi_{j,n} \rangle \quad (55.10)$$

55.2.2 Discrete Cosine Transform

The discrete cosine transform (DCT) expresses a signal in terms of cosine functions. Such transform is commonly used in the JPEG compression algorithm [8], the MP3 audio format and the VP8 video format. The discrete cosine transform for a signal $f(x)$ of length N is defined as follows [9]:

$$C(u) = \alpha(u) \sum_{x=0}^{N-1} f(x) \cos\left(\frac{\pi(2x+1)u}{2N}\right) \quad (55.11)$$

where $\alpha(u)$ is given by

$$\alpha(u) = \begin{cases} \sqrt{\frac{1}{N}} & \text{for } u = 0 \\ \sqrt{\frac{2}{N}} & \text{for } u \neq 0 \end{cases} \quad (55.12)$$

In particular, $C(u=0) = \sqrt{1/N} \sum_{x=0}^{N-1} f(x)$ is known as the direct current coefficient (DC) and the remaining coefficients are called the alternating current coefficients [10]. Most of the energy of the signal is packed in the DC coefficient.

55.2.3 Wavelet Compression

The objective of data compression is to represent a set of data with less information than the original. There are two types of compression, namely, lossy and lossless compression [10]. In lossy, some of the original data is discarded in order to achieve its goal. When the data is reconstructed it will be slightly different from the original. Lossy compression can help to achieve a better compression ratio than lossless compression if the losses are acceptable over the result. This is the standard practice on image compression such as in JPEG and in JPEG2000 formats [7, 10], and video such as MPEG4 format. When a wavelet transform is applied on an image, the resultant coefficients can then be compressed more easily because the information is statistically concentrated in just a few coefficients. Wavelet compression can reach higher compression ratio than other transforms such as the discrete cosine transform suggested for foveated compression in [3]. In wavelet lossy compression, the coefficients that contain the most amount of energy are preserved and the rest are discarded. Selecting such coefficients can be done using the wavelet energy profile and choosing a cutoff frequency.

55.2.4 Foveation

Foveated images are images which have a non-uniform resolution [6]. Results reported in [11] have demonstrated that the human eye shows a form of aliasing from

the fixation point to the edges of the image. Such aliasing increases in a logarithmic rate on all directions. This can be seen as concentric cutoff frequencies from the fixation point. When it is used in a wavelet, this can be expressed as a function [2]:

$$I_0(x) = \int I(t) C^{-1}(x) s\left(\frac{t-x}{w(x)}\right) \quad (55.13)$$

where $I(t)$ is a given image, $I_0(t)$ is the foveated image, $w(x)$ is the weight function. The function s is called the weighted translation of s by x . The function C is defined as:

$$C(x) = \left\| s\left(\frac{-x}{w(x)}\right) \right\| \quad (55.14)$$

There are several weighted translation functions such as the ones defined in [6]. In [3], the suggested weighted functions are the Hamming window (Fig. 55.1a) and the triangular (Fig. 55.1b) window. Such windows offer a smooth degradation from the fixation point. The results on a foveated image from [6] are shown in Fig. 55.2. However, in order to preserve a ROI intact, such windows are not useful. A ROI needs to be left with all its coefficients without cutoff. For well-defined ROIs windows such as Tukey window (Fig. 55.1c) or a truncated triangular window (Fig. 55.1d) can be used [12]. Such windows can be used to define a weighted function where the ratio of a fixation point is bigger than one, leaving the coefficients from the ROI untouched and right after the ROI ends the energy begins to decay in a smooth ratio.

55.3 Proposed Fovea Method

55.3.1 Fovea Window

Fovea compression is expressed through a cutoff window. Ideal cutoff window is a logarithmic function as reported in [13]. Each pixel has a compression rate that decays radially respects to the fovea center. However, such function preserves only the center pixel of the fovea region. In order to create a fovea compression with a defined ROI bigger than one pixel, a fovea window function w is proposed as follows:

$$w(n) = \begin{cases} \ln(n * (e - 1) + 1) & \text{if } a \leq n \leq N \\ 1 & \text{if } n > N \end{cases} \quad (55.15)$$

where N is the radius in pixels of the fovea area, e denotes the Euler number, a is the radius also in pixels of the ROI and $n \in \mathbb{Z}^+$. Given a fovea center $F = (F_x, F_y)$ and a compression ratio interval $[b, L]$, the individual compression ratio $C_b^L(X, Y)$ of a pixel with coordinates $P = (X, Y)$ is calculated as follows:

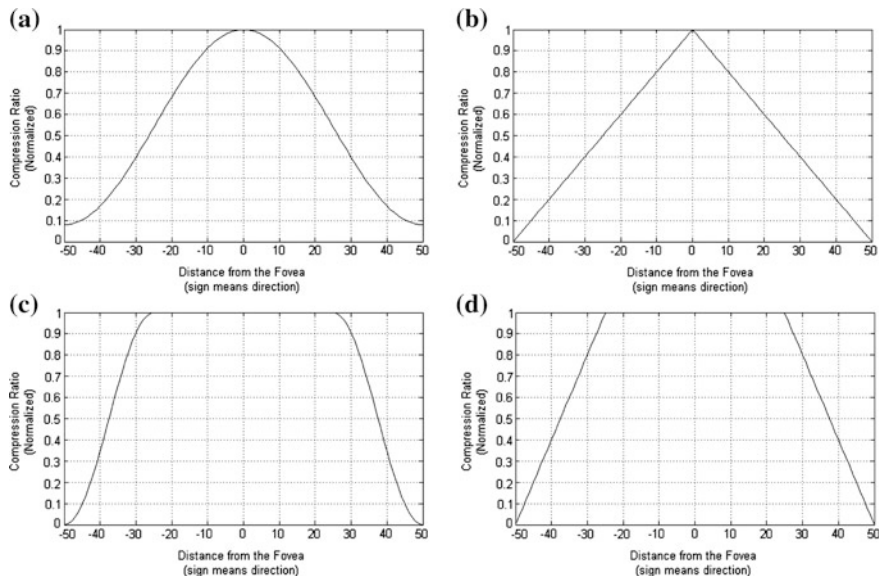


Fig. 55.1 Foveating windows. **a** Hamming window, **b** triangular window, **c** tukey window, **d** truncated triangular window



Fig. 55.2 An image and its wavelet foveated compression. **a** Original gray level image, **b** foveation point at the right eye

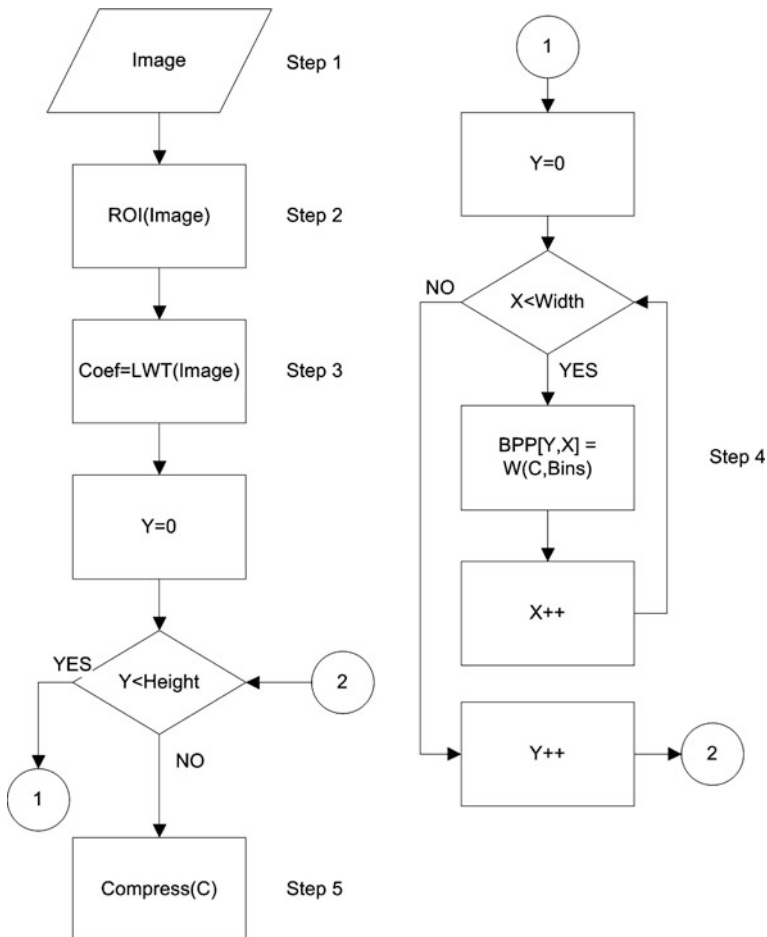


Fig. 55.3 Proposed Fovea compression Algorithm

$$C_b^L(P, F) = w \left(\frac{\|P - F\|}{N} \right) (L - b) + b \quad (55.16)$$

55.3.2 Proposed Algorithm

The proposed algorithm assumes that a method for calculating ROIs is given. The algorithm decomposes the image data into the frequency space using wavelets and the lifting wavelet transform (LWT) [14]. An image compressed through wavelets yields into a better visual quality when reconstructed than classic methods such as

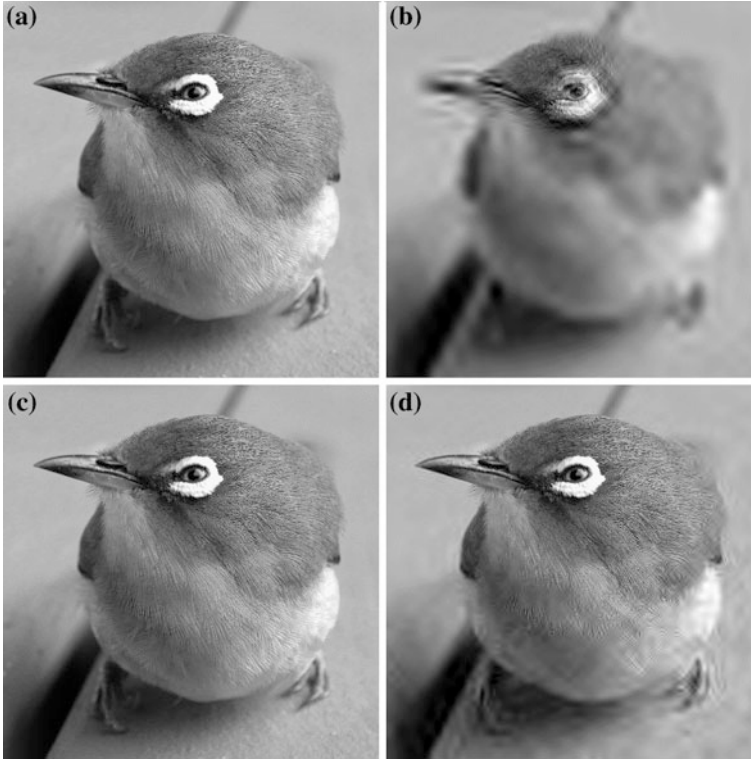


Fig. 55.4 Different foveating windows. **a** Original gray level image, **b** energy profile compression, **c** DCT-tukey window scaling quantization, **d** JPEG Type 1 quantization

the DCT [15]. The proposed algorithm is depicted in Fig. 55.3. The five main steps of the algorithm are:

1. *Image acquisition.*
2. *ROI calculation.*
3. *Wavelet coefficients calculation using the lifting wavelet transform (LWT).*
4. *The compression ratio of each coefficient is calculated.*
5. *A compression method is applied.*

The compression ratio, Eq. (55.16), is applied in step 4 to the wavelet coefficients. It is assumed that the relation of the wavelet coefficient with the pixel coordinates is given by a Quadtree with root at the 0-level of the wavelet decomposition. Given a coefficient of the wavelet located at an i -th level of the decomposition, the compression ratio in Eq. (55.16) can be rewritten as follows:

$$C_b^L(P, \frac{F}{2^i}) = w\left(\frac{2^i \|P - \frac{F}{2^i}\|}{N}\right)(L - b) + b \quad (55.17)$$

Table 55.1 Comparisons between wavelet-based foveated compression and DCT-based compression using the proposed window

| Wavelet | J | Zero coef. percentage | MSE | Quantization |
|---------|---|-----------------------|--------|-------------------|
| dct | – | 0.9383 | 2.2402 | JPEG-Type 3 |
| db7 | 2 | 0.3287 | 0.2534 | Tukey scaling |
| db7 | 4 | 0.3333 | 0.2503 | Tukey scaling |
| db9/7 | 2 | 0.3287 | 0.3279 | Tukey scaling |
| db9/7 | 4 | 0.3333 | 0.2428 | Tukey scaling |
| db7 | 2 | 0.8761 | 3.8959 | JPEG quantization |
| db7 | 4 | 0.9174 | 2.4795 | JPEG quantization |
| db9/7 | 2 | 0.8761 | 3.8167 | JPEG quantization |
| db9/7 | 4 | 0.9174 | 2.3974 | JPEG quantization |
| db7 | 2 | 0.9163 | 2.4192 | Energy profile |
| db7 | 4 | 0.9644 | 6.0816 | Energy profile |
| db9/7 | 2 | 0.9163 | 2.3897 | Energy profile |
| db9/7 | 4 | 0.9644 | 6.0116 | Energy profile |

where $P = (X, Y)$ are the coordinates of the coefficient on the matrix of the sub band of the wavelet decomposition at level i . Notice that each decomposition level has 3 sub bands and the last level of decomposition 4 sub bands [7]. In step 5, several compression methods can be used, namely, energy compression [16], JPEG quantization [17], and DCT-Tukey scaling [18]. Figure 55.4 shows an image and its reconstruction from different compression methods with the fovea point in the eye of the bird and a radius of 60 pixels, and four levels of wavelet decomposition using a Daubechies 7 wavelet.

55.4 Simulation Results

Simulations were carried out using the wavelets Daubechies 7 (db7) and Daubechies 9/7 (db9/7) suggested by the JPEG2000 standard [19] with two levels of decomposition, $J = 2$ for low image distortion using fovea compression and the JPEG200 standard level $J = 4$. To assess the performance of the wavelet-based foveated algorithm the mean squared error (MSE) metric is used. Also, an arbitrary fovea radius used was of 60 pixels. The results are shown in Table 55.1. The DCT-based foveated compression was also realized and it is shown in Table 55.1 as DCT. The compression was realized using the Type 3 JPEG variable quantization compression [8] and the proposed window in Eq. (55.16) as the quantization weight. Table 55.1 shows the percent of coefficients that becomes zero after the compression algorithm is applied to the original image. Figure 55.5 shows the DCT-based foveated image and a wavelet-based foveated image using the Daubechies 9/7 wavelet with four levels of decomposition with energy profile quantization, the proposed window-based scaling quantization and JPEG Type-1 quantization. It should be noted in Table 55.1 that the amount of zeros in the



Fig. 55.5 Foveated image with the proposed window and different compression methods. **a** DCT-based foveated compression, **b** wavelet-based foveated compression with db9/7 and energy profile quantization, **c** wavelet-based foveated compression with db9/7 and tukey window scaling quantization, **d** wavelet-based foveated compression with db9/7 and JPEG Type 1 quantization

coefficients after applying wavelet-based foveated algorithm using energy profile and JPEG Type-1 quantization schemes is lower than using the DCT-based approach. Furthermore, simulation results show that the DCT-based algorithm generates artifacts that make the image fuzzier than the wavelet-based algorithm.

55.5 Conclusion

Wavelet foveation compression offers a very good compression ratio at expenses of controlled losses. The proposed window allows isolate fovea regions over an image by choosing the slope. The fovea window has been applied over different algorithms showing the expected behavior. As stated in [3], applying foveation with wavelets yields into squared artifacts. These artifacts rise as the decomposition

levels increases. However, the DCT also showed a similar behavior in the area outside the ROI as the compression rate increases. With a good model for choosing a ROI, this kind of compression can achieve high compression ratios without losing visual quality over desired areas. Further work will focus on investigating other enhancements of the wavelet-based foveated compression algorithm and comparing with other methods such as the JPEG2000 ROI compression using the maximum shift (Maxshift) method [16].

Acknowledgments The authors gratefully acknowledge the financial support from the CONACYT Mexico and the Puebla State Government under the contract no. 109417.

References

1. N. Kehtarnavaz and M. Gamadia, "Real-Time Image and Video Processing: From Research to Reality," Morgan and Claypool, University of Texas at Dallas, USA, 2006.
2. E. C. Chang and C. K. Yap, "A wavelet approach to foveating images," In *SCG'97: Proceedings of the thirteenth annual symposium on Computational geometry*, New York, NY, USA, 1997, pp. 397–399.
3. S. Lee and A. Bovik, "Fast algorithms for foveated video processing," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.13, No. 2, 2003, pp. 149–162.
4. Guo C, Zhang L (2010) A novel multiresolution spatiotemporal saliency detection model and Its applications in image and video compression. *IEEE Trans Image Process* 19(1):185–198
5. Bogges A, Narcowich FJ (2009) A first course in wavelets with Fourier analysis. 2nd edn. Wiley
6. E. C. Chang, S. Mallat, and C. Yap, "Wavelet Foveation," in *Applied and Computational Harmonic Analysis*, Vol. 9, No. 3, 2000, pp. 312–335.
7. Mallat S (2008) A wavelet tour of signal processing. In: The sparse way, 3rd edn. Academic Press
8. Ahmad J, Raza K, Ebrahim M, Talha U (2009) FPGA based implementation of baseline JPEG decoder. In: Proceedings of the 7th international conference on frontiers of information technology (FIT '09). ACM, New York (Article 29)
9. N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, Vol. C-32, January 1974, pp. 90–93.
10. Bovik AC (2009) The essential guide to image processing. Academic Press
11. B. A. Wandell. *Foundations of Vision*. Sinauer Associates, Inc., 1995.
12. Galan-Hernandez JC, Alarcon-Aquino V, Starostenko O, Ramirez-Cortes JM (2010) Wavelet-based foveated compression algorithm for real-time video processing. *IEEE Electron Robotics Automat Mech Conf (CERMA'10)* pp 405–410
13. Chang E (2000) Wavelet foveation. *Appl Comput Harm Anal* 9(3):312–335
14. C. Jain, V. Chaudhary, K. Jain, S. Karsoliya. "Performance analysis of integer wavelet transform for image compression," *3rd International Conference on Electronics Computer Technology (ICECT'11)*, Vol.3, April 2011, pp.244–246.
15. I. Bocharova, "Compression for Multimedia," 1st ed. Cambridge University Press, New York, NY, USA, 2010.
16. M. Mrak, M. Grgic, and M. Kunt, "High-Quality Visual Experience," Signals and Communication Technology Series, Springer-Verlag, Berlin, 2010.
17. Richter T (2010) Spatial constant quantization in JPEG XR is nearly optimal. Data compression conference (DCC'10), March 2010, pp79–88

18. J. C. Galan-Hernandez, V. Alarcon-Aquino, O. Starostenko, and J. M. Ramirez-Cortes, "Foveated ROI compression with hierarchical trees for real-time video transmission," *In Proceedings of the Third Mexican conference on Pattern recognition (MCPR'11)*, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 240-249.
19. E. J. Balster, B. T. Fortener, W. F. Turri, "Integer Computation of Lossy JPEG2000 Compression," *IEEE Transactions on Image Processing*, Vol. 20, No.8, 2011, pp.2386-2391.

Chapter 56

Energy Aware Data Compression in WSN

Roshanak Izadian and Mohammad Taghi Manzuri

Abstract Wireless sensor networks (WSNs) use energy for their sensing, computation, and communication. To increase the lifetime of the network, sensor nodes are equipped with energy storage devices. Recharging of their batteries is not possible in most applications. Therefore, energy consumption needs to be monitored and limited to increase the energy consumption performance of the network. The communication module uses the highest amount of energy in a WSN. Among several methods offered to reduce the energy consumption, data compression has the highest effect on the energy usage by reducing the number of bits to be broadcasted. This paper illustrates an energy consumption reduction in WSN by compression data. Text data were tested one time with original size and one more time with compressed size with different routing algorithms in Wireless Sensor Network.

56.1 Introduction

Because of swift advancement of semiconductor fabrication technology, electronic devices have been becoming smaller, cheaper, and require less power for its operation. One of the fields, which have benefited by this technological advancement, is the field of Wireless Sensor Networks (WSNs). WSNs can be utilized for data collection in situations such as environmental monitoring, habitat

R. Izadian (✉)

Department of Science and Engineering,
Sharif University of Technology, Tehran, Iran
e-mail: izadian.roshanak@gmail.com

M. T. Manzuri

Department of Computer Engineering,
Sharif University of Technology, Tehran, Iran
e-mail: manzuri@sharif.edu

monitoring, surveillance, structural monitoring, equipment diagnostics, disaster management, and emergency response. In a WSN, nodes are connected by radio frequency, infrared, or other medium without any wire connection [1].

These devices transmit data directly from node to node. If two devices do not have direct connection, intermediate nodes located between them will transmit data packets from the source node to the destination node in a multi-hop routing scheme [2]. Because of their peer-to-peer communication style, no centralized point is required for the network [1]. The centralized node controls a network formation like a base station for a cellular network. The advantage of such a WSN is the formation of an inexpensive network without the need for fixed infrastructure. In addition, nodes can be easily added or removed from the network. Since the formation of a node in such a network can be accomplished easily, the network can be altered intensely. In addition, sensor nodes will run away over a sensor field. Therefore, their position in the field cannot be fixed. The data from sensor nodes is gathered by a sink, which may connect to the outside world thorough internet or satellite [3]. Therefore, a wireless sensory network has several parts to collect, transfer and process data.

A small sensor is equipped with miniature size power supply, sensing and processing unit, and a small transmitter–receiver unit [4]. To form a wireless sensory network, a large number of sensor nodes are being deployed to often hard to reach locations [4]. Therefore, it will not be practical to perform maintenance operations such as changing batteries on deployed sensor nodes [3]. These nodes have limited power supply, bandwidth for communication, processing speed, and memory space. Thus, WSNs, or more specifically each sensor node, are resource restricted devices. One possible way of obtain maximum use of those resource is compressing data on sensor data before transmission [1]. Most often, processing data consumes much less power than transferring data in wireless scheme, therefore, it will be more efficient to perform data compression before transferring data to reduce total power consumption and enhance the lifetime of the sensor node.

In this paper, compression algorithms, such as Run Length Encoding (Lempel–Ziv–Welch) LZW, Adaptive Huffman, Huffman, Shannon-Fano Algorithms will be utilized and their performance in transferring text data will be analyzed in MANET by AODV and DSR routing algorithm. This work is based on the signal processing; therefore by coding algorithm of compression obtain the compressed data. By coding compression algorithms the size of files calculated. All the coding compression is based on the signal processing.

56.2 Data Compression

56.2.1 Run Length Encoding Algorithm

One of the simplest data compression algorithms is Run Length Encoding (RLE). In this algorithm, the sequential of symbols are recognized as runs and non-runs.

RLE handles some sort of redundancy [6], by verifying the existence of reiteration symbols. Sequential repetitive symbols are recognized as runs and all the other sequential are contemplated as non-runs. For instance, in the text “XYXYYYYYZ” the first 3 letters are evaluated as a non-run with length 3, and the next 4 letters are considered as a run with length 4 with a reiteration of symbol Y. The main task of this algorithm is to recognize the runs of the source file, and to record the symbol and the length of each run. The RLE algorithm uses those runs to compress the original source file while holding all the non-runs without utilizing for the compression procedure [5, 7].

56.2.2 Huffman Encoding

Huffman Encoding Algorithms (HE) utilizes the probability of alphabet distribution in the source file to improve the codes for symbols. The frequency of character distributions in the source file is calculated to obtain the probability distribution. Accordingly, shorter code words for higher probabilities and longer code words for smaller probabilities are appointed. In this task, to assign effective code words, a binary tree is produced in which their leaves and paths are formed respect to the calculated probabilities. Two categories of Huffman Encoding have been suggested: Static Huffman Algorithms and Adaptive Huffman Algorithms. Static Huffman Algorithms computes the frequencies and creates a common tree for both the compression and decompression processes [6]. Details of this tree should be stored or transmitted with the compressed file. The Adaptive Huffman algorithms improve the tree while computing the frequencies. There will be two trees in both procedures. In this proposition, a tree is created with the flag symbol in the initialing and is updated as the symbols are understood.

56.2.3 Shannon Fano Algorithm

The Shannon Fano Algorithm is derived from Static Huffman Coding Algorithm. The only distinction is in the production of the code word. All other procedures are similar to Huffman Encoding Algorithm.

56.2.4 Arithmetic Encoding

In this method, a code word is not utilized to typify a symbol of the text. It utilizes a fraction to typify the complete source message [8]. The happening probabilities and the increasing probabilities of a prepared of symbols in the source message are effective into account. The increasing probability range is utilized in both

compression and decompression processes. In the encoding procedure, the increasing probabilities are computed and the limits are production in the initial. While reading the source character by character, the corresponding range of the character within the increasing probability range is chosen. Then the chosen limited is separated into sub sections according to the probabilities of the alphabet. Therefore the following character is read and the comparable sub limited is chosen. In this way, characters are read persistently until the end of the message is encountered. Finally a number should be effective from the last sub limited as the result of the encoding procedure. This will be a fraction in that sub limited. Therefore, the entire source message can be shown utilizing a fraction. To decode the encoded message, the count of characters of the source message and the probability/frequency distribution are required.

56.2.5 The Lempel Zev Welch Algorithm

Dictionary derived compression algorithms are on the basis of creating a dictionary instead of using a statistical model [8]. The Lempel-Zev Welch algorithm (LZW) is a dictionary based algorithm. To compress data, a dictionary is created and utilized to supply and index the preceding string sample. In the compression process, the index values are utilized in place of repetitive string samples. The dictionary is produced dynamically in the compression process and is not required to be transmitted for decompressing.

In the decompression procedure, a similar dictionary is produced dynamically. Hence, this technique is considered as an adaptive compression algorithm.

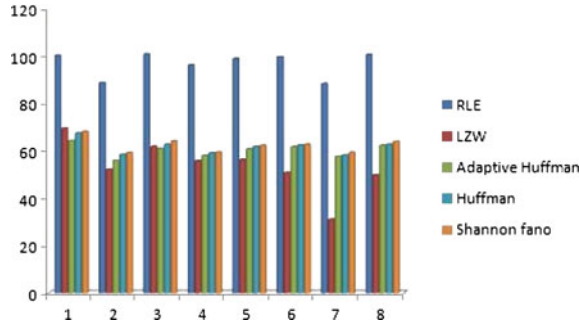
56.3 Measuring Compression Performances

The compression process depends on the redundancy of symbols in the source file. The performance of compression also depends on the kind and the organization of the data. Therefore, it is difficult to measure the performance. In addition, the compression type including the Lossy or the lossless algorithms influences the compression process. Lossy compression algorithms utilize more space and require more time than that of the lossless compression algorithms. Thus, it is difficult to measure the performance of general compression approach.

56.3.1 Comparison of Compression Algorithms

Some lossless compression algorithms are compared to analyze their performance in compression of various size files. Figure 56.1 illustrates the results for the RLE,

Fig. 56.1 Compression source files



LZW, Huffman and Adaptive Huffman. As the figure demonstrates, as the size of the source file increases, the RLE does not provide a suitable file size after compression. RLE is required to declare enormous dictionaries for compression and decompression procedures, it needs a large computing power to process, and it generates large amount of overflow. LZW on the other hand provided the best compression for larger size source files.

56.3.2 Compression Ratio

Compression Ratio is the ratio between the size of the compressed file and the size of the source file. $\text{Compression Ratio} = (\text{size after compression})/(\text{size before compression})$.

Figure 56.2 illustrates compression ratio of the size after using the algorithm. As the figure demonstrates, the compression is more effective for larger size files in algorithms such as LZW. The lower values on the vertical axis show better compression. Because of the need for dictionary files, in some instances, the RLE algorithm generated larger compressed files than the source files. As illustrated in Figs. 56.1 and 56.2, the LZW is a low efficient algorithm, since there is lot of resources required to produce the dictionary file. Compression ratio in LZW is between 32 % and 62 % which is the highest among other compression techniques. Compression ratio in Adaptive Huffman is near 60 thus this algorithm can be contemplated as an effective algorithm. According to the clarification of the code efficiency, utilized code words can be further enhanced. Shannon Fano is other modification of Static Huffman algorithm. Output achieved for this algorithm is given in Shannon Fano algorithm. The compression ratios for Shannon Fano arrive at are in the range of 58–63 % which is delicately equivalent to the preceding algorithm.

Fig. 56.2 Compression ratio of source file

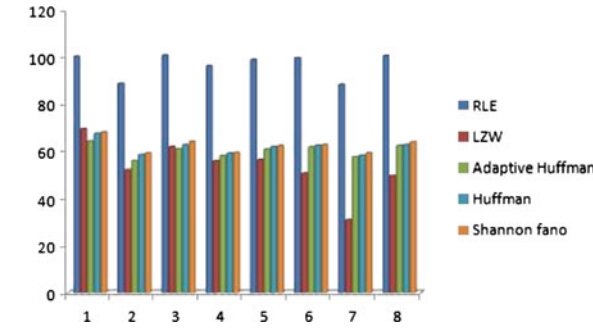


Table 56.1 Wireless network and data transmission parameters

| Parameter of network | Number |
|----------------------|------------------------|
| No. of nodes | 16 |
| Simulation time | 70, 100, 150 s |
| Environment size | 100*100 m ² |
| Transmission range | 40 m |
| Traffic size | 160 byte |
| Packet size | 1060 byte |
| Mobility model | Mesh-Random |
| Antenna type | Omnidirectional |

56.4 Wireless Network and Data Compression Energy Calculations

To calculate the energy of the wireless data transfer system, the energy tocompress the algorithm, and the energy to transfer that data through a wireless network were calculated. Two wireless routing algorithms were considered and the data was transmitted in two algorithms. All parameters of the wireless network are listed in the Table 56.1. The network was consisting of sixteen nodes of Wireless Sensor Network in the environmental of 100*100 m². These nodes were distributed in two cases of fixed (mesh) and random node format. The data transfer was analyzed in different simulation time 70, 100, 150 s. Because the FTP application utilized in this simulation, the pause time is beginning from 10 s to150 s. Traffic size is 160 byte and packet size is 1060 byte that 60 byte for header and 1000 byte for payload, transmission range for simulation 40 m is set and the kind of antenna for nodes is Omnidirectional. Transmission packet is TCP for receiving the Ack in the network. Two wireless routing protocols DSDV and AODV are used in NS2 simulator.

Fig. 56.3 Mesh wireless sensor network



56.4.1 Mesh Network Analysis

In this scenario, the mesh mobility models are shown in Fig. 56.3 in various simulation times. The Node Mica2 is used as a Sensor Network that the energy for sending and receiving is 50.4mW and 28.8mW respectively. Node 0 is source and node 15 is destination or sink.

Using DSDV routing wireless algorithm, imposed higher energy consumption in nodes 9, 10,12. The reason for higher energy consumption is their location that was close to the destination node (sink). This means that more packets has been sent from and been received by these nodes. Energy of data compression has an important effect in the overall energy consumption. Considering the energy of each instruction about 0.5 j, the Adaptive Huffman data compression algorithm consumes 177×10^{-5} j. Among data compression algorithms, the less energy was consumed by Adaptive Huffman in 3 simulations times. The most energy consuming algorithm was RLE for all simulation times. Figure 56.4 illustrates the energy consumption in 16 nodes with different algorithms in 150 s.

56.4.2 Random Network Analysis

The random mobility model is illustrated in Fig. 56.5. In this network, nodes 0, 3 and 12 used more energy rather than the other nodes. The algorithm that consumed low energy was Adaptive Huffman. The highest energy consuming algorithm in 3 time simulations was RLE.

Total energy consumption in the Wireless Sensor Network in two mobility models random and mesh is presented in Table 56.2. As the table demonstrates,

Fig. 56.4 Energy consumption comparison in WSN

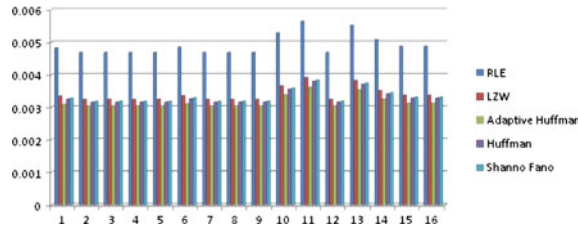


Fig. 56.5 Random wireless sensor network



the energy consumption for the raw data 11,252 byte is 0.43827 j through DSDV routing algorithm.

The Table 56.3 illustrates the various time simulations in Mesh and Random mobility model by AODV routing algorithm.

When AODV routing algorithm was used, the nodes of 0, 3, 12 consumed more energy in mesh random mobility, and 0, 5, 10 nodes consumed more energy than others. In both models of routing, the adaptive Huffman consumed lower compression energy. Energy consumption in AODV was less in mesh mobility model than DSDV in 70 s simulation time. However, less energy was consumed at time 100 s and 150 s through DSDV in mesh mobility model. In 70 s simulation time, Adaptive Huffman algorithm through AODV consumed less energy. However, at time 100 s, 150 s through DSDV Adaptive Huffman consumed less energy. In random mobility model in 70 s, DSDV consumed less energy in network, and at 100 s, 150 s the AODV consumed less in network. Adaptive Huffman consumed less energy in both cases at times 70 s through DSDV, and at 100 s, and 150 s, through AODV.

Table 56.2 Various time simulations and energy consumption (j) in Mesh and Random mobility model by DSDV

| | Time (sec) | RLE | LZW | Adaptive huffman | Huffman | Shannon-fano |
|------------|------------|--------|--------|------------------|---------|--------------|
| Mesh WSN | 70 | 0.4388 | 0.9713 | 0.2827 | 0.2971 | 0.2980 |
| | 100 | 0.1713 | 0.1186 | 0.1115 | 0.1171 | 0.1181 |
| | 150 | 0.0736 | 0.0543 | 0.0520 | 0.0545 | 0.0532 |
| Random WSN | 70 | 0.1866 | 0.1291 | 0.1212 | 0.1238 | 0.1267 |
| | 100 | 0.2134 | 0.1477 | 0.1384 | 0.1454 | 0.14498 |
| | 150 | 0.2279 | 0.1577 | 0.1477 | 0.1551 | 0.1547 |

Table 56.3 Various time simulations and energy consumption (j) in Mesh and Random mobility model by AODV

| | Time (sec) | RLE | LZW | Adaptive huffman | Huffman | Shann-fano |
|------------|------------|--------|--------|------------------|---------|------------|
| Mesh WSN | 70 | 0.1916 | 0.1326 | 0.1245 | 0.1308 | 0.13018 |
| | 100 | 0.2036 | 0.1330 | 0.1248 | 0.1311 | 0.13053 |
| | 150 | 0.1922 | 0.1330 | 0.1322 | 0.1311 | 0.13057 |
| Random WSN | 70 | 0.2031 | 0.1326 | 0.1245 | 0.1308 | 0.13019 |
| | 100 | 0.2035 | 0.1329 | 0.1247 | 0.1310 | 0.13043 |
| | 150 | 0.1919 | 0.1328 | 0.1327 | 0.1309 | 0.13036 |

56.5 Conclusion

Lifetime of WSN is the most important part in networks. This paper illustrated the energy consumption in different compression algorithms in two Wireless Sensor Networks. First, RLE was introduced as a high energy consumption algorithm, and LZW and Adaptive Huffman algorithms were used to compress the data at lower energy consumption rate. The compression ratio in LZW is almost half of the RLE, and the compression ratio of adaptive Huffman is a little more than LZW. If one node was in the path of routing algorithm, the route of node’s antenna is turned on when a packet was received or sent. This imposed higher energy rates. Adaptive Huffman consumed less energy in both AODV and DSDV routing algorithms.

References

1. Kimura N, Latifi Sh (2005) A survey on data compression in wireless sensor networks, International Conference on Information Technology, vol 2. pp 214–219
2. Ergen SC, Varaiya P (2005) On multi-hop routing for energy efficiency, IEEE Commun Lett 9(10):880–881
3. Jolly V, Latifi Sh, Kimura N (1990) Energy-efficient routing in wireless sensor networks based on data reduction, IEEE Trans On Comput 16(10):1150–1163

4. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2002) Wireless sensor networks: a survey, Elsevier Computer Networks 38(4):393–422
5. Pu IM (2006) Fundamental data compression. Elsevier, Britain
6. Belloch E (2002) Introduction to data compression, Computer Science Department, Carnegie Mellon University, Pittsburgh
7. Kesheng W, Otoo, J, Arie S (2006) Optimizing bitmap indices with efficient compression, ACM Trans Database Systems 31:1–38
8. Gawthrop J, Liuping W (2005) Data compression for estimation of the physical parameters of stable and unstable linear systems, Automatica, 41:1313–1321

Chapter 57

Energy Consumption Text and Image Data Compression in WSNs

Roshanak Izadian and Mohammad Taghi Manzuri

Abstract Wireless sensor networks (WSNs) consume energy for their sensing, computation, and communication. To extend the lifetime of the network, sensor nodes are equipped with energy storage devices. Recharging of their batteries is impossible in most applications. Therefore, energy consumption needs to be monitored and limited to extend the high performance operation of the network. In this network, the communication module consumes the highest amount of energy. This paper demonstrates that among several methods offered to reduce the energy consumption, data compression has the highest effect on the energy usage by reducing the number of bits to be broadcasted. To determine the energy efficiency of the communication module, the energy consumption of broadcasting data as text and image in original and compressed forms were measured. Black and white JPEG images and Adaptive Huffman text for both Mesh and random nodes consumed less energy.

57.1 Introduction

Because of swift advancement of semiconductor fabrication technology, electronic devices have been becoming smaller, cheaper, and require less power for their operation. One of the fields, which have benefited by this technological advancement, is the field of Wireless Sensor Networks (WSNs). WSNs can be utilized for data collection in situations such as environmental monitoring, habitat

R. Izadian (✉) · M. T. Manzuri

Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
e-mail: izadian.roshanak@gmail.com

M. T. Manzuri

e-mail: manzuri@sharif.edu

monitoring, surveillance, structural monitoring, equipment diagnostics, disaster management, and emergency response. In a WSN, nodes are connected by radio frequency, infrared, or other medium without any wire connection (1).

These devices transmit data directly from node to node. If two devices do not have direct connection, intermediate nodes located between them will transmit data packets from the source node to the destination node in a multi-hop routing scheme (2). Because of their peer-to-peer communication style, no centralized point is required for the network (1). The centralized node controls a network formation like a base station for a cellular network. The advantage of such a WSN is the formation of an inexpensive network without the need for fixed infrastructure. In addition, nodes can be easily added or removed from the network. Since the formation of a node in such a network can be accomplished easily, the network can be altered intensely. In addition, sensor nodes will run away over a sensor field. Therefore, their position in the field cannot be fixed. The data from sensor nodes is gathered by a sink, which may connect to the outside world thorough internet or satellite (3). Therefore, a wireless sensory network has several parts to collect, transfer and process data.

A small sensor is equipped with miniature size power supply, sensing and processing unit, and a small transmitter–receiver unit (4). To form a wireless sensory network, a large number of sensor nodes are being deployed to often hard to reach locations (4). Therefore, it will not be practical to perform maintenance operations such as changing batteries on sensor nodes (3). These nodes have limited power supply, bandwidth for communication, processing speed, and memory space. Thus, WSNs, or more specifically each sensor nodes, are resource restricted devices. One possible way of obtaining maximum use of those resources is compressing data on sensor data before transmission (1). Most often, processing data consumes much less power than transferring data in wireless scheme, therefore, it will be more efficient to perform data compression before transferring data to reduce total power consumption and enhance the lifetime of the sensor node.

In this paper, compression algorithms, such as Run Length Encoding, LZW (Lempel–Ziv–Welch), Adaptive Huffman, Huffman, Shannon–Fano algorithms will be utilized and their energy consumption performance in transferring text data and BMP, TIFF-LZW, PNG and JPEG with different quality will be analyzed in MANET by using AODV and DSDV routing algorithm. The size of compressed file is obtained and the energy to broadcast the file is calculated to identify the most energy saving data compression and routing algorithms.

57.2 Data Compression for Text Data

57.2.1 Run Length Encoding Algorithm

One of the simplest data compression algorithms is Run Length Encoding (RLE). In this algorithm, the sequential of symbols are recognized as runs and non-runs.

RLE handles some sort of redundancy (6), by verifying the existence of reiteration symbols. Sequential repetitive symbols are recognized as runs and all the other sequential are contemplated as non-runs. For instance, in the text “XYXYYYYYZ” the first three letters are evaluated as a non-run with length 3, and the next four letters are considered as a run with length 4 with a reiteration of symbol Y. The main task of this algorithm is to recognize the runs of the source file, and to record the symbol and the length of each run. The RLE algorithm uses those runs to compress the original source file while holding all the non-runs without utilizing for the compression procedure.

57.2.2 Huffman Encoding

Huffman Encoding Algorithms (HE) utilizes the probability of alphabet distribution in the source file to improve the codes for symbols. The frequency of character distributions in the source file is calculated to obtain the probability distribution. Accordingly, shorter code words for higher probabilities and longer code words for smaller probabilities are appointed. In this task, to assign effective code words, a binary tree is produced in which their leaves and paths are formed respect to the calculated probabilities. Two categories of Huffman Encoding have been suggested: Static Huffman Algorithms and Adaptive Huffman Algorithms. Static Huffman Algorithms computes the frequencies and creates a common tree for both the compression and decompression processes [1]. Details of this tree should be stored or transmitted with the compressed file. The Adaptive Huffman algorithms improve the tree while computing the frequencies. There will be two trees in both procedures. In this proposition, a tree is created with the flag symbol in the initialing and is updated as the symbols are understood.

57.2.3 Shannn Fano Algorithm

The Shannon Fano Algorithm is derived from Static Huffman Coding Algorithm. The only distinction is in the production of the code word. All other procedures are similar to Huffman Encoding Algorithm.

57.2.4 Arithmetic Encoding

In this method, a code word is not utilized to typify a symbol of the text. It utilizes a fraction to typify the complete source message [2]. The happening probabilities and the increasing probabilities of a prepared of symbols in the source message are effective into account. The increasing probability range is utilized in both compression and decompression processes. In the encoding procedure, the

increasing probabilities are computed and the limits are production in the initial. While reading the source character by character, the corresponding range of the character within the increasing probability range is chosen. Then the chosen limited is separated into sub sections according to the probabilities of the alphabet. Therefore the following character is read and the comparable sub limited is chosen. In this way, characters are read persistently until the end of the message is encountered. Finally a number should be effective from the last sub limited as the result of the encoding procedure. This will be a fraction in that sub limited. Therefore, the entire source message can be shown utilizing a fraction. To decode the encoded message, the count of characters of the source message and the probability/frequency distribution are required.

57.2.5 The Lempel Zev Welch Algorithm

Dictionary derived compression algorithms are on the basis of creating a dictionary instead of using a statistical model [2]. The Lempel Zev Welch algorithm (LZW) is a dictionary based algorithm. To compress data, a dictionary is created and utilized to supply and index the preceding string sample. In the compression process, the index values are utilized in place of repetitive string samples. The dictionary is produced dynamically in the compression process and is not required to be transmitted for decompressing.

In the decompression procedure, a similar dictionary is produced dynamically. Hence, this technique is considered as an adaptive compression algorithm.

57.3 Data Compression for Image Data: Lossless Image Formats

57.3.1 Bitmap

BMP (bitmap) is a bitmapped graphics format utilized intellectually by the Microsoft Windows graphics subsystem (GDI), and utilized prevalently as a simple graphics file format on that aims. It is an uncompressed format.

57.3.2 Portable Network Graphics

PNG Portable Network Graphics is a bitmap image format that operates lossless data compression and the history of this format come back to 1996. PNG was produced to both recuperate upon and substitute for the GIF format with an image file format that does not necessitate a patent license to use. It uses the DEFLATE

compression algorithm, that utilizes a union of the LZ77 algorithm and Huffman coding. PNG supports palette based (with a palette defined in terms of the 24 bit RGB colors), grayscale and RGB images. PNG was intended for transport of images on the internet not for professional graphics and as such other color spaces. This format is very good for images with big areas of one unique color, or with small variations of color.

57.3.3 Tagged Image File Format

Tiff is a file format for mainly storing images, involving photographs and line art. It is one of the most common and variable of the prevalent file formats. Originally created by the company Aldus, jointly with Microsoft, for utilize with PostScript printing, TIFF is a popular format for high color depth images, in parallel with JPEG and PNG. TIFF format is widely supported by image- usage requests, and by scanning, faxing, word processing, visional character identification, and other requests.

57.4 Lossy Image Formats

57.4.1 Joint Photographic Experts Group

JPEG is an algorithm intended to compress images with 24 bits depth or grayscale images. It is a lossy compression algorithm. One of the characteristics that make the algorithm very supple is that the compression rate can be adapted. If we compress a lot, more information will be lost, but the result image size will be smaller. With a smaller compression rate we acquire better quality, but the size of the culminate in image will be bigger. This compression be composed of in making the coefficients in the quantization matrix bigger when we want more compression, and smaller when we want less compression. JPEG is the most utilizes format for storing and transmitting images in Internet.

57.5 Measuring Compression Performances

The compression process depends on the redundancy of symbols in the source file. The performance of compression also depends on the kind and the organization of the data. Therefore, it is difficult to measure the performance. In addition, the compression type including the Lossy or the lossless algorithms influences the compression process. Lossy compression algorithms utilize more space and

require more time than that of the lossless compression algorithms. Thus, it is difficult to measure the performance of general compression approach.

57.5.1 *Compression Algorithms*

Some lossless compression algorithms are compared to analyze their performance in compression of various size files in text data. Figure 57.1 illustrates the results for the RLE, LZW, Huffman and Adaptive Huffman. As the figure demonstrates, as the size of the source file increases, the RLE does not provide a suitable file size after compression. RLE is required to declare enormous dictionaries for compression and decompression procedures, it needs a large computing power to process, and it generates large amount of overflow. LZW on the other hand provided the best compression for larger size source files.

Some lossless and lossy compression algorithms are compared to analyze their performance in compression of various size files in image data. Figure 57.2 illustrates the results for the BMP or original data in comparison of BMP, TIFF and JPEG in one for color image and two for black and white images. PNG is a good algorithm for color image and JPEG for black and white image. TIFF and JPEG for color image operate as same as each other and compress size of data almost half of the original size. However TIFF in black and white images compress data near the original size [3–8] .

57.5.2 *Compression Ratio*

Compression Ratio is the ratio between the size of the compressed file and the size of the source file.

$$\text{Compression Ratio} = (\text{size after compression})/(\text{size before compression}).$$

Figure 57.3 illustrates compression ratio of the size after using the algorithm. As the figure demonstrates, the compression is more effective for larger size files in algorithms such as LZW. The lower values on the vertical axis show better compression. Because of the need for dictionary files, in some instances, the RLE algorithm generated larger compressed files than the source files. As illustrated in Figs. 57.1 and 57.2, the LZW is a low efficient algorithm, since there is lot of resources required to produce the dictionary file. Compression ratio in LZW is between 32 and 62 % which is the highest among other compression techniques. Compression ratio in Adaptive Huffman is near 60 thus this algorithm can be contemplated as an effective algorithm. According to the clarification of the code efficiency, utilized code words can be further enhanced. Shannon Fano is other modification of Static Huffman algorithm. Output achieved for this algorithm is given in Shannon Fano algorithm. The compression ratios for Shannon Fano arrive

Fig. 57.1 Compression source files for text data

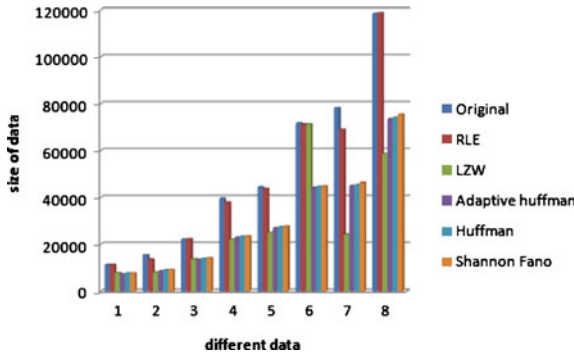


Fig. 57.2 Compression source files for image data

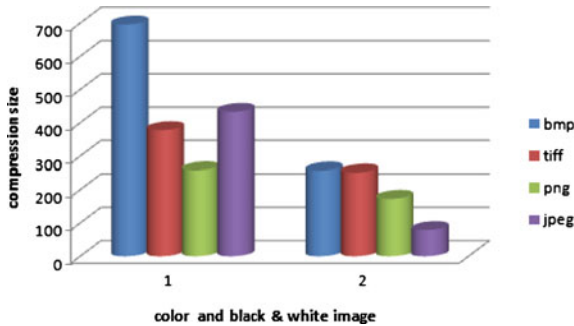
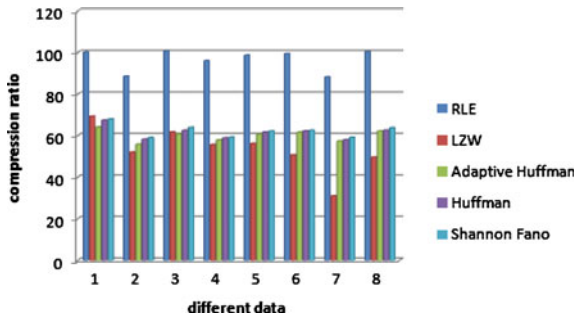


Fig. 57.3 Compression ratio of source file for text data



at are in the range of 58–63 % which is delicately equivalent to the preceding algorithm [9]. In Fig. 57.4 JPEG in one color image and two black and white images operates better than the others algorithms. PNG algorithm is near 1/3 less than the TiFF algorithm and jpeg in color images operate very well it reduces near 1/6 and in black and white 1/3 less than Tiff algorithm. JPEG is effective for big compression ratio and good for photographic image; it operates well for black and white image but not as well as color images, in addition the black and white images has a lot of details in spite each pixel has the eight bits.

Fig. 57.4 Compression ratio of source File Image Data

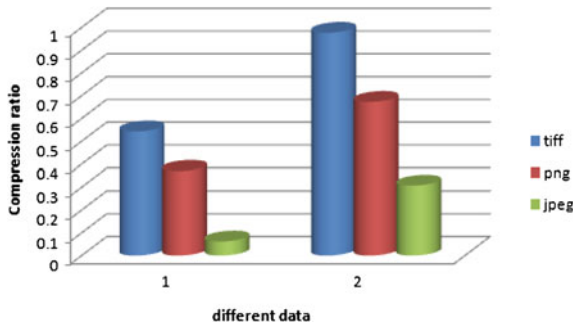


Table 57.1 Wireless network and data transmission parameters

| Parameter of network | Number |
|----------------------|------------------------|
| Number of nodes | 16 |
| Simulation time | 70, 100, 150 s |
| Environment size | 100*100 m ² |
| Transmission range | 40 m |
| Traffic size | 160 byte |
| Packet size | 1060 byte |
| Mobility model | Mesh-Random |
| Antenna type | Omnidirectional |

57.6 Energy Consumptions

To calculate the energy of the wireless data transfer system, the energy to compress the algorithm, and the energy to transfer that data through a wireless network were calculated. Two wireless routing algorithms were considered and the data was transmitted in two algorithms. All parameters of the wireless network are listed in the Table 57.1. The network was consisting of sixteen nodes of Wireless Sensor Network in the environmental of 100*100 m². These nodes were distributed in two cases of fixed (mesh) and random node format. The data transfer was analyzed in different simulation time 70, 100, 150 s. Because the FTP application utilized in this simulation, the pause time is beginning from 10 to 150 s. Traffic size is 160 byte and packet size is 1,060 byte that 60 byte for header and 1,000 byte for payload, transmission range for simulation 40 m is set and the kind of antenna for nodes is Omnidirectional. Transmission packet is TCP for receiving the ack in the network. Two wireless routing protocols DSDV and AODV are used in NS2 simulator.

57.6.1 Mesh Network Analysis

In this scenario, the mesh mobility models are shown in Fig. 57.5 in various simulation times. The Node Mica2 is used as a Sensor Network that the energy for

Fig. 57.5 Mesh wireless sensor network



Table 57.2 Various time simulations and energy consumption (j) in Mesh and Random mobility model by DSDV for text data

| | Time (s) | RLE | LZW | Adaptive Huffman | Huffman | Shannon–Fano |
|------------|----------|--------|--------|------------------|---------|--------------|
| Mesh WSN | 70 | 0.4388 | 0.9713 | 0.2827 | 0.2971 | 0.2980 |
| | 100 | 0.1713 | 0.1186 | 0.1115 | 0.1171 | 0.1181 |
| | 150 | 0.0736 | 0.0543 | 0.0520 | 0.0545 | 0.0532 |
| Random WSN | 70 | 0.1866 | 0.1291 | 0.1212 | 0.1238 | 0.1267 |
| | 100 | 0.2134 | 0.1477 | 0.1384 | 0.1454 | 0.14498 |
| | 150 | 0.2279 | 0.1577 | 0.1477 | 0.1551 | 0.1547 |

sending and receiving is 50.4 and 28.8 mW, respectively. Node 0 is source and node 15 is destination or sink.

Using DSDV routing wireless algorithm, imposed higher energy consumption in nodes 9, 10, 12. The reason for higher energy consumption is their location that was close to the destination node (sink). This means that more packets has been sent from and been received by these nodes. Energy of data compression has an important effect in the overall energy consumption. Considering the energy of each instruction about 0.5 j, the Adaptive Huffman data compression algorithm consumes 177×10^{-5} j. Among data compression algorithms, the less energy was consumed by Adaptive Huffman in 3 simulations times. The most energy consuming algorithm was RLE for all simulation times. Figure 57.6 illustrates the energy consumption in 16 nodes with different algorithms in 150 s.

57.6.2 Random Network Analysis

The random mobility model is illustrated in Fig. 57.7. In this network, nodes 0, 3 and 12 used more energy rather than the other nodes. The algorithm that consumed low energy was Adaptive Huffman. The highest energy consuming algorithm in 3 time simulations was RLE.

Fig. 57.6 Energy consumption in text data comparison in WSN

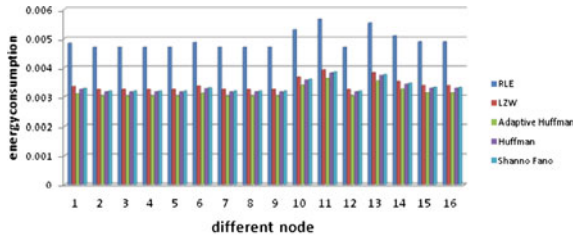


Fig. 57.7 Random wireless sensor network



Table 57.3 Various time simulations and energy consumption (j) in Mesh and Random mobility model by AODV for text data

| | Time (s) | RLE | LZW | Adaptive Huffman | Huffman | Shann–Fano |
|------------|----------|--------|--------|------------------|---------|------------|
| Mesh WSN | 70 | 0.1916 | 0.1326 | 0.1245 | 0.1308 | 0.13018 |
| | 100 | 0.2036 | 0.1330 | 0.1248 | 0.1311 | 0.13053 |
| | 150 | 0.1922 | 0.1330 | 0.1322 | 0.1311 | 0.13057 |
| Random WSN | 70 | 0.2031 | 0.1326 | 0.1245 | 0.1308 | 0.13019 |
| | 100 | 0.2035 | 0.1329 | 0.1247 | 0.1310 | 0.13043 |
| | 150 | 0.1919 | 0.1328 | 0.1327 | 0.1309 | 0.13036 |

Total energy consumption in the Wireless Sensor Network in two mobility models random and mesh is presented in Table 57.2. As the table demonstrates, the energy consumption for the raw data 11,252 byte is 0.43827 j through DSDV routing algorithm.

The Table 57.3 illustrates the various time simulations in Mesh and Random mobility model by AODV routing algorithm for text data.

When AODV routing algorithm was used, the nodes of 0, 3, 12 consumed more energy in mesh random mobility, and 0, 5, 10 nodes consumed more energy than others. In both models of routing, the adaptive Huffman consumed lower compression energy. Energy consumption in AODV was less in mesh mobility model than DSDV in 70 s simulation time. However, less energy was consumed at time 100 and 150 s through DSDV in mesh mobility model. In 70 s simulation

Table 57.4 Various time simulations and energy consumption (j) in Mesh and Random mobility by DSDV for color image

| | Time (s) | BMP | TiFF | PNG | JPEG |
|------------|----------|--------|------|-------|--------|
| Mesh WSN | 70 | 12.12 | 6.58 | 4.49 | 0.754 |
| | 100 | 31.05 | 7.67 | 5.23 | 0.878 |
| | 150 | 6.46 | 2.39 | 2.39 | 0.402 |
| Random WSN | 70 | 11.84 | 6.43 | 4.38 | 0.736 |
| | 100 | 11.873 | 6.44 | 4.401 | 0.7386 |
| | 150 | 11.876 | 6.45 | 4.402 | 0.7386 |

Table 57.5 Various time simulations and energy consumption (j) in Mesh and Random mobility by AODV for color image

| | Time (s) | BMP | TiFF | PNG | JPEG |
|------------|----------|-------|-------|-------|--------|
| Mesh WSN | 70 | 11.52 | 6.26 | 4.27 | 0.717 |
| | 100 | 13.18 | 7.16 | 4.88 | 0.820 |
| | 150 | 14.07 | 7.64 | 5.21 | 0.875 |
| Random WSN | 70 | 11.84 | 6.431 | 4.389 | 0.7367 |
| | 100 | 11.86 | 6.44 | 4.397 | 0.7380 |
| | 150 | 11.85 | 6.439 | 4.395 | 0.7376 |

Table 57.6 Various time simulations and energy consumption (j) in Mesh and Random mobility by DSDV for black white image

| | Time (s) | BMP | TiFF | PNG | JPEG |
|------------|----------|-------|-------|-------|-------|
| Mesh WSN | 70 | 4.478 | 4.37 | 3.014 | 1.376 |
| | 100 | 5.215 | 5.093 | 3.51 | 1.603 |
| | 150 | 2.387 | 2.332 | 1.644 | 0.733 |
| Random WSN | 70 | 4.372 | 4.270 | 2.943 | 1.344 |
| | 100 | 4.691 | 4.281 | 2.951 | 1.347 |
| | 150 | 4.385 | 4.282 | 2.952 | 1.348 |

Table 57.7 Various time simulations and energy consumption (j) in Mesh and Random mobility by AODV for black white image

| | Time (s) | BMP | TiFF | PNG | JPEG |
|------------|----------|--------|--------|-------|-------|
| Mesh WSN | 70 | 4.256 | 4.157 | 2.865 | 1.308 |
| | 100 | 12.666 | 12.666 | 12.37 | 1.496 |
| | 150 | 5.1985 | 5.077 | 3.49 | 1.598 |
| Random WSN | 70 | 4.372 | 4.270 | 2.943 | 1.344 |
| | 100 | 4.380 | 4.278 | 2.984 | 1.346 |
| | 150 | 4.378 | 4.276 | 2.947 | 1.345 |

time, Adaptive Huffman algorithm through AODV consumed less energy. However, at time 100, 150 s through DSDV Adaptive Huffman consumed less energy. In random mobility model in 70 s, DSDV consumed less energy in

network, and at 100, 150 s the AODV consumed less in network. Adaptive Huffman consumed less energy in both cases at times 70 s through DSDV, and at 100, and 150 s, through AODV. Energy consumption for Color image Tables 57.3, 57.4 in JPEG algorithm in all simulation times was less than the TIFF and PNG in both Mesh and Random nodes in both DSDV and AODV routing algorithms. However, JPEG algorithm for black and white Tables 57.5, 57.6 images for both Mesh and random nodes consumed less energy consumption for both routing algorithm. JPEG compresses the size of color image near 57 % of the original image size, but it compresses black and white image about 27 % of its original image size. Therefore JPEG operates well for both image and black and white image but it compresses black and white image better than color image (Table 57.7).

In Tables 57.4 and 57.5, respectively energy consumption for color image by DSDV and AODV routing algorithms in mesh and Random nodes are illustrated.

57.7 Conclusion

This paper illustrated the energy consumption using various data compression and routing algorithms Wireless Sensor Networks. First, RLE was introduced as a high energy consumption algorithm, and LZW and Adaptive Huffman algorithms were used to compress the data at lower energy consumption rate. The compression ratio in LZW was almost half of the RLE, and the compression ratio of adaptive Huffman was a little more than LZW. If one node was in the path of routing algorithm, the route of node's antenna was turned on when a packet was received or sent. This imposed higher energy rates. Adaptive Huffman consumed less energy in both AODV and DSDV routing algorithms.

Energy consumption for Color image in JPEG algorithm in all simulation times was less than the TIFF and PNG in both Mesh and Random nodes in both DSDV and AODV routing algorithms. However, JPEG algorithm for black and white image for both Mesh and random nodes consumed less energy consumption for both routing algorithm.

References

1. Belloch E (2002) Introduction to data compression. Computer Science Department, Carnegie Mellon University
2. Gawthrop J, Liuping W (2005) Data compression for estimation of the physical parameters of stable and unstable linear systems. *Automatica* 41:1313–1321
3. Kimura N, Latifi Sh (2005) A survey on data compression in wireless sensor networks. In: International conference on information technology
4. Ergen SC, Varaiya P (2005) On multi-hop routing for energy efficiency. *IEEE Commun Lett* 9(10):880–881

5. Jolly V, Latifi SH, Kimura N (2011) Energy-efficient routing in wireless sensor networks based on data reduction. *IEEE Trans Computv* 8(2/3):169–185
6. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2002) Wireless sensor networks: a survey. *Elsevier Comput Netw* 38(4):393–422
7. Pu IM (2006) *Fundamental data compression*. Elsevier, Amsterdam
8. Kesheng W, Otoo J, Arie S (2006) Optimizing bitmap indices with efficient compression. *ACM Trans Database Syst* 31:1–38

Chapter 58

New QoS Framework for Mobile Ad hoc Networks Based on the Extension of Existing QoS Models

Peter Magula and Margaréta Kotočová

Abstract With the expansion of real-time applications for mobile ad hoc networks the need for quality of service (QoS) support has become essential. Providing QoS in this kind of networks is a big challenge requiring a complex QoS model or framework. This paper deals with quality of service models in mobile wireless ad hoc networks, provides a brief overview of existing models and describes a new QoS framework based on the extension of existing QoS models. Proposed QoS framework increases service differentiation in the network and provides them various QoS levels. It also improves one-way delay as a QoS network parameter. We validated our proposal by means of network simulation. This paper also presents simulation scenarios and evaluation details.

58.1 Introduction

Mobile wireless ad hoc networks (MANETs) have become very popular nowadays due their simplicity by using and the ubiquity of wireless technology. MANET is a set of wireless network nodes creating a dynamic ad hoc topology without any central control, administration of fixed infrastructure. In this kind of network a node can act as a host and also as a router at the same time. As the growth of multimedia applications continues and the requirements for real traffic increase,

P. Magula (✉) · M. Kotočová
Informatics and Information Technologies, Slovak University
of Technology, Bratislava, Slovakia
e-mail: magula@fiit.stuba.sk

M. Kotočová
e-mail: kotocova@fiit.stuba.sk

there is a big challenge for researchers to find technologies and approaches to satisfy these requirements. In the field of wireless ad hoc networking this is more challenging due to the dynamic network topology and bandwidth constraint.

Generally, the goal of QoS in communication networks is to achieve a more deterministic network behaviour so that data carried by the network can be better delivered and the network resources can be better utilized [1]. In wired networks, there are many approaches, techniques and protocols used to achieve this goal. Also, there exist QoS models which are complex frameworks widely accepted and used to satisfy required QoS level for chosen application. Typically, QoS is measured or represented by set of qualitative network performance metrics such as bandwidth, one-way network delay, jitter (delay variance) and packet loss [2, 3]. There are applications sensible to one or two metrics with requirements to minimum or maximum level of certain metric. Unlike wired networks, MANETs have specific characteristics which make QoS provisioning more difficult [4]. In particular, it is a dynamic network topology, bandwidth constraint, power issues, and wireless medium. Due to these special attributes of MANET network, we can define some other QoS metrics, e.g. service coverage area or power consumption [5–7]. In order to satisfy QoS requirements, there is a need to have a complex framework or QoS model which can provide basic architecture of QoS provisioning. There should be the cooperation among all QoS components, e.g. routing, queuing, packet scheduling, admission control, signalling and traffic engineering [8–10]. In wired networks there are two QoS models widely used: IntServ (Integrated Services) providing hard QoS, but with low scalability, and DiffServ (Differentiated Services) used in the Internet [4]. Unfortunately, both are not suitable for MANETs due to their specific characteristics. Several QoS models for MANETs have been proposed recently. In the next section, we cover most used ones and discuss their advantages and disadvantages. The proposed extension of SWAN model is in the Sect. 58.3 with our simulation results and the comparison with SWAN model in Sect. 58.4. Finally, the last section contains conclusion and description of some open issues in this area and our future plans.

58.2 Related Work

58.2.1 Flexible QoS Model for MANETs

Flexible QoS Model for MANETs (FQMM) is the first QoS model designed for MANETs. It combines the advantages of IntServ and Diffserv models and provides a hybrid scheme of per-flow provisioning as in IntServ and per-class provisioning as in DiffServ. FQMM operates at the IP layer with the cooperation with Medium-Access layer. It is divided into data forwarding and control plane. The main purpose of data forwarding plane is to classify incoming packets going through traffic conditioner and packet scheduler. The control plane handles preparation for

data forwarding operation with specific protocols and algorithms cooperation. This model defines three categories of nodes: ingress, interior and egress node. This nodes differentiation is borrowed from DiffServ model from wired networks. Ingress node is a source node sending data to destination. Interior nodes are nodes forwarding data to other nodes according to some routing decisions. And lastly, egress node is actually the destination node. Interior nodes forward data packets by certain PHB (Per Hop Behaviour) according to the Diffserv field in the packet header. We can look at MANET as one DiffServ domain bounded with the ingress and egress nodes [11]. It is important to note that due to the mobility of nodes in MANETs, the nodes can have different roles as they move. FQMM can provide per flow QoS provisioning for high-priority flows. The question is how many high-priority flow sessions can coexist at the same time in the network. Another open issue is the scheduling performed by intermediate nodes.

58.2.2 Integrated Mobile Ad hoc QoS Framework

The Integrated Mobile Ad hoc QoS framework (iMAQ) is another QoS model or framework for MANETs [12]. The main idea of this model is based on a cross-layer communication approach involving network and application layer by means of so called middleware service layer.

As nodes are mobile, the network can become partitioned which leads to missing data. Predictive location-based QoS routing protocol with middleware layer cooperation can predict network partitioning and provide necessary information to the application layer. Thus the main role of middleware layer is to replicate data among different network groups in order to provide better data accessibility before network partitioning occurs. The disadvantage of this QoS model is its high overhead and lack of resource reservation [4].

58.2.3 INSIGNIA Model

INSIGNIA model represents the first signalling framework designed for MANETs. It is based on in-band signalling approach. That means that control information is carried in data packets along the same communication path in contrast to out-of-band signalling approach where control data are carried separately in control packets sometimes even along different path than data packets. In wired networks, Resource reSerVation Protocol (RSVP) is used as a standard for resource reservation and management. RSVP is an example of out-of-band signalling. For MANETs in general, this kind of signalling is not very suitable because it consumes network bandwidth. Thus, it is a better idea to use in-band signalling which consumes less bandwidth and if control overhead is simple the information can be carried in each packet.

INSIGNIA signalling framework uses the options part of IP packets within all control information is carried. For each active flow in the network there is a soft state stored in all related hosts. The soft state is periodically refreshed every time when packets from the particular flow arrive at the hosts or are forwarded by the host to their destination. INSIGNIA, with admission control cooperation, reserves network resources, mainly available bandwidth, to the particular flow if the resource requirement coming from the source node can be satisfied. In order to keep INSIGNIA very simple and to not conserve much bandwidth, there are no error messages and thus no negative notification among network nodes. For example, if the resource requirement request cannot be satisfied, no error message is sent to source node. Due to the dynamic topology of MANETs, INSIGNIA needs to respond fast to the topology changes. It is done by periodical informing the source node with the status of the data flow. The destination node gathers statistical information such as throughput and loss rate and sends the report to the source node. With this kind of feedback, the source node can adapt the transmission of data packet belonging to the particular flow. Due to these attributes of INSIGNIA, it can provide assured adaptive QoS provisioning to real-time flows based on the source node's requirements and resource availability in the MANETs [13]. On the other hand, the research shows that INSIGNIA has problems with scalability since state information about data flows is stored in network nodes. Another QoS framework making use of INSIGNIA and TORA routing protocol is INORA. TORA provides multiple routes between a given source and destination, and together with INSIGNIA signalling, provide QoS requirements for a flow. INORA also combines congestion control with routing [14].

58.2.4 Design of an Efficient QoS Architecture Model

Unlike the previous model, Design of an Efficient QoS Architecture (DEQA) model is very scalable and stable [15]. It consists of three parts. The first one is routing protocol that search several parallel communication paths. Data packet are fragmented in source node and sent along these paths independently to the destination where they are assembled again. The goal of such a routing approach is stability increase. The next part of a model is admission control with definition of two thresholds, minimum and maximum. If the incoming QoS requirement is under the maximum limit, it is denied. On the other hand, if the requirement is below the minimum, it is allowed. But in case the QoS requirement is between minimum and maximum, the probe packet has to be sent along the communication path to the destination in order to get information about available network resources. Then, based on this information, the request is either allowed or denied. The last part of a model is congestion control which is important after admission control part allows the data flow request. After that congestion control block periodically monitors whether communication links begin to be congested. If so, Explicit Congestion Notification (ECN) technique is used with the goal to decrease

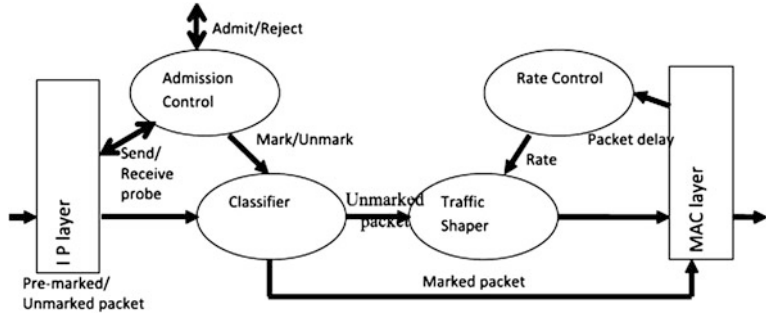


Fig. 58.1 SWAN model [19]

transmission rate of the network traffic that do not require QoS provisioning. Some authors discuss that this behaviour is the biggest disadvantages of the model because there can be a strong unbalance between best-effort traffic and traffic requiring some level of QoS.

58.2.5 Cross Layer Interactions and Service Mapping QoS Model

The main idea of Cross Layer Interactions and Service Mapping (CLIASM) model is to build a shared database with data from all protocols of the network model with information about QoS [16]. Thus, it is another example of cross-layer QoS model. The goal is that each layer has the same information about the network itself, network performance parameters and available resources. There are four groups of such data corresponding to the layers of network model: application, transport, network, and link layer data. The nodes along communication path are not involved in QoS provisioning. The whole overhead is managed by source node including state information handling. This can lead sometimes to unwanted interactions and stability issues with the overall performance decrease.

58.2.6 Stateless Wireless Ad hoc Network QoS Model

The last QoS model described in this paper is called Stateless Wireless Ad hoc Network (SWAN) model. It is a distributed network QoS model with stateless approach using rate control for UDP and TCP best-effort traffic based on AIMD (Additive Increase Multiplicative Decrease) [17]. Like DEQA model, it also uses ECN (Explicit Congestion Notification) to regulate real-time traffic in order to react dynamically to topology changes. Figure 58.1 describes the architecture of SWAN model.

The two main functional block of SWAN model are Classifier and Traffic Shaper which both operate between IP and MAC layer. The role of Classifier is to distinguish real-time traffic that should not go through Shaper. The traffic shaper in this model is represented by simple Leaky bucket shaper which is used to shape best-effort traffic based on the information from Rate Controller in order to delay best-effort packets and thus provide more bandwidth to real-time traffic.

Admission Controller is a block located at source node. Its function is to send a probe request toward the destination node to estimate resources availability. Based on this information, Admission Control module decides whether admit or reject the request. The advantage of SWAN is that all nodes regulate best-effort traffic independently and each source node uses admission control for real-time sessions [18]. When a new real-time flow is allowed by admission control block, all packets, belonging to the particular flow, are marked as a real-time packets. Due to this marking, classifier bypasses shaper and packets remain unregulated [19].

The fact that SWAN is a stateless model and thus it does not require maintaining information at network nodes makes it very scalable and robust QoS model solution for MANETs. The lack of reservation and signalization mechanism means that this QoS model is not suitable for hard QoS provisioning but it was not the design goal of this model [20].

58.3 An Extension of SWAN QoS Model

As stated above, Stateless Wireless Ad hoc Network QoS model is suitable for dynamic MANET topologies. It provides soft QoS in a scalable and robust manner by means of distributed network approach with traffic rate control. SWAN model [17].

We consider the ability of the model to differentiate only between two types of traffic as a drawback. Typically, there is a need to provide service differentiation in a more precise way than only real-time traffic and best-effort traffic. In many scenarios, real-time traffic needs to be differentiated according to various parameters, e.g. priority. Therefore, this paper proposes an extension to Stateless Wireless Ad hoc Network QoS model with a scheduling module and rate control improvement. The architecture of our proposal is illustrated in Fig. 58.2.

The scheduling module has been added to the former SWAN model, between Classifier and Medium Access Control block. Then, the functionality of SWAN model has been modified in the following manner.

If Admission Controller admits the request, Classifier differentiates packets according to their marking to five classes: Platinum, Gold, Silver, Bronze, and Best-effort. Then, packets are queued in respective queues and wait for the transmission. There is a special queue for best-effort traffic which can be shaped by traffic shaper, based on the information from rate controller, in a similar way like in the former SWAN model. The scheduling algorithm is a combination of Strict Queuing, Weighted Fair Queuing and Probability called Probabilistic Priority

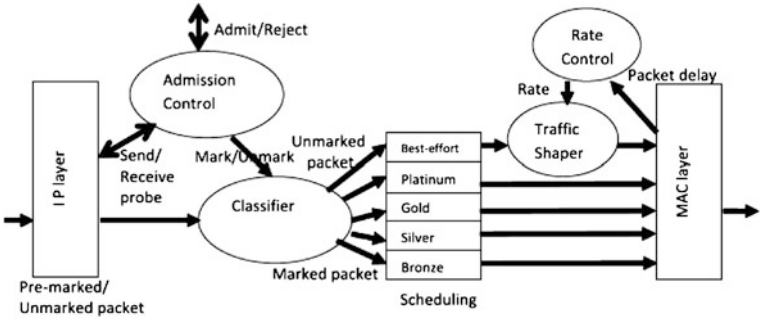


Fig. 58.2 Proposed extension of SWAN model

Queuing proposed in [13]. Each queue has a parameter p which is a probability with which a particular queue is served. The basic algorithm is as follows:

1. If queue i is not empty, and if all other queues with higher priority than queue i are empty, transmit one packet from queue i with probability 1.
2. If at least one higher priority queue is not empty, generate a random number between 0 and 1.
3. If random number from step 2 is higher or equal than p , transmit one packet from queue i , else go to step 4.
4. $i =$ the next queue and continue with step 1.

Next section presents discussion and simulation results of proposed SWAN extension.

58.4 Simulation and Evaluation

We simulated several scenarios using network simulator ns-2 to validate our proposed solution and compare its performance with the former SWAN model [21]. Network simulator ns-2 was chosen due to its extensive support for MANETs and an existing implementation of SWAN model. We used SWAN implementation for ns-2 from [18] and Probabilistic Priority Scheduling [22]. We performed many experiments with various scenarios and simulation parameters. Based on the evaluation of these simulation results we can argue that our proposed extension of an existing QoS model does not depend on different simulation scenarios or network parameters. In this paper, we provide results based on simulation process with typical network scenario with following simulation parameters:

- Simulation area: 500×500 m
- Number of nodes: 20
- Nodes mobility: Random Waypoint Model - RWM
- Node velocity: 0–5 m/s

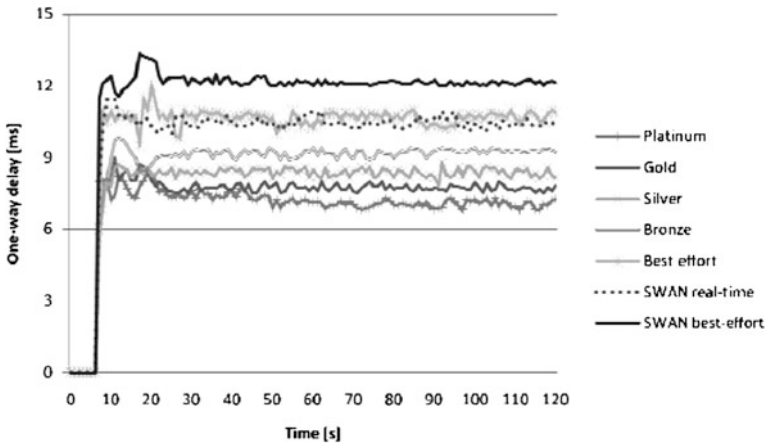


Fig. 58.3 One-way delay comparisons

- Routing protocols: AODV, DSDV
- MAC protocol: IEEE 802.11
- Simulation time 120 s

Source communication nodes generated five types of data flows with fixed packet size. In our proposed extension model, we classified these flows from Platinum to best-effort traffic, and in SWAN model, we classified them only to two groups, real-time and best-effort. Figure 58.3 shows simulation results in order to compare network performance of both models.

Simulation results show that our proposal increase the ability of SWAN model to provide service differentiation in a more precise way. It is capable to differentiate between five data flows and provide them different QoS level. Based on various simulation experiments with different scenarios, we argue that our SWAN modification and scheduling extension is a significant improvement of SWAN model in terms of service differentiation and end-to-end delay performance. It is important to note that we also simulated scenarios with both routing approaches, reactive and proactive. The results show no difference and independence of our proposal on routing protocols.

This was expected because Stateless Wireless Ad hoc Network QoS model itself is not dependent on any particular routing technique.

58.5 Conclusions

This paper deals with QoS models in MANETs. It describes most used ones and presents their advantages and disadvantages. We focus on SWAN model because due to its simplicity, robustness and scalability, it can provide soft QoS in MANET

networks. We propose an extension of SWAN model in order to increase level of service differentiation and by adding probabilistic scheduling approach also one-way delay.

Simulation experiments show that our proposal improves the performance of MANETs in terms of one-way end-to-end delay. In addition, it is compatible with different routing approaches, proactive and reactive. Our next idea for future work is to analyze the use of Random Early Detection (RED) technique in each queue to increase more network performance and to avoid congestion in the network..

Acknowledgments This work was partially supported by the Science and Technology Assistance Agency under the contract No. 1/0243/10.

References

1. Marwaha S et al (2008) Challenges and recent advances in QoS provisioning, signaling, routing and MAC protocols for MANETs. In: Proceedings of telecommunication networks and applications conference (2008), pp 97–102
2. Mohapatra P et al (2003) QoS in mobile Ad hoc networks. *IEEE Wirel Commun* 4(3): 291–302
3. Ilyas M (2003) The handbook of Ad hoc wireless networks. CRC Press, Boca Raton. ISBN 0-8493-1332-5
4. Nahrstedt K (2011) An integrated mobile Ad-hoc QoS framework [Online; accessed 20 Feb, 2011]. Available at: <http://ftp.rta.nato.int/public//PubFull-Text/RTO/TR/RTO-TR-IST-030///TR-IST-030-W3/TR-IST-030-W3-Presentations+Discussions/TR-IST-030-W3-14.pdf>
5. Satyabrata C, Amitabh M (2001) QoS issues in Ad hoc wireless networks. *IEEE communication magazine*, pp 142–148
6. Hekmat R (2006) Ad-hoc networks: fundamental properties and network topologies. Springer, Berlin, ISBN 1-4020-5165-4
7. Tavli B, Henzelman W (2006) Mobile Ad hoc networks—energy efficient real time data communications. Springer, Berlin, ISBN-10 1-4020-4632-4
8. Basagni S et al (2004) Mobile Ad hoc networking. Wiley-IEEE press, New York, ISBN: 978-0-471-37313-1
9. Enzai N, Farhat A, Omer M (2008) Evaluation study of QoS-enabled AODV. In: Proceedings of the international conference on computer and communication engineering, pp 1254–1259
10. Wu K, Harms J (2001) QoS support in mobile Ad hoc networks. *Crossing Boundaries Interdiscip J* 1(1):92–107
11. Xiao H et al (2000) A flexible quality of service model for MANET. In: Proceedings of vehicular technology conference, pp 445–449
12. Nahrstedt K (2010) An integrated mobile Ad-hoc QoS framework. [Online; Accessed 17 Feb, 2010]. Available at: <http://ftp.rta.nato.int/public//PubFull-Text/RTO/TR/RTO-TR-IST-030///TR-IST-030-W3/TR-IST-030-W3-Presentations+Discussions/TR-IST-030-W3-14.pdf>
13. Lee SB et al (2000) INSIGNIA: an IP-based quality of service framework for mobile Ad hoc networks. *Parallel Distrib Comput Special Issue Wirel Mobile Comput Commun* 60(4):374–406
14. Dharmaraju D, Chowdhury A (2002) INORA—a unified signaling and routing mechanism for QoS support in mobile Ad hoc networks. In: Proceedings of the 2002 international conference on parallel processing workshops, pp 86–93
15. Sulthani M, Rao D (2009) Design of an efficient QoS architecture (DEQA) for mobile Ad hoc networks. *ICGST-CNIR J* 8:49–57

16. Krishna P, Iyengar S (2007) A cross layer based QoS model for wireless and mobile ad hoc networks. *Mobile Communication* 1:114–120.
17. Ahn G et al (2002) Supporting service differentiation for RT and best-effort traffic in stateless wireless Ad hoc networks (SWAN). *IEEE Trans Mobile Comput* 1:192–207
18. Zhang N, Anpalagan A (2010) Sensitivity of SWAN QoS model in MANETs with proactive and reactive routing: a simulation study. *Telecommunication Systems* 44(1–2):17–27
19. Ahu GS et al (2002) SWAN: service differentiation in stateless wireless Ad hoc networks. In: *Proceedings of IEEE INFOCOM'2002*, New York
20. Zhang N, Anpalagan A (2009) Sensitivity of SWAN QoS model in MANETs with proactive and reactive routing: a simulation study. *J Telecommun Syst* 10:307–311
21. The Network Simulator ns-2. [Online; accessed 20 Feb, 2011]. Available at: <http://www.isi.edu/nsnam/ns/>
22. Yuming J, Chen-Khong T, Chi-Chung K (2002) A probabilistic priority scheduling discipline for multi-service networks. In: *Sixth IEEE symposium on computers and communications*, Elsevier Science, vol 25. No. 13, pp 1243–1254

Chapter 59

Method for Data Collection and Integration into 3D Architectural Model

L. Kurik, V. Sinivee, M. Lints and U. Kallavus

Abstract Rapid development in the field of smartphones over past years has enabled using their multimedia capabilities in virtually reconstructing and visualizing architectural objects by creating 3D-models of such objects. Several authors have tested cell phone cameras accuracy. For example Nokia N93, Sony Ericsson K750i, iPhone placed under test could be successfully used in applications not requiring ultimate precision [Gruen A, Akca D (2009) Evaluation of metric performance of mobile phone cameras. Institute of Geodesy and photogrammetry, ETH, Zurich. doi:10.3929/ethz-a-005749738; Takahashi Y, Chikatsu H (2009) 3D Modelling and visualization of cultural heritage using mobile phone cameras. In: Proceedings of the 3rd ISPRS international workshop 3D-ARCH, Trento]. Rapid technical development in the field of unmanned aerial vehicles (UAV) and lowering of their cost has made them available practically for everyone interested and eased tremendously use of such machines in aerial photography and photogrammetry [Eisenbeiss H (2009) UAV photogrammetry. Ph.D thesis, Dissertation ETH no. 18515, ETH Zurich; Remondino F, Barazzetti L, Nex F, Scaioni M, Sarazzi D (2011) UAV photogrammetry for mapping and 3D . In: Proceedings of the international conference on unmanned aerial vehicle in geomatics (UAV-g), Zurich].

L. Kurik (✉) · V. Sinivee · M. Lints
Department of Physics, Tallinn University of Technology,
Ehitajate tee 5, 19086 Tallinn, Estonia
e-mail: lembit.kurik@ttu.ee

V. Sinivee
e-mail: veljo.sinivee@ttu.ee

M. Lints
e-mail: martin.lints@mail.ee

U. Kallavus
Centre for Materials Research, Tallinn University of Technology,
Ehitajate tee 5, 19086 Tallinn, Estonia
e-mail: urka@staff.ttu.ee

We used Nokia N8 smartphone in our study since it had best optical, computing and communications capabilities at the time of this writing. In present paper we illustrate two ways of using smartphone's camera in mapping object's properties. One is a well known method of photogrammetric 3D-modelling. Another is a less used method of photogrammetric positioning [Dillon MJ, Bono RW, Brown D. L (2004) Use of photogrammetry for sensor location and orientation. IMAC-XXII: conference and exposition on structural dynamics, Jacksonville; Mautz R, Tilch S (2011) Survey of optical indoor positioning systems. In: Proceedings of the international conference on indoor positioning and indoor navigation, Guimarães]. Firstly we demonstrate mapping of moisture content using photogrammetric positioning of sensor. Mechanically rigidly connected moisture probe and phone (together with built-in camera and special software) are linked via a bluetooth adapter enabling thus synchronisation of moisture measurement results to exact location. Individual results are stored directly to accompanying image file used for positioning each measurement. Such approach should eliminate possible human errors common in long and tedious measurements. Secondly a 3D model of a Tallinn Observatory tower is created by traditional means using phones built-in 12 megapixel camera. Low weight of the phone (about 120 g) enabled transporting it to suitable photographing positions onboard an airborne UAV. Described in present paper method is suitable for 3D-mapping of a large variety of physical properties of objects in interest.

59.1 Introduction

Availability and constantly dropping prices of phones equipped with camera has made using them in photogrammetric applications easy even in not planned beforehand cases. Immense computational power, various built-in sensors and diverse connecting capabilities of smartphones create good basis for registering various measurement results and for integrating them into photogrammetrically created virtual models.

In essence new smartphones are small computers running their own operating systems. Their manufacturers have created multiple tools for writing and even virtually testing own applications easily and quickly. This gives smartphones a big advantage over SLR-cameras and gives user a chance of using cell-phone in 3D-mapping of various measurement results.

Location of a (smartphone) camera can be determined photogrammetrically. Results of cameras own built-in sensor or remote sensor mechanically and rigidly attached to the camera and transmitted to it via some media (i.e. Bluetooth) can then be bound to location image.

A data collection containing locations of measurement points, results and even sensor (s) orientation (s) can be created in such way. A 3D map of results will be constructed of acquired data later. One can use different sensors depending on the nature of a physical property of object that needs to be mapped.

The goal of present research was creating a method for visualizing moisture content of walls of buildings having a historic and heritage value in order to simplify monitoring changes in “health” of such objects in time and help raising quality of decisions concerning restoration of mentioned buildings.

Traditional approach (choosing measurement points with a constant step and creating a 2D-map of moisture content) works well in case of objects with smooth and even walls. This method is not suitable for historical buildings typically having very uneven and complex-shaped walls. A 3D mapping of each measurement point would be needed in order to increase measurements repeatability and overall quality.

59.2 System Description

Relying on needs discussed above we created a system consisting mainly of a smartphone Nokia N8 (12 Mpx camera, Symbian 3 operating system), an application “6D measure” written in Python (communication with sensors, storing data and binding it with a picture file shot for photogrammetric positioning), measurement sensor (microwave moisture content measurement device), Bluetooth module for communications between phone and sensor and PC program “PhotoModeler” (Fig. 59.1). The system enables storing 6D positioning data (three space coordinates plus orientation of the sensor) and measurement results into EXIF-data part or filename of image file used for positioning. Details of the process for binding camera positions and measurements results for creating 3D maps are described in greater detail in earlier works [7]. Method for calculating Euler angles of measurement probe (s) could be found from mentioned paper as well. We used similar method for creating 3D models on complex shaped objects earlier as well. Orientation of measurement probe could be crucial for some types of sensors. For example a moisture content probe with working depth up to 3 cm, uses polarized microwave radiation making determining its orientation obligatory when used with test objects having a layered structure.

59.2.1 Data Filter for Extracting Measurement Results

Described in present paper method for collecting data of architectural objects is not very demanding on equipment. In most cases only a camera and device for measuring moisture (or virtually any other physical property) is needed. Our group used a microwave moisture meter “Moist-200” for measuring moisture contents of walls for building moisture maps. Since one of our goals was automation of measurements, “Moist-200” was slightly modified. The aim of modifications was to send measured value to mobile phone via Bluetooth link instantly after measurement took place. Original “Moist-200” was not capable for outputting

Fig. 59.1 Main components of measurement system

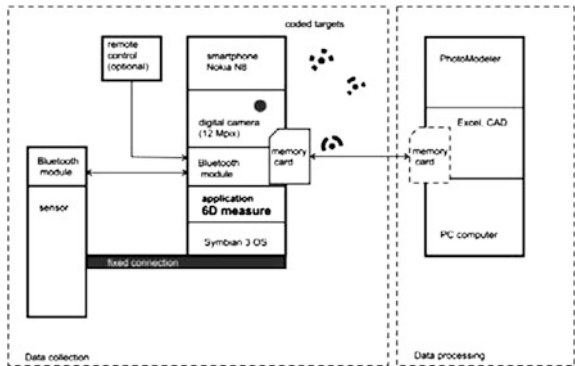


Fig. 59.2 Data collection part of the system in work



individual results. It could only transmit data series containing values of the whole measurement cycle (Fig. 59.2).

We could not extract data directly from the measurement probe although it seems to be using standard SPI data bus. Commands for starting measurements were not known as well as data processing algorithm. Luckily enough there was another possibility—let the instrument calculate result. After this operation has been completed, one can read values electronically from LCD screen of the meter.

For this purpose we added a small microcontroller listening to communications between “Moist-200” processor and display. Since LCD is connected via a 4-bit bus the controller has to perform two read operations to get one byte of data. Each read is triggered by a pulse on Enable pin of the LCD. “Moist-200” sends various messages on screen (menus etc.). Controller looks for measurement data only, filters it out of the whole stream and sends to a mobile phone via a Bluetooth data link. Controller also initializes Bluetooth adapter and opens connection to the phone. Circuit diagram of the controller is given on Fig. 59.3.

One can see, that connection is really elementary—it consists of a microcontroller and clock crystal only. One could even use internal calibrated oscillator of the controller if small temperature-dependent drift of serial communications speed is not relevant.

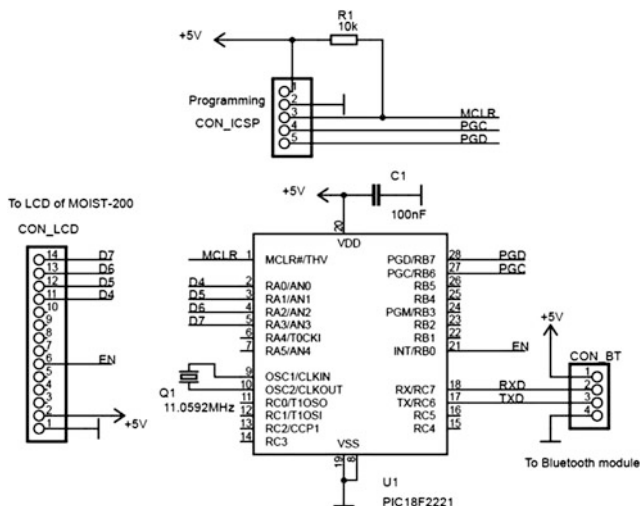


Fig. 59.3 Electrical circuit of controller for extracting measurement data from “Moist-200”

Some pins of the controller are reserved for future communications between smartphone and different measurement sensor(s).

Although not used together with “Moist-200”, controller offers a possibility for connecting a GPS engine. If it is enabled in configuration, geographic coordinates of each measurement location are output together with measured values.

This feature could be handy in measurement systems lacking an integrated GPS engine (Nokia N8 has one).

Circuit consumes about 3 mA from “Moist-200” stabilized +5 V power supply. The Bluetooth module consumes about 100 mA in peak, average consumption is 30–40 mA.

In order to achieve high-speed and flawless operation, firmware for the controller is written entirely in assembly language. Firmware is designed using a “state machine” approach.

59.2.2 Smartphone Software

For ease of programming, the software is written in Python for S60 (PyS60). Mobile phone is connected with the bluetooth module by creating a serial port with PyS60 socket module [8]. A function listens to the serial port and when data stream is received and fully in the buffer, it calls another function which takes a picture using the mobile phone’s camera. That picture is then saved along with the data received. Initial version of the program saves the data in the filename.

Fig. 59.4 UAV on a mission near our astronomy observatory



The process of measuring and taking the picture is initiated solely by pressing a button on the measuring device. All following processes take place automatically and there is no need for further user input.

In the future, the script can be improved to save the data in EXIF header, to check and handle errors in received data and to improve user interface (picture save path, Bluetooth module address selection, data saving location and method, etc.)

59.2.3 3D Modelling

The camera of the smartphone could be used in traditional ways even for creating 3D models of objects. Accessibility of suitable photographing positions in working with architectural objects is a common problem. In reality some parts of a building could be hidden if watched from ground or they may be not accessible for photographing (too high). In order to take photos of such “dark corners” a radiocontrolled multirotor copter built and piloted by one of the co-authors was used.

Our quadcopter implements four electric brushless motors (distance from motor shaft to motor shaft is 60 cm) consuming an average of 12–14 A from onboard 11,1 V Li-Polymer battery in normal hovering mode. Quick manoeuvres can rise current consumption even to 48 A (maximum current that machines ESC’s can hold, flight logs reveal a maximum of 42,5 A consumed)! Two Li-Po batteries with a 2,3 A/h capacity gives approximately 12 min of “normal” flight time (Fig. 59.4).

All-up weight of copter is about 1.5 kg. A similar tricopter with tilt/pan camera mount was also tested and found suitable for our mission.

Flight controller of the copter is built around a cheap microcontroller Atmega48. Circuit uses three MEMS gyroscopes with analog output for determining deviation of the platform.

Presently acceleration sensors are not implemented and keeping copter stable is the main task of the pilot. “Stable mode”, altitude- and position hold and other automatic features will probably be implemented in next version of flight controller.

59.3 Indoor Tests

Photogrammetrical suitability of Nokia N8 phones camera was tested both in single-shot (resolution: 4000×3000) and video (resolution: 1280×720) modes with the aid of a PhotoModeler Pro software suite. Our vertically placed test field (1.1×2.2 m) had $12 \times 22 = 264$ targets, 108 of them were coded targets (Fig. 59.5).

Tests were carried out in 4 different modes using video and single-shot modes of the phone, holding the phone in hand (not on a tripod!) and from board of an airborne UAV. Our goal was not to determine the maximum photogrammetric accuracy of the camera, but rather getting an estimate of accuracy obtainable in real measurement process. Shooting commands were sent to the camera on board an UAV using a standard Bluetooth mouse (operating system of smartphone used has built-in functions for that purpose). It is possible to send commands to the camera of the phone via a standard PC as well.

Figure 59.6 displays images of coded targets acquired during different test conditions.

Best results were obtained, like expected, using a photo mode and holding camera in hand. Deviation of centres of coded targets in 3D space was calculated for all targets using a program “PhotoModeler Pro”. Average deviation was found to be about 1:500 which is much less than could be achieved theoretically but still usable in moisture measurement applications. It was possible to find frames from video files with quality similar to those shoot from hand suggesting that video mode is also applicable for photogrammetric positioning. It is clear that using video mode from board the UAV needs improvement, especially in outdoors conditions.

In present experiment the UAV was piloted purely manually (only 3 low-end gyros helped holding the machine in horizontal position). Building a flight stabilisation system (autopilot) should improve results tremendously. Designing such controller is presently going on.

59.4 Tests on an Astronomy Observatory

Tests of the described system were carried out on an astronomy observatory of Tallinn University of Technology. The observatory is located near the university in a sight-seeing tower built in 1910 by an eccentric baron von Glehn.

Fig. 59.5 View of the test field from UAV

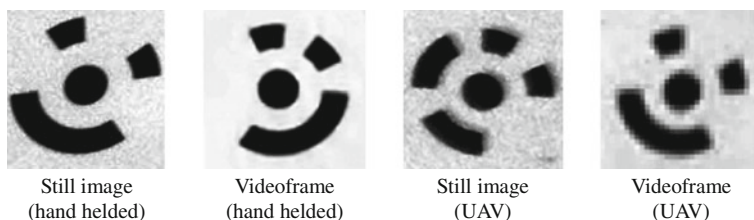


Fig. 59.6 Quality of the images of the targets

Presently the tower is a heritage protection object. Its outer wall having many different-sized windows, storm ladder, various indents and salient parts is clad of natural unprocessed granite rocks.

Numerous horizontal surfaces and joints of different width make moisture entrance to the construction relatively simple. Complicated history of the tower [over lived two wars, was abandoned for years, was even used as a fire place of the midsummer feast (Fig. 59.7)] has created big moisture problems involving gathering of salts, corrosion of steel armature of internal floors, destruction of internal refinement, decay of windows etc.

In 2010 the outer wall of the tower was capitally reconstructed, doors and windows were replaced. A special technology was used for filling joints of the wall. Quality of reconstruction is presently not yet known.

A 3D-model of the tower was constructed photogrammetrically and results of moisture content measurements of inner walls were bound with it in order to document the present state of the building and monitoring its health in future (Fig. 59.8).

The complex architectural model thus acquired with moisture data on it gives us a snapshot of the state of the building right after the reconstruction and creates basis for judging the quality of repair works and suitability of the technology used for Estonian climate.



Fig. 59.7 Tallinn Observatory at different times

Fig. 59.8 Part of the 3D model of the tower

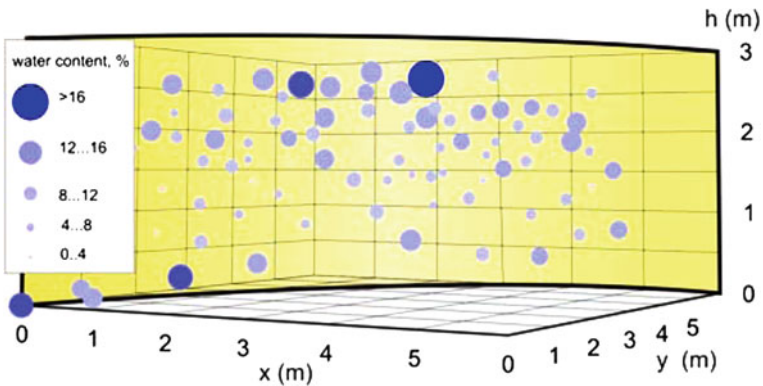


Fig. 59.9 Moisture content map of the wall from inner side, diameter of the bubbles is proportional to moisture content

Simplified (2D) moisture maps of some parts of the tower made in 2005 and 2008 serve as a comparison base. Those early measurements indicate that moisture in the walls is not water induced by capillary forces but originates rather from rainwater that came in through broken joints in the wall. Now there is a hope that walls may start losing moisture (Fig. 59.9).

59.5 Conclusion

The method proposed and tested in current paper for connecting an external sensor (probe) to a device for photogrammetric positioning is a perspective tool for 3D mapping of different physical parameters. A smartphone with large memory, good camera and powerful computing capabilities (Nokia N8) was used as a photogrammetric device in present work. Light weight and possibility for remote controlling of the phone makes it especially suitable candidate for a central photogrammetric unit in various applications (i.e. as an on-board camera of an UAV like in this work).

Obtainable accuracy depends of many factors and is theoretically sufficient for different applications. Improving accuracy calls for diligence and extra time as well as technical aids like extra lighting, camera stabilisers etc. that are not needed in majority of situations. Proposed method binds result of every single measurement with accompanying image file used for positioning. Although amount of data acquired in such way is massive, it should not be an issue to modern computing devices. On the other hand every measurement is well documented and improvement of accuracy is possible even later (by taking extra photos). Improving software and user interface for described method should make it an affordable and available photogrammetric tool.

References

1. Gruen A, Akca D (2009) Evaluation of metric performance of mobile phone cameras. Institute of Geodesy and photogrammetry, ETH, Zurich. doi:[10.3929/ethz-a-005749738](https://doi.org/10.3929/ethz-a-005749738)
2. Takahashi Y, Chikatsu H (2009) 3D Modelling and visualization of cultural heritage using mobile phone cameras. In: Proceedings of the 3rd ISPRS international workshop 3D-ARCH, Trento
3. Eisenbeiss H (2009) UAV photogrammetry. Ph.D. thesis, Dissertation ETH no. 18515, ETH Zurich
4. Remondino F, Barazzetti L, Nex F, Scaioni M, Sarazzi D (2011) UAV photogrammetry for mapping and 3D . In: Proceedings of the international conference on unmanned aerial vehicle in geomatics (UAV-g), Zurich
5. Dillon MJ, Bono RW, Brown DL (2004) Use of photogrammetry for sensor location and orientation. IMAC-XXII: conference and exposition on structural dynamics, Jacksonville
6. Mautz R, Tilch S (2011) Survey of optical indoor positioning systems. In: Proceedings of the international conference on indoor positioning and indoor navigation, Guimarães

7. Kurik L, Kallavus U (2007) Low Cost Photogrammetric System For 6D Positioning Contact Type Sensors. In: Proceedings of 8th conference on optical 3D measurement techniques, ETH Zurich, vol II, pp 202–206
8. <http://epx.com.br/artigos/pys60.php> (Accessed on 23.11.2011)

Chapter 60

Statistical Analysis to Export an Equation in Order to Determine Heat of Combustion in Blends of Diesel Fuel with Biodiesel

C. G. Tsanaktsidis, V. M. Basileiadis, K. G. Spinthoropoulos,
S. G. Christidis and A. E. Garefalakis

Abstract Herein we try to export equations which describe the variation of heat of combustion in blends of Diesel fuel with Biodiesel. Using specific volume of these blends, we determine heat of combustion, in order to study the contribution of each kind of Biodiesel and we convert the statistical data into mathematic relations as a specific formula, attempting to achieve an empirical evaluation.

60.1 Introduction

The new living conditions that include the upgrade of level of life, the growth of cities, by reason of the displacement of people, and the increased living needs lead to the expanding market for petroleum products. This necessity has a great impact on the country's trade balance and the financial growth due to the increased crude oil imports. The damaging environmental results of harmful gases due to the use of diesel oils in the means of transport made the European government decide to forbid their use and protect the environment from polluted atmosphere which causes the acid rain, the green house effect and noxious health impacts.

C. G. Tsanaktsidis (✉) · V. M. Basileiadis · S. G. Christidis
Laboratory of Qualitative Fuel Control, Department of Pollution Control and Technologies,
Technological Education Institute of Western Macedonia, Kila, 501 00 Kozani, Greece
e-mail: felch@staff.ttu.ee

K. G. Spinthoropoulos
Department of Financial Applications, Technological Education Institute of Western
Macedonia, Kila, 501 00 Kozani, Greece

A. E. Garefalakis
Department of Accounting, Technological Education Institute of Crete Estauromenos,
710 04 Heraklion, Greece

Taking the above issues into account, the conclusion could be made that the biodiesel is the only fuel that is able to have the properties of the diesel oil and meet the today's engines demands of any vehicle for proper, satisfying energy. It can be blended in diesel fuels and provide the same burning and stability under any circumstances [1].

The biodiesel production can be succeeded by the four main ways which are the followings: microemulsions, thermal cracking (pyrolysis), transesterification (alcoholysis) and naturally the immediate use and blending [2].

However, among them, the transesterification of natural and fats oils are mostly applied. The procedure involves the use of alkalis, acids or enzymes in order to be the catalysts to the transesterification chemical reaction of triglycerides with alcohols (methanol, ethanol, propanol, butanol or amyl alcohol) [3, 4].

In this framework, the fuel's energy output is a significant characteristic and it is known as heat of combustion [5]. The scientists attempt to experiment and derive the heat of combustion (HOC) by working with an oxygen bomb calorimeter and applying specified rules (e.g. ASTM D4809-09a). In addition, among the relevant methods applied to biodiesel fuels in order the HOC to be measured and evaluated are the ultimate analysis and the approximate analysis which are based on basic analysis data [6]. The physical characteristics of the vegetable oils were examined so as the HOC to be evaluated and the scientists experimented on the viscosity and density [7] and saponification [8] as well. Further to this the carbon number and molecular weight were the main bases for Sadramela et al. in 2008 [9] to make research on them and come to accurate outcome for the HOC of the saturated fatty acids. With reference to VC estimation, several experimental procedures can be applied [5].

Moreover, a necessity for other types of fuels for engines that use different and blended types of oils arises due to environmental and energy matters, is claimed by Greene [10]. Thus, a scientific research has held regarding the enterprises environmental impact by the ICT branch, Stiakakis and Fouliras (2009) [11] report applying the Data Envelopment Analysis (DEA) method. Following to this, the variables will be tested by econometric methods focusing on the relevance of the fuels density and the blend in the question cases.

The estimation of proper properties of the new fuels is the scientific target and it is achieved by this methodology with the use of the extracted equations that account the fuels density in advance and the experimental actions are not needed. So next to this, the $Y = ax + b$ is the pointed linear equation that shows the variables reaction in every blend.

60.2 Experimental and Statistical Process

60.2.1 Production and Physicochemical Properties of Biodiesel

60.2.1.1 Production of Biodiesel

Biodiesel is produced by transesterifying the parent oil or fat with an alcohol, usually methanol, in presence of a catalyst, usually a strong base such as sodium or potassium hydroxide, or, preferably and increasingly more commonly, alkoxides. The resulting product therefore can contain not only the desired alkyl ester product but also unreacted starting material, residual alcohol, and residual catalyst [12].

Glycerol is formed as by-product and separated from Biodiesel in the production process, however, traces thereof can be found in the final Biodiesel product (Fig. 60.1).

60.2.1.2 Physicochemical Properties of Biodiesel

After the production of Biodiesel, we have to certify the appropriateness of the end product, measuring its physical properties and compare their values to its specifications. For example, Biodiesel can absorb a certain amount of water during storage. Such issues are addressed in Biodiesel standards.

Some specifications in Biodiesel standards are carryovers from Diesel standards. However, not all test methods carried over from Diesel standards into Biodiesel standards are well suited for Biodiesel analysis.

Although there are many standard methods used for analyzing the various properties, emphasis is placed on methods determined in our laboratory.

60.2.1.3 Physicochemical Properties of Diesel Fuel

The Diesel oil constitutes blends of many hydrocarbons with different properties. Diesel of internal combustion should have attributes that ensure auto ignition of fuel and furthermore sure and smooth combustion without problems in the conditions of booth combustion. The properties of fuel depend on the type of hydrocarbons that it contains, as well as from their contents.

60.2.2 Measurements of Heat of Combustion of Fuel Samples

In this study we used pure Diesel and two kinds of Biodiesel; Biodiesel by Animal Fats and Biodiesel by Vegetables (vegetable oil fuel). These samples met the specifications of Diesel fuel and Biodiesel Standards (Tables 60.1, 60.2).

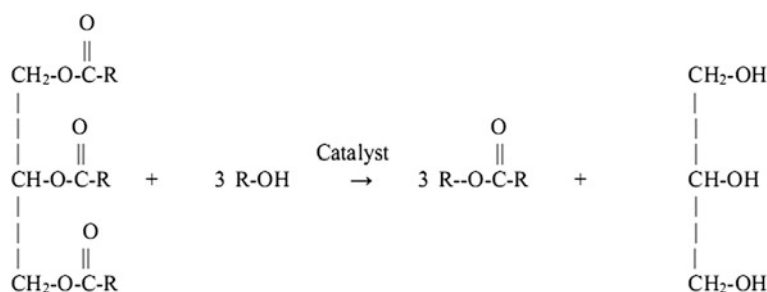


Fig. 60.1 Transesterification reaction [15]

Table 60.1 Laboratory values of vegetable biodiesel and animal fats biodiesel and biodiesel standards (European Biodiesel Standards EN 14214 for Vehicle Use) [16]

| Physicochemical property | Animal fats biodiesel laboratory values | Vegetable biodiesel laboratory values | Biodiesel standards | Methods of determination ASTM |
|---|---|---------------------------------------|---------------------|-------------------------------|
| Density 15 °C (g/mL) | 0.883 | 0.884 | 0.860–0.900 | D1298-2005) [17] |
| Kinematic viscosity, (40 °C) (mm ² /s) (cSt) | 4.36 | 4.54 | 1.9–6.0 | D 445-06 [18] |
| Humidity (mg/kg) | 322 | 296 | <500 | D 1744-92 (2000 [19] |
| Heat of combustion (kJ/kg) | 41,111 | 40,036 | >35,000 | D 4809-09a [13] |
| Total acid number (mg/ g KOH) | 0.08 | 0.24 | <0.50 | D 664 [20] |

The total volume of each blend was 100 mL (for example 20 % blend included: 80 mL Diesel and 20 mL Animal Fats Biodiesel or 20 mL Vegetable Biodiesel) and in every measurement the volume of Biodiesel was changing.

We measured heat of combustion via the method ASTM D 4809-09a [13] (with calorimeter IKA C200) firstly in pure Diesel and then in all blends using 0.5 g of each fuel (accurately weighed with a Precisa YT 220A analytical balance) and the results are presented in Table 60.3.

60.2.2.1 Statistical Empirical Findings

The selection of the suitable model will be realized by using the Regression Analysis targeting to be used for the prices evaluation of the dependent variable in the specific prices of the independent variable. The preliminary statistical analysis of the series, that are the Heat of Combustion Value (Y) and blend (X) (fuels), has led us to the accurate conclusion that the relation that connects the two variables is linear as it is defined in Figs. 60.2 and 60.3).

Table 60.2 Laboratory values of pure diesel and greek specifications (enarmonized with European Community) of diesel fuel (FEK 332/B/11-2-2004, EN 590:1999)

| Physicochemical property | Diesel laboratory values | Diesel standards | Methods of determination ASTM |
|---|--------------------------|------------------|----------------------------------|
| Density 15 °C (g/mL) | 0.829 | 0.820–0.845 | D1298-99 (2005) [17] |
| Kinematic viscosity (40 °C) (mm ² /s) (cSt) | 2.49 | 2.00–4.50 | D 445-06 [18] |
| Humidity (mg/kg) | 50.0 | <200.0 | D 1744-92 (2000) [19] |
| Heat of combustion (kJ/kg) | 46,428 | >42,600 | D 4809-09a [13] |
| Total acid number (mg/g KOH) | – | – | D 664 [20] |

Table 60.3 Values of heat of combustion of diesel and its blends with biodiesel (vegetable and animal fats)

| 1st blend (diesel + animal fats biodiesel) | | 2nd blend (diesel + vegetable biodiesel) | | Method of determination |
|--|----------------------------|--|----------------------------|-------------------------|
| Blend (% v/v) | Heat of combustion (kJ/kg) | Blend (% v/v) | Heat of combustion (kJ/kg) | |
| 0.00 | 46,428 | 0.00 | 46,428 | ASTM D 4809-09a [13] |
| 10.00 | 45,828 | 10.00 | 45,668 | |
| 20.00 | 45,278 | 20.00 | 44,908 | |
| 30.00 | 44,678 | 30.00 | 44,152 | |
| 40.00 | 44,128 | 40.00 | 43,488 | |
| 50.00 | 43,472 | 50.00 | 42,955 | |
| 60.00 | 42,961 | 60.00 | 42,399 | |
| 70.00 | 42,461 | 70.00 | 41,819 | |
| 80.00 | 41,941 | 80.00 | 41,219 | |
| 90.00 | 41,511 | 90.00 | 40,619 | |
| 100.00 | 41,111 | 100.00 | 40,036 | |

The variables relation could be stated by the simple linear regression model as it is showed in the following equation:

$$Y_i = b_o + b_iX + U_t \tag{60.1}$$

where Y_i : the i is the observation of the dependent variable (Heat of Combustion Value)

X_i : the I the observation of the independent variable blend

b_0, b_1 : the straight regression coefficients

U_t : the equation error.

So proceeding to our research, we have created two blends which are blend 1, that refers to the pure diesel with the animal fats biodiesel and the blend 2 that refers to the pure diesel and the vegetable biodiesel. The two blends were studied

Fig. 60.2 Dispersion of heat of combustion of blend 1

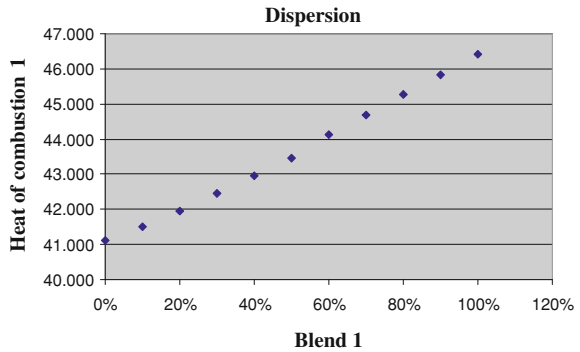
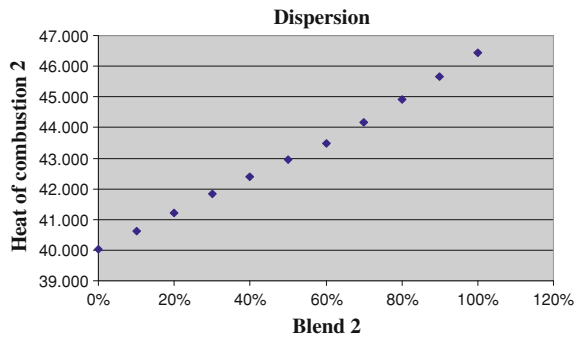


Fig. 60.3 Dispersion of heat of combustion of blend 2



in terms of the heat of combustion value in order to find the equation that will define the blends heat of combustion value.

As far as the two blends are concerned (blend 1 & blend 2), we expected that the upgrade of the animal and vegetable biodiesel would be proportional to the increase of the heat of combustion value arithmetic outcome due to the linear relation that connects the two variables. So studying on the regression, we notice that the Eqs. 60.2 and 60.3 have correct signals as the blend price increase should quote positively the heat of combustion variable.

$$\begin{aligned} \text{heat of combustion 1} &= 40.914 + 0.0540 \text{ blend1} \\ \eta \\ Y &= 40.914 + 0.0540X \end{aligned} \quad (60.2)$$

$$\begin{aligned} \text{heat of combustion 2} &= 39.927 + 0.062 \text{ blend2} \\ \eta \\ Y &= 39.927 + 0.062X \end{aligned} \quad (60.3)$$

With reference to the blend 1, the accounted price of the statistical element F is 2,844.730 and the p value is 0.000,000. Similarly to the blend 2, the price of the

Table 60.4 Values of t-statistic and R^2

| 1st blend | | | 2nd blend | | |
|---------------------|--------------------|----------------|---------------------|--------------------|----------------|
| b_0 | b_1 | Adjusted R^2 | b_0 | b_1 | Adjusted R^2 |
| 40.91 | 0.054 | 0.99 | 39.92 | 0.062 | 0.99 |
| 682.33 ^a | 53.33 ^a | | 527.76 ^a | 49.03 ^a | |
| ≈0.00* | ≈0.00* | | ≈0.00* | ≈0.00* | |

^a t-statistic; * p value

p value is 0.000,000 therefore the tested model is statistically significant disallowing the zero case that demands at least one $\beta_k \neq 0$ in significance level α . (meaning 0.05 or 0.01). Next to this, we terminate that the selected regression model for the two blends is able to clear an important part of the heat of combustion variability, namely the relation $Y = \sum (Y - \bar{Y})^2$.

The examination of the single variables by the t test suggests the significance of the blend variable.

Aiming to examine the relation between the dependent and the independent variables, we study the multiple correlation coefficient R . As we have already substantiated the fuel blend alteration is proportional to the Eqs. 60.2 and 60.3 output alteration and eventually to the fuel’s heat of combustion value. This is statistically supported with regard to the b_1 parameter modification. The regression model explanatory value is upgraded by the price increase of the multiple correlation coefficient which tends to one. As long as the $R^2 = 1$, the straight line goes through all the points (xi, yi) of the dispersion diagram. The adaptation or definition index pinpoints the dependent variable rate that is defined by the independent variable variations.

We notice in Table 60.4 that, in both of the blends, the adjusted R^2 is high and close to one which indicates the fact that the independent variable defines at about 99 % in both of the blends the dependant variable prices, namely the fuel’s heat of combustion value.

The nearly perfect positive relation between the two blends variables, by applying the R^2 adjusted, ensures the dispersion diagrams that were used to support the linear relation that involves the variables. Speaking specifically, based on the R^2 adjusted, the approximately 99 % of the dependant variable variability is attributed by the regression. The 1 % residual is proceeded by foreign factors and are not included in this question model. This suggests that the approximately the 48 % of the dependant variable variability is indicated by the regression. The remainder variability is based on other factors that are not included in this model. The t test is comparable to the p value examination and taking into account that the rate of the question index should be very low so as to be statistically significant, the Table 60.4 supports the statistically significance of 1st & 2nd blends is high provided that the p value is lower than 1 % (based on the regression output).

The model suitability test was realized by the application of diagnostic tests and the existence examination of standardized residuals serial correlation. Next to the

Table 60.5 LB test for residuals and squared residuals from the regression estimation of the heat of combustion and the blend

| Standardized residuals | | | | Squared standardized residuals | | | |
|------------------------|------------------|---------------------|--------|--------------------------------|------------------|---------------------|--------|
| Lags | Auto correlation | Partial correlation | LB(n) | Lags | Auto correlation | Partial correlation | LB(n) |
| <i>Blend 1</i> | | | | | | | |
| 1 | 0.482 | 0.482 | 33.191 | 1 | -0.002 | -0.002 | 9.E-05 |
| 2 | 0.108 | -0.161 | 35.052 | 2 | -0.191 | -0.191 | 0.5826 |
| 3 | -0.175 | -0.211 | 40.525 | 3 | -0.196 | -0.205 | 12.694 |
| 4 | -0.416 | -0.305 | 75.872 | 4 | -0.002 | -0.051 | 12.694 |
| 5 | -0.523 | -0.292 | 14.113 | 5 | 0.353 | 0.297 | 42.393 |
| 6 | -0.249 | 0.106 | 15.888 | 6 | -0.078 | -0.119 | 44.121 |
| 7 | -0.125 | -0.207 | 16.451 | 7 | -0.127 | -0.037 | 49.897 |
| 8 | 0.042 | -0.067 | 16.534 | 8 | -0.145 | -0.078 | 59.915 |
| 9 | 0.150 | -0.123 | 18.152 | 9 | -0.159 | -0.243 | 78.013 |
| <i>Blend 2</i> | | | | | | | |
| 1 | 0.576 | 0.576 | 47.362 | 1 | 0.172 | 0.172 | 0.4219 |
| 2 | 0.065 | -0.399 | 48.024 | 2 | -0.281 | -0.320 | 16.766 |
| 3 | -0.339 | -0.287 | 68.573 | 3 | 0.018 | 0.157 | 16.822 |
| 4 | -0.468 | -0.111 | 11.325 | 4 | 0.229 | 0.114 | 27.511 |
| 5 | -0.368 | -0.090 | 14.556 | 5 | -0.089 | -0.151 | 29.381 |
| 6 | -0.255 | -0.266 | 16.421 | 6 | -0.144 | 0.019 | 35.343 |
| 7 | -0.088 | -0.068 | 16.697 | 7 | -0.119 | -0.211 | 40.422 |
| 8 | 0.068 | -0.053 | 16.914 | 8 | -0.129 | -0.142 | 48.345 |
| 9 | 0.156 | -0.107 | 18.665 | 9 | -0.121 | -0.098 | 58.829 |

Note LB(n) are the n-lag Ljung–Box statistics for the residual series. LB(n) follows Chi-square variable with n degree of freedom; the series of residual contains 9 observations

Table 60.6 Breusch–Godfrey serial correlation LM test [21]

| | | | |
|----------------|--------|-------------|-------|
| <i>Blend 1</i> | | | |
| F-statistic | 3.6545 | Probability | 0.081 |
| Obs*R-squared | 5.6188 | Probability | 0.060 |
| <i>Blend 2</i> | | | |
| F-statistic | 3.1814 | Probability | 0.10 |
| Obs*R-squared | 5.2377 | Probability | 0.072 |

Table 60.5, this took place using the statistical elements LB for the standardized residuals and the squared standardized residuals [14].

It is remarkable that, according to Table 60.5, the standardized and the squared standardized residuals are not independent inter each other since a thinkable line is imprinted without special fluctuations and not a prices blend so that a positive residuals autocorrelation is submitted. The reasonable conclusion is the heredos-edicity existence. Attempting to reject the autocorrelation and heteroskedasticity

Table 60.7 ARCH–LM test for squared residuals

| Squared residuals lag (−1) | Squared residuals lag (−2) | Squared residuals lag (−3) | Squared residuals lag (−4) | F-statistic | Observation* R-squared |
|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------|---------------------------|
| 1 blend | | | | | |
| 0.001 (−0.03) | 0.001 (−0.93) | 0.001 (−0.60) | 0.0008 (−0.08) | 0.001 | 0.0016 |
| 2 blend | | | | | |
| 0.001 (0.72) | 0.0009 (−1.59) | 0.0005 (1.75) | 0.0003 (0.32) | 0.525 | 0.6161 |

existence suspicion in this model, we proceed to the above diagnostic test such as the Breusch-Godfrey and ARCH LM tests as it is showed in Tables 60.6 and 60.7.

We observe that the Probability in the 1st but also in the 2nd blend is >5 % which implies that we do not face autocorrelation issue.

Additionally, the residuals independence and the LB test results should be examined so we apply the ARCH Lm diagnostic test as it is represented in Table 60.7.

Next to the use of the ARCH-LM test with reference to the 4 lags in the residuals, we concluded that the variance does not show heteroskedasticity in the both specifications.

60.3 Conclusions

In the framework of our research on the fuels relation (Diesel + Animal Fats Biodiesel = blend 1 & Diesel + Vegetable Biodiesel = blend 2), we have come to the conclusion that there is an actual strong linear relation that connects the above blends with the heat of combustion value.

It is remarkable that the blend independent variable blend (1st and 2nd blend) has a strong linear relation to the heat of combustion as the index of adjusted $R^2 = 0.99$ and particularly in both cases is equal to 0.99 %. Normally, before the question index testing, we have already established the statistical model significance basing on the price p value and the statistical F. Likewise, we substantiated that the model variables are statistically significant as the H_0 case was rejected. Furthermore, the LB elements suitability test showed that there is residuals correlation with heteroskedasticity existence suspicion. However, the diagnostic test Breusch–Godfrey indicates that the suggested model does not include autocorrelation. The autocorrelation absence that is combined to the strong linear positive relation of the variables makes us conclude that the Eqs. 60.2 and 60.3 are able to preview the heat of combustion of the model fuels.

References

1. Ramesh D, Samapathrajan A, Venkatachalam P Production of biodiesel from *Jatropha curcas* oil by using pilot biodiesel plant. Agricultural Engineering College and Research Institute, India
2. Ma F, Hanna AM (1999) Biodiesel production: a review. *Bioresour Technol* 70(1):1–15
3. Furuta S, Matsuhashi H, Arata K (2004) Biodiesel fuel production with solid superacid catalysis in fixed bed reactor under atmospheric pressure. *Catalysis Commun* 5(12):721–723
4. Du W, Xu D, Liu D, Zeng J (2004) Comparative study on lipase-catalyzed transformation of soybean oil for biodiesel production with different acyl acceptors. *J Mol Catal B: Enzym* 30:125–129
5. Fassinou W, Sako A, Fofana A, Koua K, Toure S (2009) Fatty acids composition as a means to estimate the high heating value (HHV) of vegetable oils and biodiesel fuels. In: The 3rd international conference on sustainable energy and environmental protection, SEEP 2009
6. Parikh J, Channiwal SA, Ghosal G (2005) A correlation for calculating HHV from proximate analysis of solid fuels. *Fuel* 84(5):487–494
7. Demirbas A (2000) A direct route to the calculation of heating values of liquid fuels by using their density and viscosity measurements. *Energy Convers Manag* 41(15):1609–1614
8. Demirbas A (1998) Fuel properties and calculation of higher heating values of vegetable oils. *Fuel* 77(9–10):1117–1120
9. Sadramela SM, Seames W, Mann M (2008) Prediction of higher heating values for saturated fatty acids from their physical properties. *Fuel* 87(10–11):1776–1780
10. Greene D (1989) Motor fuel choice: an econometric analysis? TWLC ~ II. Ru.-A. vd. DA, NO. 3, pp x3–2.53. 1989 Rimed in Great Britain
11. Stiakakis E, Fouliras P (2009) The impact of environmental practices on firms' efficiency: the case of ICT-producing sectors. *Oper Res: Int J* 9(3):311–328
12. Tsanaktsidis CG, Christidis SG, Tzilantonis GT (2010) Study about effect of processed biodiesel in physicochemical properties of mixtures with diesel fuel in order to increase their antifouling action. *Int J Environm Sci Develop* 1(2):205–207
13. ASTM D4809-09a Standard test method for heat of combustion of liquid hydrocarbon fuels by bomb calorimeter (precision method)
14. Ljung GM, Box GEP (1978) On a measure of a lack of fit in time series models. *Biometrika* 65:297–303
15. Van Gerpen JJ, Shanks B, Pruszek R, Clements D, Knothe G (2004) Biodiesel production technology, subcontractor report pirated for the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy by Midwest Research Institute, National Renewable Energy Laboratory NREL/SR-510-36244, Battelle, July 2004
16. Knothe G (2006) Analyzing biodiesel: standards and other methods review. *JAOCS* 83(10):823–833
17. ASTM D1298-99 (2005) Standard test method for density relative density (Specific Gravity), or API gravity of crude petroleum and liquid petroleum products by hydrometer method
18. ASTM D445-06 standard test method for kinematic viscosity of transparent and opaque liquids (and calculation of dynamic viscosity)
19. ASTM D1744-92 standard test method for determination of water in liquid petroleum products by Karl Fischer reagent (Withdrawn 2000)
20. ASTM D664 Standard test method for acid number of petroleum products by potentiometric titration
21. Engle R (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom Inflation. *Econometrica* 50:987–1007

Chapter 61

The Retail Banking Adverse Selection: RCBS Calculator Solution

M. Hedvicakova, I. Soukal and J. Nemecek

Abstract The paper is focused on the field of retail core banking services market (RCBS) in the Czech Republic. The surveys of the European Union repeatedly identified fundamental market imperfection—the asymmetry of information. The specific manifestations of this problem are clients' high costs on search or the market overview, tariff opacity, product-tying and lack of comparison tools. At the end of the 2010 there was launched the web-accessible system of RCBS calculator that help the client to get the market overview individually adjusted to his or hers RCBS usage profile. The paper uses the BPM to introduce the system's data acquisition module that is a key part of process. The rest of the paper shows the possibility how to employ the data acquired during the process to even more help to reduce the costs on search for the optimal product according the price and range of the demanded services.

61.1 Introduction

This paper is focused on the retail core banking services market (therein after only as RCBS abbreviation) in the Czech Republic. As in other countries in the European Union, RCBS market in Czech Republic is known as non-transparent filed with very high frequency of changes and many hidden fees. Many foreign studies point at this fact.

At first we have to notice in 2005 the White paper of Services Policy 2005–2010 [1] where the sub goal of removing the undue barriers associated with all types of bank accounts and to improve the competition between service providers was

M. Hedvicakova (✉) · I. Soukal · J. Nemecek
Department of Economics, The Faculty of Informatics and Management, University of
Hradec Kralove, Rokitanskeho 62, 500 03, Hradec Kralove 3, Czech Republic
e-mail: tsanaktsidis@teikoz.gr

declared. The second part of this goal is synergy goal from an article 17 of the Regulation (EC) no. 1/2003 (for this paper purposes there can be shortly said that Commission understands that information asymmetry reduces the sector competition). After the suggested surveys and wide discussion there were compiled several documents, please see [2–4]. The main findings of mentioned sources will be shortly introduced and commented in the next two paragraphs. It was clearly stated that RCBS clients' decisions are constrained by information asymmetry and high switching costs. Mainly there is shown that there is very difficult product comparison of offers. The main factors can be found when combining the results of mentioned studies—the complexity of products being sold and the transparency of prices [5].

Various institutions and individuals are trying to build up a tool for easy comparison of banking fees on the market of the Czech Republic. These banking fees are often very unclear and include hidden charges. The most known projects are “RCBS Calculator” or “Chytrýhonzá.cz”. Topic of this article is focused on analysis and benefits of RCBS Calculator. This project runs longest time and each month 2,000 clients fills questionnaire.

61.2 The Expert System the Calculator

There was created The Calculator in the Czech Republic. This service compares costs for clients in individual offered accounts on the basis of their RCBS use and recommends the lowest costs products. Knowledge base of the Calculator contains the tariff data of 12 banks (more that 98 % of the RCBS market in the CZ) and their 45 accounts. Client only fills the detailed form concerning the usage of specific services—52 questions in total (25 questions with attached sub questions and three additional questions). All fills are saved so the data for this study were acquired from 18,000 respondents who used the form of the Calculator in the year of 2010. The main tool for further help to choose the optimal cost account requires the identification of the most important RCBS client profiles [6–10].

The application is focused on the calculation of monthly costs of RCBS and other banking services of the client on the basis of the client's use of banking services. Frequencies of monthly use of the services, or amounts utilized, are entered by the client into an electronic form, which is then saved on the server. The form is divided into logical chapters. It includes 52 questions in total (25 questions with attached sub questions and three additional questions) in chapters:

- I. Account,
- II. Statements,
- III. Card services,
- IV. Electronic banking,
- V. Payments—direct payments,
- VI. Payments—standing orders,
- VII. Payments—authorization for encashment (including SIPO),

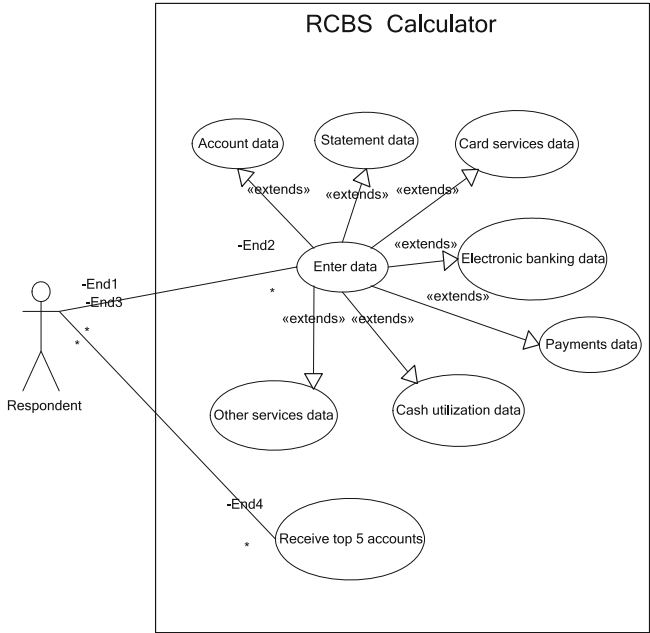


Fig. 61.1 RCBS calculator

VIII. Cash utilization,
IX. Other services.

Communication between Respondent and RCBS Calculator system is shown on Fig. 61.1 in UML notation. UML is a standard language for specifying, visualizing, constructing, and documenting the artifacts of software systems. Although UML is generally used to model software systems but it is not limited within this boundary. It is also used to model non software systems as well like process flow in a manufacturing unit etc.

There are a number of goals for developing UML but the most important is to define some general purpose modelling language which all modelers can use and also it needs to be made simple to understand and use.

UML diagrams are not only made for developers but also for business users, common people and anybody interested to understand the system. The system can be a software or non software. So it must be clear that UML is not a development method rather it accompanies with processes to make a successful system.

The goal of UML can be defined as a simple modelling mechanism to model all possible practical systems in today's complex environment [11].

Calculator also monitors "if—then" conditions, when for example clients are exempt from some charges, if the balance on their account is higher than the set limit, or if the turnover on the accounts exceeds the set limit. These conditions are for example in the offers of Raiffeisenbank, GE Money bank, Citibank and others.

The Calculator includes data about imposition of charges for the offers of 12 banks, or rather 44 various types of current accounts or the so-called package accounts offered in the Czech Republic.

When all necessary data are entered, the Calculator will compute the costs and arrange bank offers in a transparent manner from the most cost advantageous to the least cost advantageous with regard to the type of use of services entered by the client. Data about frequencies, amounts related to the account turnover and balance and the particular calculated amount of costs are saved on the server. All fills are saved so the database holds more than 20,000 respondent's answers. From the marketing research point of view there are gathered data:

- a. Multivariate—there has been monitored 53 variable concerning RCBS usage, two system variables for respondent identification and 45 variables containing the calculated costs for each of monitored RCBS product,
- b. Primary—data were gathered directly from the client,
- c. Subjective—data came from respondent himself, respectively it is his or hers subjective seem.

Due to specific data gathering process the data analysis outcome cannot be applied on the whole CZ population. Main limits that characterize the population of RCBS Calculator are connection to the Internet and basic IT and banks literacy. Still there can be expected that passive client with desk service preference are presented in the Calculator's database much less than e.g. internet banking preferred clients. For the adult low-cost banking where there is presumption of high internet banking preference, mentioned limitations are not that strict, still they must be respected when discussing the field of data relevance [6–10].

61.3 Modeling of the Calculator Questionnaire

There was described the expert system The Calculator in last chapter. There is described structure of communication while filling the questionnaire in next chapters for easier understanding.

There are modeling six chapter of questionnaire (IV. Electronic banking, V. Payments—direct payments, VI. Payments—standing orders, VII. Payments—authorization for encashment (including SIPO), VIII. Cash utilization, IX. Other services) for practical show. First three were published in [8].

All these questions from the questionnaire are crucial for the correct calculation of the total costs that clients pay to bank for a banking account. Currently there is being prepared more ergonomic version of the questionnaire, where clients will be asked only the most common used services. This version will be less time consuming, but not as accurate as current version. According latest analysis, most clients use only basic services from their bank accounts and these clients are target group of this more ergonomic version of the questionnaire. For example whole step 9 (Other services) will be skipped [13–15].

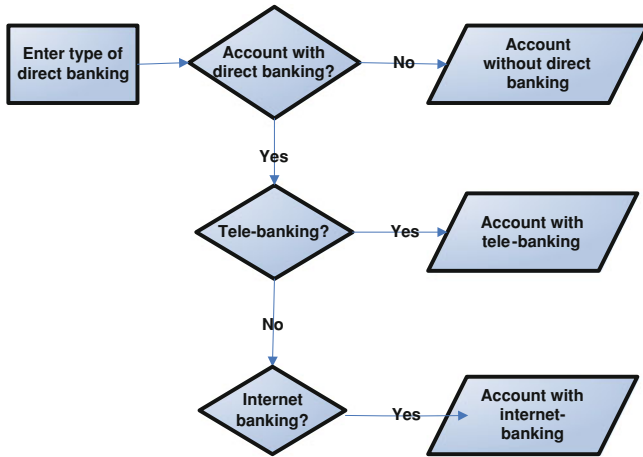


Fig. 61.2 Electronic banking

61.3.1 Electronic Banking

Fourth part is focused on electronic banking (see Fig. 61.2). Client decides, if he wants to communicate electronically with the bank. He can choose from internet banking or telebanking (both possibilities are possible only with some banks) [8].

61.3.2 Payments: Direct Payments

In this part of questionnaire clients fill number of incoming payments from other banks and number of incoming payment from own bank. He also needs to enter the way how direct payments are entered (see Fig. 61.3).

- Direct payments to own/other bank at desk,
- Direct payments to own/other bank by telebanking,
- Direct payments to own/other bank by box,
- Direct payments to own/other bank by Internet.

61.3.3 Payments: Standing Orders

In this part of questionnaire clients fill number of standing orders to own bank and to other banks. He also needs to enter the way how standing orders are entered (see Fig. 61.4).

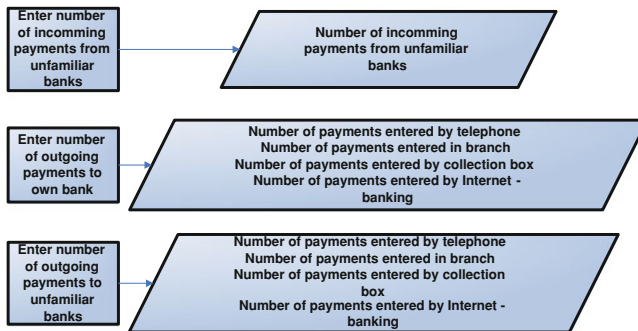


Fig. 61.3 Payments—direct payments

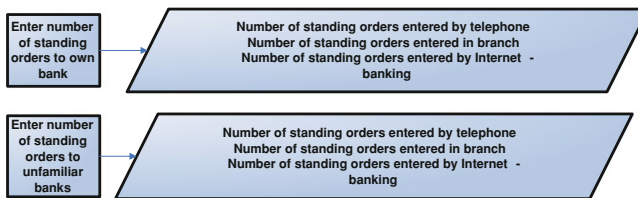


Fig. 61.4 Payments—standing orders

- Standing orders to own/to other bank at desk,
- Standing orders to own/to other bank by Internet,
- Standing orders to own bank/to other bank by telebanking.

61.3.4 Payments: Authorization for Encashment (Including SIPO)

In this part of questionnaire clients fill number of encashment to own bank and to other banks. He also needs to enter the way how encashment are entered (see Fig. 61.5):

- Encashment to own bank/to other bank at desk,
- Encashment to own bank/to other bank by Internet,
- Encashment to own bank/to other bank by telebanking.

61.3.5 Cash Operations

In this part of questionnaire clients fill number of cash operations, which he uses (see Fig. 61.6):

- Number of Cash deposit at desk,
- Number of Cash withdrawal at desk,

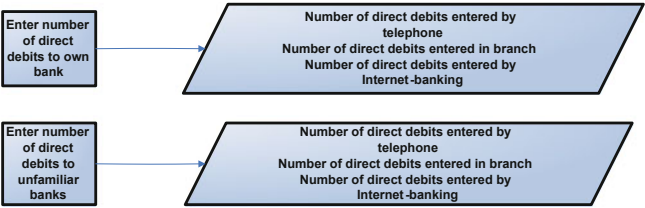


Fig. 61.5 Payments—authorization for encashment

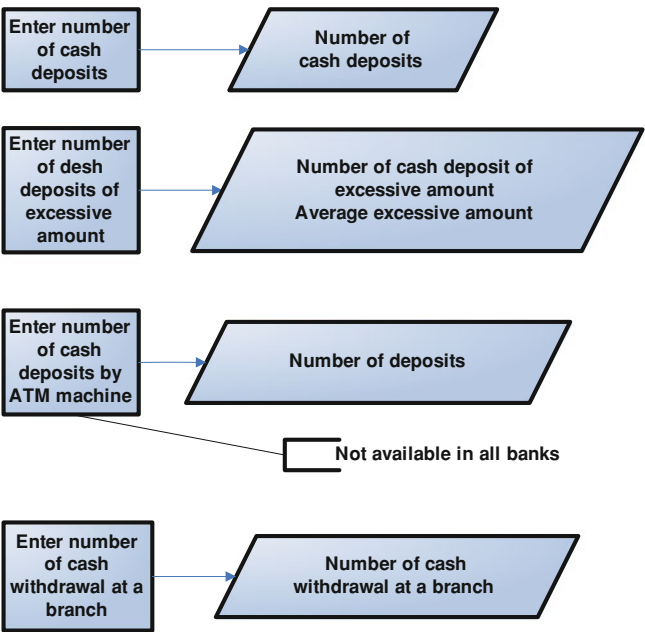


Fig. 61.6 Cash utilization

- c. Number of Cash deposit at ATM (this service is not available in all banks, available in for example GE Money Bank, UniCredit Bank and Ernste Bank).
- d. Number of cash deposits of excessive amount (number of deposits, excessive amount, and average excessive amount).

61.3.6 Other Services

Last part of questionnaire is focused on specific operations which most of clients do never use. But when these operations are used, they are charged very high. These operations are as follows (see Fig. 61.7—Other services):

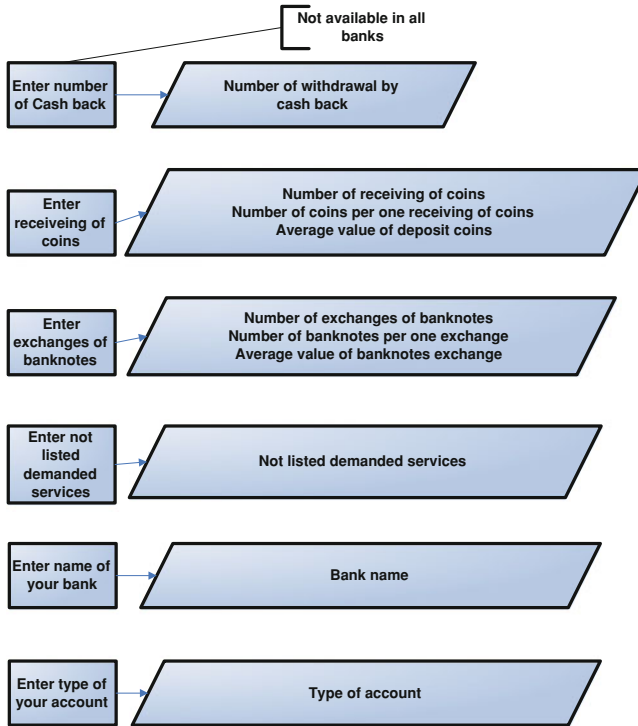


Fig. 61.7 Other services

- e. Cash back,
- f. Receiving of coins,
- g. Exchange of bank notes,
- h. Not specified services.

Clients need to fill in the questionnaire also the name of their own bank and the type of their account, which they are actually using. These questions have strong impact on calculating of their monthly cost of their potential banking account.

After filling all the questions, the Calculator offer clients five accounts (according entered specifications) with lowest monthly cost. These records are compared with entered values.

61.4 Banking Index Computation

As stated above, the Calculator usage brings a collateral benefit—consumer data. Those data can be, due to very specific data gathering process that offers consumer an added value, can be after the validation/verification, considered as high-quality

mainly because of very good level of detail. In [7] there was introduced cluster analysis output from the pilot run of the Calculator system. Because of the type of the data and the number of respondents there was chosen k-means clustering algorithm. There were four clusters identified by the demanded services and by the usage frequency. This can be and is used again to help the client to make a decision or get the market overview with almost no costs on search. Every quarter there are on the web side of the Calculator computed and presented charts of the 15 best RCBS accounts for each of the profiles, respectively for each of the cluster centroids. As an example, see Table 61.1.

Client just identifies the most suitable cluster where he or she belongs and then all what is needed is to see the table. Still there has to be mentioned that due to very specific data gathering process we can assume that all the analysis output concerns the population of RCBS client with activated ebanking, Internet connection and possessing the basic ICT literacy.

Computation of client profiles costs also gives clients an opportunity to estimate the pricing policy. Every account is more or less charged by fixed costs and the variable one. Typical fee from the fixed is the account management or card fee. Variable costs come mostly from the services such as ATM withdrawal, standing order etc. Let us compare passive profile and the average one costs on the best-price seven accounts, (see the Fig. 61.8). It is important to notice that both profiles has the same range of demanded services, they differ in the usage frequency.

There can be seen that the difference passive \times average profile costs is not constant. Account from mBank is free of charge in disregard of the profile (still the condition of card turnover has to met but the amount of 150€ per month is very easy to comply). PS ERA online has very low difference but the costs of the average profile are three times higher than for the passive one at Airbank. This overview is useful also in the situation when the client is not able to decide whether he or she belongs to the passive cluster or not. In case of a bit less active client but still not the passive one the differences between the two profiles using on the same product can help to identify the best one in case of nondescript usage pattern or higher usage variability.

61.5 Conclusion

In the Czech Republic there are one of the highest bank account charges in the European Union. The banks make use of the practices on the given banking market and the conservativeness of citizens of the Czech Republic who are not willing to flexibly adjust to market conditions.

For the reasons of clarifying charges at individual institutions, a number of projects are being raised, such as bank charges Calculator, which constantly monitors the level of bank charges and aims to select the most appropriate accounts according to specified criteria at the lowest monthly account maintenance costs.

Table 61.1 Average client cluster best 15 RCBS accounts in the third quarter of 2011

| Average client bank and account name | Month costs in € |
|--------------------------------------|------------------|
| mBank mAccount ^a | 0 |
| PS ERA online ^b | 1, 2 |
| Fio current account | 1, 9 |
| Airbank small tariff | 2, 4 |
| PS ERA online | 3, 2 |
| ZUNO account | 3, 2 |
| UniCreditBank partners account | 3, 9 |
| ZUNO account plus | 3, 9 |
| LBBW account 5 for 50 | 4 |
| PS ERA personal account | 4, 9 |
| ČSOB active account | 5 |
| LBBW IQ account | 5, 1 |
| Equa current account | 5, 2 |
| Equa bank On-line account | 5, 2 |
| Equa bank Prima account | 5, 4 |
| Passive client bank and account name | Month costs in € |
| mBank mAccount ^a | 0, 0 |
| PS ERA online ^b | 0, 6 |
| Airbank small tariff | 0, 8 |
| Fio current account | 1, 5 |
| ZUNO account | 2, 2 |
| PS ERA online | 2, 6 |
| LBBW account 5 for 50 | 3, 0 |
| Equa bank On-line account | 3, 3 |
| LBBW IQ account | 3, 4 |
| ZUNO account plus | 3, 4 |
| Equa bank Prima account | 3, 4 |
| UniCreditBank partners account | 3, 6 |
| PS ERA personal account | 3, 8 |
| KB my account ^c | 4, 1 |
| ČSOB active account | 4, 2 |

^a Meeting the condition of card over 150 “Euro Sing” and more

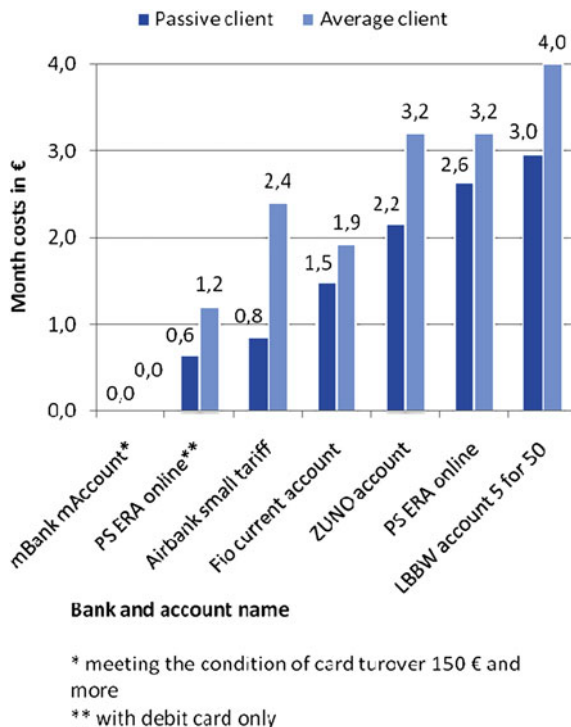
^b With debit card only

^c My account

Currently, the largest competitor of RCBS Calculator on the market in Czech Republic is “Chytryhonza.cz” [12] (available from URL: <http://supermarket.chytryhonza.cz/bezne-ucty-formular> # Step-1). Its structure is very similar to described RCBS Calculator. After filling all required data, respondent will receive overview of 39 types of banking accounts with their monthly fees.

There are also starting other projects on the internet, which keeps focus on this very actual topic in the Czech Republic. Major disadvantage, which is found on these public servers is, that they concentrate only on monthly fees. But this is not the only one factor. There are other factors, that customer must follow, while

Fig. 61.8 Comparison of 7 best accounts by costs for passive and average client cluster centroid



taking a decision where to keep money. Top factors can be as follows: number of ATM machines, number of branches, bank stability and so on.

The Calculator is very useful tool for choosing the most suitable banking account according the customer needs. Calculator also offers an information service for the customer (each month the customer is informed about actual development of banking costs on the market) so the customer can react very easily on the development on the banking market. This area on the market in the Czech Republic develops very rapidly and local banks adjust banking fees several times per year.

First condition of customer economical optimization of banking fees is to be properly, clearly and fast informed. When customers have possibility of receiving this kind of information, they can build up a constant pressure on banking market to stabilize so hectic and not predictable situation on the market.

Using the results of cluster analysis we identified the most common usage patterns. For each of them we regularly compute and publish on the Calculator website 15 best RCBS accounts considering the price and the services demanded. The costs vary from zero to five € per month. Still the client has to consider his or hers usage variability and so to compare not just the closet cluster but also the second closest one. Some banks almost do not differ from the passive usage pattern and the average one (such as mBank or Fio) but some does and very significantly (such as Airbank).

Popularity of this project also refers more than 2,000 filled questionnaires each month. There are also pointed its advantages and future perspectives of this useful service that was by more than 33,000 consumers so far.

It is up to customers of banks to start using available public information about banking fees on the market in the Czech Republic. Only customers can make a pressure on banks to start a change. Unfortunately, citizens of the Czech Republic are not really willing to change their habits and they are very rigid with their way of thinking. Calculator and similar project can only bring information, but only customer can transform information on action.

Acknowledgments This paper is written in the frame of specific research “Adverzní výběr v prostředí retailového bankovníctví”, translated as “Retail banking adverse selection”, project number 2105, funded by Czech Republic Ministry of Education, Youth and Sport.

References

1. Commission of the European communities (2005) EU policy paper—financial services policy 2005–2010. http://ec.europa.eu/internal_market/finances/policy/index_en.htm#Financial_Services_Policy_2005-2010. Accessed 26 July 2011
2. Commission of the European communities (2007) The green paper on retail financial services in the single market. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2007:0226:FIN:EN:PDF>. Accessed 26 July 2011
3. Commission of the European communities (2009) Study on the data collection for prices of current accounts provided to consumers. http://ec.europa.eu/consumers/strategy/docs/prices_current_accounts_report_en.pdf. Accessed 26 July 2011
4. Commission of the European communities (2011) Sector inquiry under article 17 of regulation (EC) No 1/2003 on retail banking (final report). eur-lex.europa.eu/LexUriServ/site/en/com/2007/com2007_0033en01.pdf. Accessed 26 July 2011
5. Soukal I, Hedvicakova M, Draessler J (2011) Central register systems: the information asymmetry reduction tool. In: World conference on information technology, Bahcesehir University & Near East University, 23–26 Oct 2011, Antalya, pp 15–21
6. Soukal I, Hedvicakova M (2011) Možnosti sledování trendu ve zpoplatnění základních bankovních služeb. *J Compet* 1:131–144
7. Soukal I, Hedvicakova M (2010) Retail core banking services e-banking client cluster identification. *Proc Comput Sci* 3:1205–1210. [s.l.]: Elsevier, ISSN: 1877-0509
8. Hedvicakova M, Soukal I (2011) Retail core banking services costs optimization. In: World conference on information technology, Bahcesehir University & Near East University, 23–26 Oct 2011, Antalya, pp 79–85
9. The Calculator (2011) <http://www.bankovnipoplatky.com/kalkulator.html>. Accessed 26 July 2011
10. Hedvicakova M, Soukal I (2010) Expert system as the tool for information asymmetry reduction on RCBS market in EU. In: International joint conferences on computer, information, and systems sciences, and engineering, SCSS 10, pp 155–164
11. Tutorials Point (2011) Simply easy learning. http://www.tutorialspoint.com/uml/uml_tutorial.pdf. Accessed 28 Aug 2011
12. Chytryhonza.cz (2011) <http://supermarket.chytryhonza.cz> (Online). Accessed 28 Aug 2011
13. Němecek J (2011) CRM a výsledky hospodaření vybraných firem. *J Compet* 1:75–81
14. Gordon AD (1999) Classification, 2nd edn. Chapman & Hall, London
15. Brian S et al (2001) Cluster analysis, 4th edn. Arnold Hodder Headline Group, London

Chapter 62

Project Management in Public Administration Sector

M. Hedvicakova

Abstract Administrative arrangements in the Czech Republic are divided into different regions, districts and municipalities with extended powers. Each region differs from each other. The biggest differences are in economic and social development and in ability of obtaining grants. Also very important is qualified knowledge of elective offices, because only qualified stuff can maximize financial effectiveness of economy of region. That is why University of Hradec Kralove began offering e-learning courses to enhance knowledge skills of employees of elected offices. Primary target of these courses are economical and managerial knowledge. At the University of Hradec Kralove there was originated a new course Project Management for Public Administration which is determinate to help agencies to get grants, streamline its operations and reduce regional disparities. Aim of this paper is to introduce e-learning course Project Management for Public Administration.

62.1 Introduction

The last 10 years have brought new requirements on corporate and public institution governance. More and more Czech (especially manufacturing) companies begin to apply project management for the needs of managing their projects. The state and public sector has also registered this trend. Even here are the guidelines and principles of project management beginning to apply more and more. Unfortunately, unlike the commercial sector for example, only in the

M. Hedvicakova (✉)

Department of Economics, Faculty of Informatics and Management, University of Hradec Kralove, Hradecka 1249/6, 500 03 Hradec Kralove 3, Czech Republic
e-mail: martina.hedvicakova@uhk.cz

smaller-scale projects (such as the realization of cultural events, etc.). In the public sector often takes place combining several separate projects, which reduces the resulting effectiveness of the projects. If the public sector will utilize project management to develop quality projects in line with the strategies of its bureau, region or the priorities of the Czech Republic, it will be able to make better use of the European Union funds. Project management provides a tool to a better and more efficient use of financial resources.

Project management is one of the tools to increase efficiency in the commercial, non-profit, state and public sector. By using these guidelines and principles it is possible to manage the projects more effectively and achieve the desired goals. For this reason, more and more private educational institutions and universities seek to offer training courses made-to-measure to the commercial and non-commercial organizations. The University of Hradec Kralove has also developed and implements the course “Project Management for Public Administration”, aimed at reducing disparities of Hradec Kralove and Pardubice regions.

62.2 Definition of Public Administration

We understand public administration as (a) certain type of activity (management of public affairs) and (b) institutions (organization, office) conducting public administration. In the materialistic (functional) approach it is public administration of state or other institutions activities, which by its content is not a legislative nor judicial activity. In the formal (institutional, organizational) approach is public administration defined as activity of the authorities designated as the administrative offices. The concept of public administration is a common (superior) term for the concept of state administration (which is performed primarily by state authorities), local administration (performed by local government authorities or bodies of interstitial/professional self-administration) and other public administration (performed mainly by institutions with legal subjectivity—such as General Health Insurance Company, Czech National Bank, Czech Television or Czech Press Agency) [1].

The local administration is a form of public administration consisting in the administration of territorially or differently organized commonwealth of people, which solves its own affairs independently and makes decisions directly or through elected bodies [2].

62.3 Description of Current Situation, Reasons to Resolve

According latest analysis of data from OECD, governance in Czech Republic is exercised in way, which is not very effective. In present time there is no comprehensive overview of what public administration is carried out, nor sophisticated

concept of what and how public administrations have to provide. Each agenda of public administration is often realized ill-conceived and executed spontaneously and are under ever-changing political priorities. There are only slow and partial optimizations without any inter-agency or more sophisticated approach.

The OECD recommends that the Czech government continues with its reforms, especially with the effort to increase competitiveness and the long-term growth. The organization also praises the government's target to improve the business environment, strengthen the education system, support innovation and reform the pension, social and health care systems. It especially welcomes the government's recently-passed Competitiveness Strategy. The document responds to the current situation and offers a comprehensive approach to structural reforms. According to Ángel Gurría, it will also help the Czech Republic in its transition to an innovative economy based on energy conservation [3].

62.4 Specifics of Public Administration

Public administration is specific compared to the commercial sector because it often combines investment with non-investment projects, which are primarily social in nature. It also mostly differs in diverse ways of financing. Public administration is currently largely using European Union funds to fund its projects. For this reason must be the municipality able to prepare quality projects, which can draw on these subsidies. State institutions unfortunately often face a lack of human resources and employees without the necessary knowledge and experience or, conversely, have no spare capacity to manage a further project.

This is the reason why at the University of Hradec Kralove, Faculty of informatics and management has originated a e-learning course "Project management for public administration" [4–6].

62.5 Project Management Methods

The goal of project management is to keep control over all activities that are needed for the project to create the expected product. Projects are the means for carrying out change, they are unique by their nature and so they involve higher risk compared to "business as usual" activities of the organisation. Therefore the implementation of a reliable, logical and proven project management approach is a good investment.

Project management methods are twofold:

Proprietary—these are neither available nor usable without the permission of their owners.

Non-proprietary—publicly accessible and freely usable for managing projects. Thanks to their free availability, non-proprietary methods are much more

Table 62.1 Most Important Project Management Standards—Comparison

| | PRINCE2 | PMI PMBOK guide | IPMA ICB | Critical chain |
|---|----------------------|----------------------|-------------------------|--|
| Process model | + | + | – | Just for time and resource management |
| Roles and responsibilities definition | + | + | Just Project Manager | + |
| Tools and techniques | + | + | – | Just for time and resource management |
| Document outlines/templates | + | – | – | – |
| Certification system for project management professionals | + | + | + | + |
| Accreditation of training materials | + | + | – | + |
| Accreditation of trainers | + | – | – | + |
| Accreditation of training organizations | + | + | – | – |
| Number of certified project management professionals worldwide | More than 400,000 | More than 400,000 | ca. 70,000 | Hundreds |

Source [7]

+ included, – not included

widespread, what bears multiple advantages. The most notable non-proprietary project management methods are PRINCE2, PMI and Critical Chain. We could also add the IPMA standard. However, as can be seen from Table 62.1, the IPMA standard is not a full project management method [7].

There are three ways how to gain knowledge about project management methodologies:

studying by yourself,
taking a training on the chosen project management method,
combination of the previous two alternatives [7].

62.6 Course Description and Goals

The e-learning course Project management for public administration is, based on the analysis of market needs, focused just on the public administration. This course focuses on beginners and does not require any initial knowledge. It serves mainly mayors and executive public authorities to obtain basic overview about the nature of project management. The e-learning course Project management for public administration was created within the solving project Management for solving regional and local disparities initiated by the Ministry of Regional Development in the Czech Republic. All course costs were reimbursed from funds of the solved research project [8–11].

The goal of this course is to provide basic findings on project management, its contents, extent and practical use. Upon successful completion of the course the trainees will be versed in the problematic of product management from project definition to its final evaluation [15–17].

One of the main goals is to enable the trainees to connect and use the gained knowledge and skills in practice (for instance to increase efficiency of current projects/investment actions, subsidy programs, grants/, in finding internal savings, streamlining the implementation of investment projects, creating high-quality timetables and schedules, expanding business possibilities in public administration, effectively manage project risks, etc.).

62.7 Course Structure

Absolving the whole course takes two and half month and the initial run was realized from 11th of April 2011 till 29th June 2011. Three non-compulsory meetings took place during this time with participation of 90 %.



Fig. 62.1 E-learning course “Project Management for Public Administration”

- The first meeting is devoted to the familiarization with form, content and progression of distant education, logging into the system and getting familiar with the WebCT environment. Then a course presentation takes place and the completion requirements are defined (a discussion and short final treatise on the use of project management at a particular workplace of individual trainee).
- The second meeting takes place in the middle of the course term and serves foremost to solving problems emerging during the studies and to practical examples of project management applications in public administration. Special attention is given to practical examples of projects realized in public administration.
- The third and last meeting is realized at the end of the course and serves to exchange of experience of individual course participants, defense of individual project and certificate presentation.

The course consists of 15 individual chapters devoted to the most important areas of project management and provides the trainees with theoretical and practical view of the given problematic with focus on public administration (see Fig. 62.1).

The advantage of this course is that the participants choose the depth of subject matter on their own and may use the supplied presentations, case studies, links to current articles and other according to their own initiative. The basic recommended texts are announced to individual participants every week by the tutor via e-mail. Other study materials are dependant on the activity and temporal possibilities of individual participants.

All the texts fulfill the distance material creation requirements and contain unified graphic structuring with stressing the most important constructs.

At each chapter is stated its goal, the most important constructs, keywords, temporal requirements, distinguished main and subsidiary text, summary, questions to ponder and used literature. Each chapter contains a link to a PDF file, which can be used offline.

Mean temporal requirement of each chapter is 20–30 min [12, 13].

62.8 Continue E-Learning Course Project Management for Public Administration II

After a successful pilot run of the course a continuation is planned for the first quarter of 2012 with creation of follow-up course for the participants of the basic course.

Follow-on Project Management Course for Public Administration II. will focus on theoretical knowledge of soft skills. These skills are now necessary for each employee in any professional direction, including in public administration. They help to better adapt to social situations in the workplace, to negotiate with people and promote their views, submit and present their ideas and projects to cope in crisis situations or cope with difficulties, such as a job interview into a job. The area under consideration will include, for example time management, teamwork and communication in the team, managerial skills, assertiveness and presentation skills [14].

The main objective of the soft skills in project management for public administration is to extend and complement the hard skills (course I.—Project management for public administration), including the know-how, the necessary theoretical knowledge of social skills that could be subsequently used in the workplace (course II.—continue). The layout of the individual modules of soft skills will be organized according to life of the project [14].

Students here learn also work in Microsoft Office Project 2010 and will be able to follow-on after successful completion of the course project manager to work in public administration.

There will be created short video spots, which will help students with orientation in the field [14].

62.9 Conclusion

The guidelines and principles of project management should be the same for both commercial and non-commercial sector. Unfortunately, this does not occur in practice. The non-commercial sector faces foremost the lack of financial and human resources. Another disadvantage is mainly due to the lower qualification of employees in the public administration, who are elected for a certain period of

time regardless of their qualification, knowledge and skills. For this reason emerges a number of training courses on the educational market that are specialized, according to current market needs, on the problematic or needed areas. This “Project Management for Public Administration” e-learning course was created at the University of Hradec Kralove as the seventh in the row after the previous e-learning courses for public administration (Marketing, Management, Municipal Management, Legislation, Local and Regional Development after the entry of Czech Republic into the EU, and Information and Communication Technologies in the Municipalities).

After the completion of the course and certificate presentation, the trainees completed a voluntary questionnaire on their satisfaction with the course. 8 trainees out of 10 have successfully completed the course. The remaining 2 trainees have not completed the course because of temporal reasons. The survey conducted shows that the course met the expectations and demands of the participants and that they would welcome, if the course had a sequel for advanced trainees or a similarly oriented course. The main objection was only to the time-consumption of the individual chapters, although the chapters took only 20–30 min. Due to the demands of participant’s individual work positions, who were mainly managers, it is difficult for them to find time for studies every week. Despite these problems most of them have found some time to study and devoted their attention even to the recommended supplementary materials. The course was beneficial for the group selected and they will put the knowledge acquired to practical use.

On this e-learning course of Project management for public administration will be followed by other e-learning courses in economics, management and marketing. Students will be able to comprehensively deepen the knowledge in different areas.

Acknowledgments This paper was prepared with the support of the grant WD-48-07-1—Management pro řešení regionálních disparit (translated as Management form solving regional disparities)—donated by the Ministry for Regional Development of the Czech Republic.

References

1. Hendrych D a kol (2009) *Správní právo—obecná část*, 7. Vydání. C.H.Beck, Praha, 838 pp, ISBN: 978-80-7400-049-2
2. Hendrych D (2009) *Právníký slovník*, 3. podstatně rozšířené vydání. C. H. Beck, Praha, 1460 pp, ISBN: 978-80-7400-059-1
3. BusinessInfo.cz (2011). <http://www.businessinfo.cz/cz/clanek/strategie-mezinarodni-konkurenceschopnosti/smk-efektivni-verejna-sprava/1001958/62119/>, 20 Sept 2011
4. Kadavova M, Slaby A (2006) ICT in education and models of virtual university. WSEAS Trans Adv Eng Educ 3(11): 649–656, ISSN 1790-1979
5. Kadavova M, Slaby A, Maly F (2008) Key factors involving the design of the system of virtual university. In: 7th WSEAS international conference on applied computer and applied computational science (ACACOS '08), Hangzhou, China, 6–8 Apr 2008. WSEAS Press, Lisbon, pp s.678–s.683, ISBN: 978-960-6766-49-7, ISSN: 1790-5117
6. E-learning course project management (2010). <http://www.oliva.uhk.cz/>, 20 Oct 2010

7. Potifob (2011). <http://www.potifob.com/>, 20 Sept 2011
8. Kadavova M (2007) Virtual education relation to economic profit of educational subjects. In: E-activities: networking the world, Proceedings of the 6th WSEAS international conference on e-activities (e-learning, e-communities, e-commerce, e-management, e-marketing, e-governance, tele-working/e-activities '07), Puberto De La Cruz, Tenerife, Canary Islands, Spain, 14–16 Dec 2007, ISBN: 978-960-6766-22-8, ISSN: 1790-5117
9. Slaby A, Kadavova M (2006) Process based modeling of virtual university. In: WSEAS 2006, proceedings of the 5th WSEAS international conference on education and educational technology, Tenerife, Canary Islands, Spain, 16–18 Dec 2006, CD: ISSN: 1790-5117, ISBN: 960-8457-57-2
10. Mohelska H, Hedvicakova M (2009) The influence of macroeconomics entities on city and village budgets. In: Proceedings of the 9th international conference liberec economics forum 2009, Sept 2009, Technical University of Liberec, pp 234–243, ISBN 9788073725235
11. Kadavová M (2007) Nové progresivní trendy ovlivňující ekonomiku vzdělávacích subjektů. In: Mezinárodní konference: finance a výkonnost firem ve vědě, výuce a praxi, Univerzita Tomáše Bati ve Zlíně, Fakulta managementu a ekonomiky, Zlín, ISBN 978-80-7318-536-7
12. Kadavova M (2008) Identifying the break-even point in distance courses. WSEAS Trans Adv Eng Educ 5(5):s.282–s.294, ISSN: 1790-1979
13. Hedvicakova M, Maresova P (2010) Actual and future trends of the E-education and its costs specification in the Czech Republic. In: Sobh T et al (eds) Technological developments in networking, education and automation. Springer, Dordrecht, pp 103–108, ISBN 978-90-481-9150-5
14. Inflow Roč.2, Number 11 (2009). <http://www.inflow.cz/kpm-kurz-projektoveho-managementu>, 20 Sept 2011
15. Maresova P, Hedvicakova M (2010) The state of knowledge management in Czech companies. In: Sobh T et al (eds) Innovations in computing sciences and software engineering. Springer, Dordrecht, pp 161–166, ISBN 978-90-481-9111-6
16. Maresova P, Hedvicakova M (2010) Costs and benefits in knowledge management in Czech enterprises. In: Sobh T et al (eds) Innovations and advances in computer sciences and engineering. Springer, Dordrecht, pp 433–437, ISBN 978-90-481-3657-5
17. Hedvicakova M, Soukal I (2010) Expert system as the tool for information asymmetry reduction on RCBS market in EU. In: Sobh T et al (eds) Innovations and advances in computer sciences and engineering. Springer, Dordrecht

Chapter 63

Access Point Checking to Improve Security in Wireless Infrastructure Networks

Ammar Odeh and Miad Faezipour

Abstract Security issues are taken into consideration for many applications and are also an integral part of computer networks. On one hand, security especially comes into picture when transferring data from one device to another. On the other hand, rapid migration of networks from wired to wireless increase security complications among these networks. This research paper, initially describes the security issues over wireless infrastructure networks, and identifies different types of attackers. The 802.11 standard security mechanisms as well as some encryption techniques such as RSA are discussed. A new methodology called Access Point Checking is then proposed which relies on a checksum-bit check at the access point before completing the data transfer. This technique outperforms traditional security mechanisms in terms of timing characteristics.

63.1 Introduction

63.1.1 Background

With rapid development in security techniques and the use of computers, automated tools that protect files and other information stored in the computer are highly in demand. Especially, distributed systems such as time-sharing systems or access over public telephone and/or data networks call for this necessity. The

A. Odeh (✉) · M. Faezipour
Department of Computer Science and Engineering, University of Bridgeport, Bridgeport,
Bridgeport, CT 06604, USA
e-mail: aodeh@bridgeport.edu

M. Faezipour
e-mail: mfaezipo@bridgeport.edu

generic name for such tools designed to protect data from hackers; is computer security [1, 2].

In distributed systems, the use of networks and standard communication protocols facilitate data transmission between a terminal user and a computer—and between a computer and another computer [2, 3]. Network security measures the need to protect data during transmission. Clearly, wireless networks are less secure compared to wired networks. So, the most important question here is how to protect data transmission in wireless networks.

63.1.2 Main Contribution and Paper Organization

An efficient wireless infrastructure algorithm is presented in this paper. The goal is to save the time consumed during message travel from one host to another in the network, while maintaining message security. We employ a checksum mechanism to enhance message integrity. In addition, access point (AP) will check the message and decide whether the message should be sent back to the original sender or not. Our algorithm divides the connection into two paths. The first path is from the sender to AP and the other one is from AP to the receiver. Depending on the path that establishes a successful connection, the algorithm will automatically identify which path possibly has an intruder/attacker. Moreover, the reliability of the system will be increased using the access point checking. In case an attacker is identified through incorrect message transfer, the insecure channel will be removed and will be saved in a history file.

The rest of this paper is organized as follows. In [Sect. 63.2](#), we briefly glance at the main structure of different wireless networks. Security requirements are then discussed in [Sect. 63.3](#). In [Sect. 63.4](#) we describe different classes of attackers. General standards for security mechanisms are discussed in [Sect. 63.5](#) and encryption algorithms are described in [Sect. 63.6](#). In [Sect. 63.7](#), we describe our proposed algorithm for access point checking, and evaluate the timing performance with respect to the conventional technique. Finally, concluding remarks are in [Sect. 63.8](#).

63.2 Wireless Networks

Wireless networks can be divided into two categories. The first category is ad-hoc networks. IEEE defines ad-hoc as a network having an independent basic service set (IBSS) [4, 5]. In the ad-hoc mode, each client communicates directly with the other clients within the network ring, as shown in [Fig. 63.1](#).

The second mode is the infrastructure mode. In IEEE standard, it is defined as a basic service set, which means each client sends all of its communications to a central station, or access point (AP) [5, 6]. The access point acts as an Ethernet

Fig. 63.1 Ad-hoc network**Fig. 63.2** Infrastructure network

bridge and forwards the communications onto the appropriate network—either the wired network, or the wireless network, as shown in Fig. 63.2.

63.3 Security Requirements

Before explaining the main security mechanisms in wireless networks, one must know the three main security goals:

- **Integrity:** this means that data can be modified by only authorized parties. Modifications include writing, changing, deleting, and/or creating messages. One popular way to achieve this is to have an integrity checksum field, as shown in Eq. 63.1.

$$\text{Checksumbit} = \begin{cases} 1 & \text{if even \# of 1s} \\ 0 & \text{if odd \# of 1s} \end{cases} \quad (63.1)$$

For instance, consider the following:

Message: 1101

Check sum bit will be: 0

Then, the sent message will be: 1101 0

- **Availability:** data should be available to authorized users.
- **Confidentiality:** only authorized users can access to data.

63.4 Categorization of Attackers

Since networks use air as the transmission medium, many unauthorized users may interfere. These users, also known as attackers/intruders, try to catch data transferred between stations. Attackers can be classified into two categories which are explained in the next subsections.

63.4.1 *Passive Attackers*

The objective is to obtain information that is being transmitted without any modification. So, security mechanisms should be able to deal with both types of passive attacks:

- *Release of message contents*: This passive attack means easily understanding the message content such as telephone connection or electronic mail message, and/or any file containing sensitive data in any explicit way.
- *Traffic analysis*: This passive attack is a process to analyze encrypted message to deduce the plain text. It also depends on language. For example, in the English language, letter {E} is the most common, followed by {T, R, N}, where {Z} rarely appears.

63.4.2 *Active Attackers*

Unlike passive attackers, active attackers involve some modification of the data stream or the creation of a false stream.

63.5 Standard Security Mechanisms

In this section, we refer to some standard security mechanisms in wireless networks.

63.5.1 *Open System Authentication*

In this protocol, the sender and destination receiver do not share a secret key. Each one generates a key-pair and sends a request to the other. If request is accepted, the connection is established for a short time. The key generation process will occur frequently and randomly, which prevents the attacker to predict the key or analyze the message [5, 7, 8].

63.5.2 Closed Network Access Control

The network administrator will determine the network strategy; whether it is open or closed. An open network will accept any foreign connection depending on the administrator's decision. The foreign user will send a join request to the network manager, who will in turn grant or revoke the connection. In a closed network, the first step is to determine the group whose members are acceptable participants of the network. In a closed network, the network members are updated at each time slot depending on the network range [9].

63.5.3 Access Control Lists

It is a static mechanism, where a list of MAC addresses for authoritative users are registered before connection setup. After registration, no one else can be a member of the network. In other words, the connection will be closed to other members whose MAC addresses are not registered. So the permission will depend on MAC addresses [10].

63.5.4 Wired Equivalent Privacy Protocol

Wired Equivalent Privacy (WEP) is a security protocol for local area networks. It is employed at the two lowest layers of the OSI model—data link layer and physical layer [4]. However, wireless networks use radio waves as a transmission media, do not have the same physical structure like wired LAN, and are therefore more vulnerable to tampering.

63.6 Encryption Algorithms

Encryption algorithms can be divided into two categories:

- *Conventional Encryption*: As shown in Fig. 63.3, the conventional encryption scheme has five steps [7]:
 1. Plain text: This is the original message that the sender wants to send. Plain text is fed into the algorithm as the input.
 2. Encryption algorithm: Represents a set of steps to convert the plain text to cipher text.
 3. Secret key: Represents an agreement key between sender and receiver which is kept secret to encrypt or decrypt messages.

Fig. 63.3 Conventional encryption

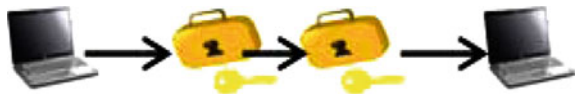


Fig. 63.4 Basic WEP encryption

| | | |
|--------------------|------------|------|
| <i>IV+KEY</i> | <i>RC4</i> | 0101 |
| <i>Plain Text</i> | | 1100 |
| | | = |
| <i>Cipher Text</i> | | 1001 |

4. Cipher text: This is unreadable data that will be a result of the encryption process and will be sent through the routing path from sender to receiver.
5. Decryption algorithm: It is a set of steps to transform cipher text to plain text.

There are two requirements for secure use of conventional encryption:

1. A strong encryption algorithm is needed.

2. Sender and receiver must have obtained copies of the secret key in a secure fashion and must keep the key secure.

For example WEP relies on a secret key (*K*) shared between the communicating parties to protect the body of a transmitted frame of data. Encryption of a frame proceeds as appeared in Fig. 63.4 [11].

In this technique, an initialization vector (*IV*) and a key *K*, representing the RC4 algorithm are chosen. This key stream is denoted by RC4 (*IV*, *K*). Then, an exclusive-or (XOR) operation of the plain text (*P*) with the key stream is performed to obtain the cipher Text (*C*):

$$C = P \otimes RC4(IV + K)$$

(63.2)

Finally, the *IV* and the cipher text (*C*) are transmitted over the radio link where all parties have the key.

- *Public-key Encryption Algorithm:* One of the most famous encryption algorithms is RSA, which stands for Rivest, Shamir and Adleman, who first publicly described it. Public key is distributed to all users on the network, which allows them to encrypt the sending message. Private key will be kept secret to transfer cipher text to plain text [11]. Basically, RSA enhances two goals:

1. *Data Confidentiality:* Data confidentiality ensures that information is accessible only to those authorized to have access. In other words, the data will be encrypted (denoted as *E*) using the public key of the receiver as shown in Eq. 63.3, and decrypted (denoted as *D*) by the private key of the receiver, as shown in Eq. 63.4.

$$E_{K_{Receiver}}(P)$$

(63.3)

$$D_{K_{Receiver}}(C)$$

(63.4)

In the above equations, KU stands for public key, P stands for plain text, KR is the private key, and C is cipher text.

2. *Authentication*: Authentication is a procedure to be sure about the source of the message (finger print). The receiver can recognize the source by his private key. In other words, sender will encrypt plain text by his private key (Eq. 63.5). At the receiver end, decryption process will employ public key of the sender (Eq. 63.6).

$$E_{KR_{\text{sender}}}(P) \quad (63.5)$$

$$D_{KU_{\text{sender}}}(C) \quad (63.6)$$

63.7 Proposed Algorithm

Our proposed algorithm suggests adding checksum to the encrypted message. In infrastructure wireless networks, the access point (AP) will arrange the connection between two parties and control the connections of all clients in the network. In our method, we add a new duty for AP to check the message and determine if it can be sent, or if it contains unsecure data. This task is performed by applying a checksum bit.

63.7.1 Algorithm Pseudo code

Our algorithm involves a few steps which is shown in the following pseudo-code:

Step 1. Apply encryption to plain text by using public key algorithm.

Step 2. Add check sum to cipher text.

Step 3. The encrypted message will be sent by AP address, which will act as receiver and checker.

Step 4. Access Point will test check sum ($Csumbit$) and set counter = 0

Test($Csumbit$)

If correct go to *Step 5*

Else do

While (counter < 3){

Counter = counter +1

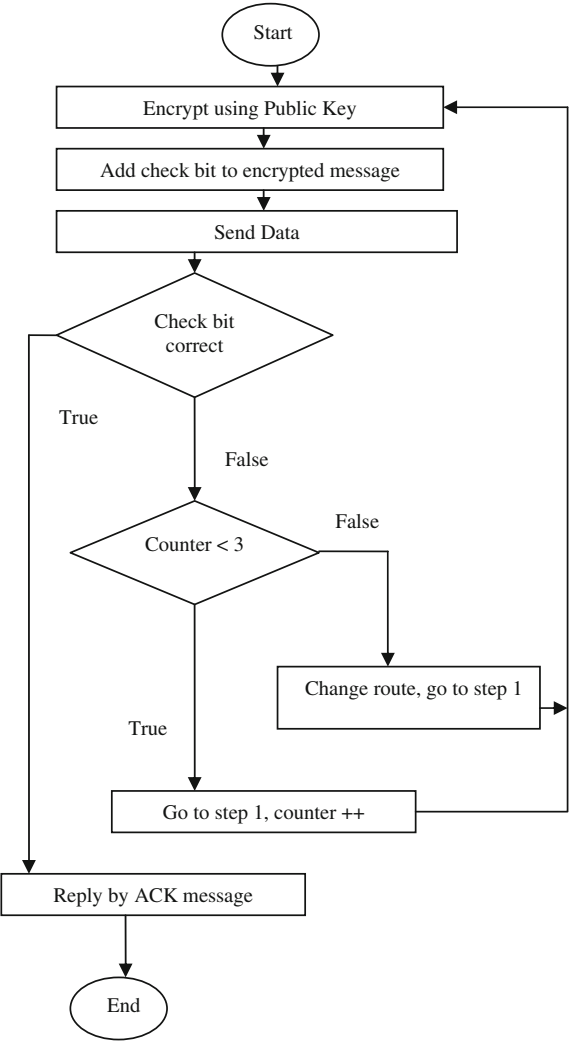
Go to *Step 1*}

If counter == 3 then route is unsecure, change the route. Go to *Step 1*

Step 5. Send message to end receiver. Decrypt it, then reply using any acknowledgment to confirm message delivery.

Step 6. End connection.

Fig. 63.5 Flowchart of the access point check algorithm



The flowchart of the algorithm is also shown in Fig. 63.5. Our algorithm has certain properties which are explained hereafter.

63.7.2 Advantages

The algorithm is rather simple, as it depends on a linear function. Our technique will have the following advantages:

1. A more reliable connection will be established compared to the conventional approach, since the unsecure route will be easily discovered. If there is an attacker in the link, it will be discovered and will be removed when the connection is re-established.
2. Time will be reduced in case of incorrect message transfers. The time required to deal with incorrect messages is:

$$t_{\text{incorrect}} = 2t_{s-ap} + t_{\text{detect}} \quad (63.7)$$

where:

$t_{\text{incorrect}}$: time consumed if message is incorrect.

t_{s-ap} : time consumed from sender to access point.

t_{detect} : elapsed time to detect message status (correct or incorrect).

In our approach, the time elapsed from the sender to AP is computed twice, because the message is returned to the sender, if the message is incorrect. In the previous (conventional) approach; check over plain text, the elapsed time for incorrect messages can be calculated from the following equation:

$$t_{\text{incorrect}} = 2(t_{s-ap} + t_{ap-r}) + t_{\text{detect}} \quad (63.8)$$

where:

t_{ap-r} : time required to transfer from AP to the receiver

In the conventional approach, the total time is considered to be the message travel time from sender to AP and from AP to receiver, plus the time consumed to detect the message status. However, the time the message travels from sender to AP and from AP to receiver is doubled if the message is incorrect, simply because the message would return.

63.7.3 Disadvantages

The proposed technique will have the following disadvantages:

- (1) Checksum is a linear function, therefore, the error cannot be easily detected:

$$CRC(X \text{ XOR } Y) = CRC(X) \text{ XOR } CRC(Y) \quad (63.9)$$

If the attacker modifies the message such that the checksum bit is the same as the correct message checksum, AP cannot detect it. For example, if the encrypted message is 1001, then $(C, c(C)) = 1001 \ 1$. If the attacker modifies the message to 10101, there is no way to figure out the message was incorrect, because the check sum bit is the same for both messages.

- (2) It is a more complicated mechanism because the access point requires more intelligence and more resources to store the error message and to send the acknowledgment to the sender.
- (3) If there is no attacker in the link, the message will be transferred correctly, but the consumed time will be:

$$t_{\text{correct}} = t_{s-ap} + t_{ap-r} + 2t_{\text{detect}} \quad (63.10)$$

In our approach, the status detect time is computed twice for correct messages, since the checksum analysis (status detection) is also performed at the receiver end. However, in the conventional approach, the elapsed time will be:

$$t_{\text{correct}} = t_{s-ap} + t_{ap-r} + t_{\text{detect}} \quad (63.11)$$

From the above analysis, it can be seen that our access point checking algorithm outperforms the conventional encryption method, in the event that an attacker/intruder interferes the link with incorrect messages.

63.8 Conclusion

Developments in wireless networks require developments in security methodologies to ensure secure data transmission. Furthermore, the time/complexity/cost should simultaneously be taken into consideration in the design of such security methodologies. In this paper, we introduced an Access Point Checking algorithm that reduces the overall data transfer timing, while ensuring a secure message transfer. In the future, we plan to run sophisticated simulations for emulating wireless infrastructure networks in the presence of attackers to confirm our proposed improved performance.

References

1. US-CERT (2006) A Government Organization, Using Wireless Technology Securely pp 1–9
2. Kumar V, Sharm R (2011) Behavioral study of dynamic routing protocols for MANER. *Inter J Comput Bus Res* 2(2):1–10
3. Liu X, Goldsmith A (2004) Wireless network design for distributed control. In: *Proceedings of the 43rd IEEE Conference on Decision and Control*, vol 3, pp 2823–2829
4. National Institute of Standards and Technology (2008) *Guide to Securing Legacy IEEE 802.11 Wireless Networks*
5. Arbaugh WA, Shankar N, Wan YCJ, Zhang K (2002) Your 802.11 Wireless Network Has No Clothes. *IEEE Wireless Commun* 9(6):44–51
6. Mahan R (2001) *Security in wireless network sans institute*, pp 1–16
7. Laing, A (2001) *The security mechanism for IEEE 802.11 wireless networks*, version 1.2f

8. Shin M, Mishra A, Arbaugh W (2006) Wireless network security and interworking. In: Proceedings of the IEEE, 94(2):455–466
9. Third Generation Partnership Project (2003) 3G Security; security architecture (Release 6), 3GPP TS 33.102 v6.0.0,” 3GPP Technical Specifications
10. Hottell M, Carter D, Deniszczu M (2006) Predictors of home-based wireless security. In: Proceedings of the 5th Workshop on the Economics of Information Security, pp 1–12
11. Stallings W, Brown L (2007) Computer security: principles and practice”, 2nd edn, Prentice Hall

Chapter 64

Comparison of Fractional PI Controller with Classical PI using Pareto Optimal Fronts

O. J. Moraka

Abstract A comparison between the fractional order PI (FOPI) controller and integer-order PI (IOPI) controller, and also between FOPI and IOPI with lead circuit, is conducted in this paper. The performance comparisons are conducted for a servo system. The FOPI controller is a controller with fractional-order integral, where the order parameter can be used to adjust the closed loop response of the servo system. The controllers are compared using the Pareto Optimal Front (POF) of each controller. The FOPI is approximated using Oustaloup's approximation method. The approximated FOPI cannot achieve the positive values of the ideal FOPI in the phase-magnitude Bode plot. The POFs show that the FOPI is not always better than IOPI. The IOPI with lead circuit achieves performance which is similar to FOPI. Thus that the IOPI controller with lead circuit can be used to obtain a performance that closely matches the performance of the FOPI controller. The implemented FOPI controller did not track the setpoint. The IOPI controller, and also with a lead circuit, tracked the setpoint.

64.1 Introduction

Fractional calculus deals with the investigations and applications of integrals and derivatives of arbitrary order (real or complex order) [1]. In 1999, Podlubny proposed the fractional-order PID controller which is an extension of the classical PID controllers with fractional (real) order derivatives and integrals [2]. The fractional orders can be used to enhance the system's performance and adjust dynamics of a control system [3].

O. J. Moraka (✉)
Department of Electrical Engineering, University of Cape Town,
Rondebosch, South Africa
e-mail: mrkots001@uct.ac.za

In 2006, C. Zhao, Y. Q. Chen and D. Xue showed that under the same optimization conditions, the optimal fractional-order PID controller outperforms the optimal integer-order PID [4]. In 2009, C. Onat, N. Tan and C. Yeroglu proposed a tuning rule for fractional PID controller. The rule was shown to outperform the classical PIDs, which were designed using the Ziegler-Nichols rules and Åström-Hägglund method [5]. In 2010, S. Jesus, S. Barbosa and J.A Tenreiro Reference used the Ziegler-Nichols heuristic rules to find the gains of the integer-order PID. Afterwards, experiments on a servo system were conducted to show that derivative and integral order can enhance the system's performance [3].

In this paper, the fractional-order PI (FOPI) controller's performance is compared with the integer-order PI (IOPI) controller's performance, on a servo system, using Pareto Optimal Fronts. The simulation results show that the FOPI controller is not always better than the IOPI controller. This paper also compares the FOPI and IOPI with lead circuit because S. Jesus, S. Barbosa and J.A Tenreiro [3] showed that the FOPI improved the phase margin, which means that the dynamic control action of the FOPI controller could be achieved by using an IOPI controller with lead circuit. The simulation results reveal that performance of the IOPI controller can be improved by a adding a lead compensator. The POF of the IOPI controller with circuit is closer to the POF of the FOPI, which means the lead circuit parameters can be used for achieving similar results to FOPI and tuning. The FOPI is approximated with an integer-order system. The phase-frequency plot from the Bode plot of the approximated FOPI shows that positive values of the phase, of the exact FOPI, cannot be approximated.

64.2 Fractional Order Systems

64.2.1 Fractional Calculus

A fractional-order system is characterized by a transferfunction of real order. The transfer function can be as expressed as

$$G(s) = \frac{1}{a_n s^{\beta_n} + a_{n-1} s^{\beta_{n-1}} + \dots + a_0 s^{\beta_0}}, \quad (64.1)$$

where β_k ($k = 0, 1, \dots, n$) is an arbitrary real number $\beta_n > \beta_{n-1} > \dots > \beta_0 > 0$ and a_k ($k = 0, 1, \dots, n$) is an arbitrary constant [2].

In the time domain, the fractional-order system (64.1) corresponds to a fractional-order differential equation

$$a_n D^{\beta_n} y(t) + a_{n-1} D^{\beta_{n-1}} y(t) + \dots + a_0 D^{\beta_0} y(t) = u(t) \quad (64.2)$$

where $D^\gamma \equiv {}_0 D_t^\gamma$ is the Caputo's fractional derivative of order γ with respect to variable t and with initial point of $t = 0$. Caputo's fractional derivative is expressed as

$${}_0D_t^\gamma = \frac{1}{\Gamma(1-\delta)} \int_0^t \frac{y^{m+1}(\tau) d\tau}{(t-\tau)^\delta}, \quad (64.3)$$

$$\gamma = m + \delta, m \in \mathbb{Z}, 0 < \delta < 1$$

where $\Gamma(z)$ is the gamma function [2].

In control systems, systems are analyzed using Laplace transforms. The Laplace transform of the fractional derivative is given as

$$\mathcal{L}\{{}_0D_t^\gamma y(t)\} = s^\gamma Y(s) - \sum_{k=0}^{m-1} s^{(\gamma-k-1)} y^{(k)}(0), \quad (64.4)$$

where $Y(s) = \mathcal{L}\{y(t)\}$ [2]. When the initial conditions (64.4) are zero (64.4) becomes

$$\mathcal{L}\{{}_0D_t^\gamma y(t)\} = s^\gamma Y(s). \quad (64.5)$$

64.2.2 Fractional PID

In 1999, Podlubny [2] proposed a generalization of the classical PID controller. The controller is called the $PI^\lambda D^\mu$ -controller because it has the integrator of order λ and a differentiator of order μ . The controller's transfer function has the form

$$K(s) = \frac{U(s)}{E(s)} = K_p + K_I s^{-\lambda} + K_D s^\mu \quad (\lambda, \mu > 0) \quad (64.6)$$

where $K(s)$ is the transfer function of the controller, $E(s)$ is the error and $U(s)$ is the control action [2].

The time domain equation of the $PI^\lambda D^\mu$ -controller's output is

$$u(t) = K_p + K_I J^\lambda e(t) + K_D D^\mu e(t) \quad (64.7)$$

where J^λ is the integration operator and D^μ is the differentiation operator. If $\lambda = 1$ and $\mu = 1$, a classic PID-controller is obtained. Taking $\lambda = 1$ and $\mu = 0$ a PI-controller is obtained and for $\lambda = 0$ and $\mu = 1$, gives a PD-controller. Lastly, for $\lambda = 0$ and $\mu = 0$ a proportional-controller is obtained. The classical types of PID-controllers are the particular cases of the fractional $PI^\lambda D^\mu$ -controller [2].

64.3 Oustaloup's Approximation Method

To implement a fractional order derivative and integral operator, a frequency domain approximation with integer-order transfer function is used. The Oustaloup filter approximates the operators in a specified frequency range (w_b, w_h) and of order N [4]. Oustaloup's recursive filter is expressed as

$$s^\gamma = K \prod_{k=-N}^N \frac{s + w'_k}{s + w_k}, \quad 0 < \gamma < 1 \quad (64.8)$$

where w'_k , w_k , and K are obtained from [4]

$$\begin{aligned} w_k &= w_b \left(\frac{w_h}{w_b} \right)^{\frac{k+N+\frac{1}{2}(1-\gamma)}{2N+1}} \\ w'_k &= w_b \left(\frac{w_h}{w_b} \right)^{\frac{k+N+\frac{1}{2}(1-\gamma)}{2N+1}}, \quad K = w_h^\gamma. \end{aligned} \quad (64.9)$$

64.4 Experimental Apparatus

The servo system that was used in experiments is shown in Fig. 64.1. A computer program, in Visual Studio C# 2010, communicates with the DAC and ADC which connect the servo system. The transfer function of the servo system is

$$G(s) = \frac{1.65}{0.61s + 1} \frac{[Volts]}{[Volts]} \quad (64.10)$$

64.5 Pareto Optimal Sets/Fronts

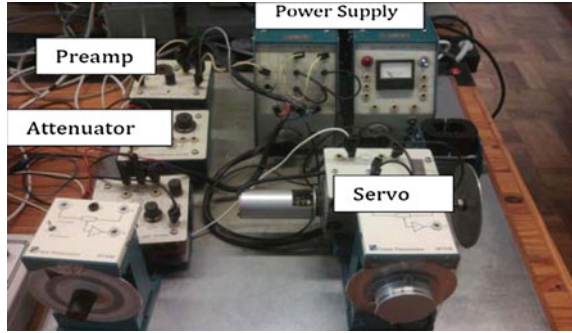
Pareto Optimal Fronts are used in control systems for optimization. To optimize a control system there are normally multiple objectives that are usually in conflict with one another. When there are conflicting requirements a set of optimal solutions are found and the solution that suits the particular problem can be selected. The set is known as Pareto Optimal Set. Each solution in the set is optimal because no improvement can be attained on one optimization objective that does not lead to the degradation in at least one of the other objectives [5].

Multi-objectives Optimisation problems that are not constrained can be described as:

$$\text{Minimize: } f_i(x), \quad i = 1, \dots, m \quad (64.11)$$

where, $f_i(x)$, $i = 1, \dots, m$ are the cost functions or objective functions and $x = (x_1, \dots, x_n)$ are the decision vectors. A decision solution x_1 is non-dominated if there exists no other candidate solution x_2 for all i values, $f_i(x_2) \leq f_i(x_1)$ and at least one i , $f_i(x_2) < f_i(x_1)$. The Pareto Optimal Set is the non-dominated solutions and the corresponding set of cost function values for the set is called the Pareto Optimal Front [5].

Fig. 64.1 Servo system for experiments



The cost functions that are used in this investigation are the integral of squared errors (ISE) and integral of squared inputs (ISI).

64.5.1 Decision Space for IOPI

The IOPI that is used for comparison has the form

$$K_{IOPI}(s) = \frac{K_p(s + I)}{s}, \quad (64.12)$$

where I is the zero position of the controller and K_p is the gain of the of the controller. The parameters, I and K_p , form a 2-dimensional decision space. The parameters of the IOPI are constrained to satisfy the following specification of the closed response to step inputs:

- The transients must have $0.1s < t_{\pm 37\%} < 0.61s$.
- The damping factor must exceed 0.4.

The upper limit of the time response specification means the closed loop poles must be to the left of the open loop poles in the s-plane. The lower limit of the time constant was chosen to be a pole at 10. The decision space was found by using Routh-Hurwitz Criterion and root locus methods. The Decision space is shown in Fig. 64.2.

The I values are rounded to one decimal place and K_p values are rounded to two decimal places. I and K_p are independent of one another.

64.5.2 Decision Space for FOPI

The FOPI that is used for comparison has the form

$$K_{FOPI}(s) = \frac{K_p(s^\alpha + I)}{s^\alpha}, \quad (64.13)$$

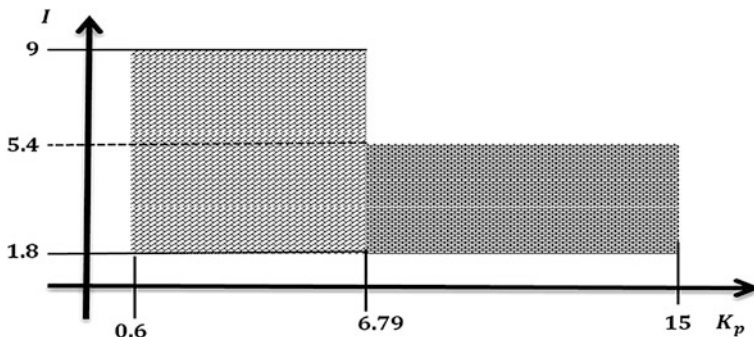


Fig. 64.2 Decision space for IOPI controller

where I is a parameter that affects the zero position of the controller, K_p is the gain of the controller and α is the order of the FOPI controller. The parameters, I , K_p and α , form a 3-dimensional decision space. The parameters I and K_p , are the same as in Fig. 64.2 and $0 < \alpha < 1$, because higher values will mean a controller with more degree than the IOPI which is not reasonable [7, 8].

64.5.3 Lead Compensator

The lead circuit is a fixed design. The lead circuit was designed to improve the Phase Margin of the open loop (64.10). The designed lead circuit's transfer functions is

$$K_{lead}(s) = \frac{(1 + s \cdot 0.8418)}{1 + 0.4209s} \quad (64.14)$$

64.6 Simulation Program Designin

64.6.1 Pseudo Code

The pseudo code of the program is easily demonstrated by referring to Table 64.1.

The pseudo code for comparison of IOPI and FOPI is as follows:

- (1) Creates M IOPI and FOPI controllers with the fields in Table 64.1.
- (2) An M length decision vector is randomly generated from the decision for each controller's parameter, with the field of α for the IOPI controller left empty.
- (3) For every entry of the IOPI and FOPI controller's parameters.

Table 64.1 Fields of controllers that have to be evaluated

| K_p | I | α | ISE | ISI |
|-------|-----|----------|-----|-----|
|-------|-----|----------|-----|-----|

- The closed loop is simulated and the ISI and ISE are evaluated and filled in the appropriate field.
- (4) The ISE and ISI are plotted in a two-dimensional scatter plot to view the objective space.
- (5) The Pareto Optimal Front for the two -dimensional scatter is computed and plotted separately.

This program was easily extended to include the lead compensator, by adding the controller in the simulator.

64.6.2 Simulation of Closed Loop Response

A fourth-order Runge–Kutta algorithm was used for simulation of the closed loop response to unit step responses. The total simulation time was set to 3 s. This is because the parameters of the controllers are constrained to give a closed loop response to step inputs that has a maximum settling time of 2.44 s.

64.6.3 Cost Function Evaluation

The cost functions are evaluated over the entire simulation time. The FOPI has small steady state errors [3] which will cause the value of the ISE to accumulate, even though the system has settled. This was regarded as reasonable because the FOPI is properly penalized when it does not track the setpoint.

64.6.4 Oustaloup’s Approximation Method for FOPI

Oustaloup’s approximation method of approximation was used to approximate the FOPI controller. The selected values for Oustaloup’s approximation method parameters were $N = 5$, $w_b = 0.001$ rad/s and $w_h = 500$ rad/s.

Code plots for the ideal FOPI and approximated FOPI were plotted, using Visual Studio C# 2010, to observe how well the two match. The Code Plots are shown in Fig. 64.3.

Reference [6], shows that Bode (1945) showed that the phase-frequency and magnitude-frequency of Bode plots has the following relationship

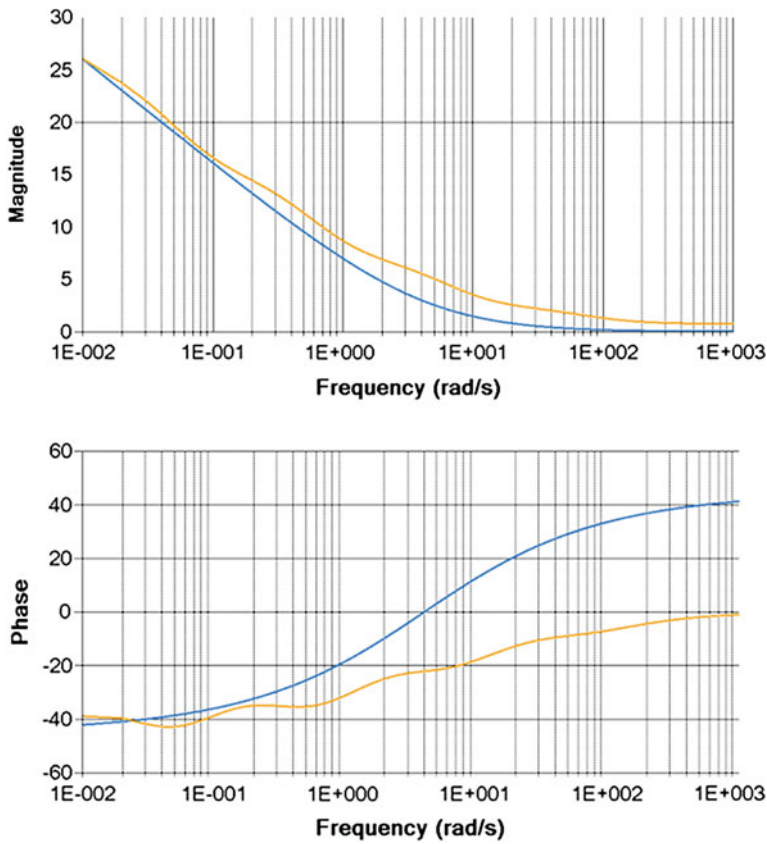


Fig. 64.3 Bode plot of FOPI where *blue* (exact FOPI) and *orange* (approximated FOPI)

$$\varphi(\omega) \approx \frac{\pi}{2} \frac{dG}{d\lambda} \quad (64.15)$$

where φ is the phase and G (magnitude) and λ (logarithmic frequency) are defined as

$$\begin{aligned} G &= \log(G(j\omega)) \\ \lambda &= \log(\omega) \end{aligned} \quad (64.16)$$

where ω is the frequency (rad/s).

This shows that for frequencies $\omega > 100$ rad/s, in Fig. 64.3, the phase of 40° for the FOPI cannot be attained with the approximation because the slope of the magnitude-frequency is zero. Figure 64.3 also shows that Oustaloup's filter cannot approximate positive values of the phase magnitude because the slope of the magnitude-frequency is negative.

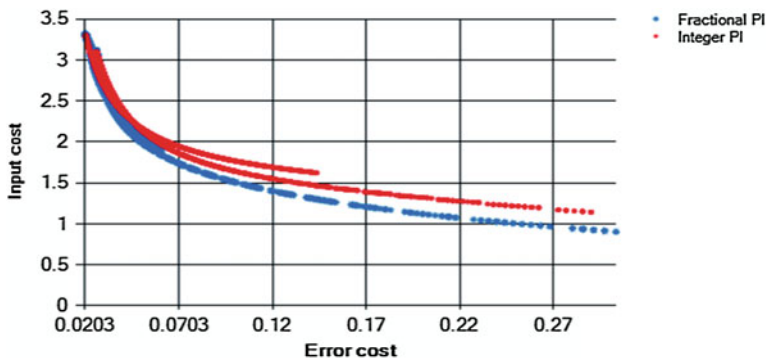


Fig. 64.4 Objective Space for FOPI (blue) and IOPI (red) controllers

64.7 Simulations Results

The lengths of the decision vectors for I , K_p and α were set to 5,000 for these simulation results.

64.7.1 Comparison of IOPI and FOPI

Figure 64.4 shows the objective space of the IOPI and FOPI controllers. In region 1 of Fig. 64.5, the FOPI's POF is closer to the origin than the POF of the IOPI, which means that the FOPI controller is better than the IOPI controller. In region 2, the POF of the IOPI is closer to the origin than the POF of the FOPI controller, which means the IOPI controller is better FOPI controller. A zoomed in version of Region 2 is shown in Fig. 64.6.

64.7.2 Comparison Between FOPI and IOPI with Lead Circuit

Figure 64.7 shows the objective space of the FOPI controller and IOPI with lead circuit. Region 1, from Fig. 64.8, indicates that FOPI is better than lead compensated IOPI. However, there is a decrease in distance between the POFs of the controllers as compared to Fig. 64.5. This result shows that there has been an improvement of performance for the IOPI controller. In Region 2, from Fig. 64.8, the lead compensated IOPI has better performance than the FOPI because the POF of the lead compensated IOPI is closer to the origin than the POF of the FOPI. These results show that a lead compensator can improve the performance of the IOPI controller. The parameters of the FOPI were constrained and this might have

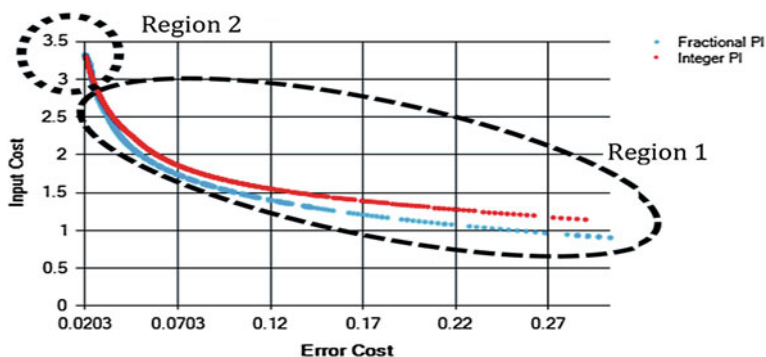


Fig. 64.5 POFs for FOPI (blue) and IOPI (red) controllers

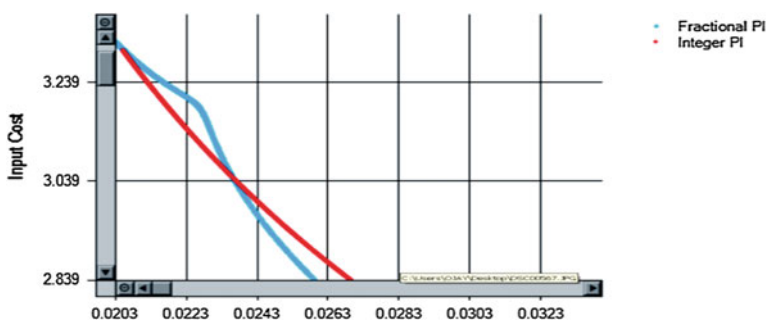


Fig. 64.6 Region 2 zoomed in

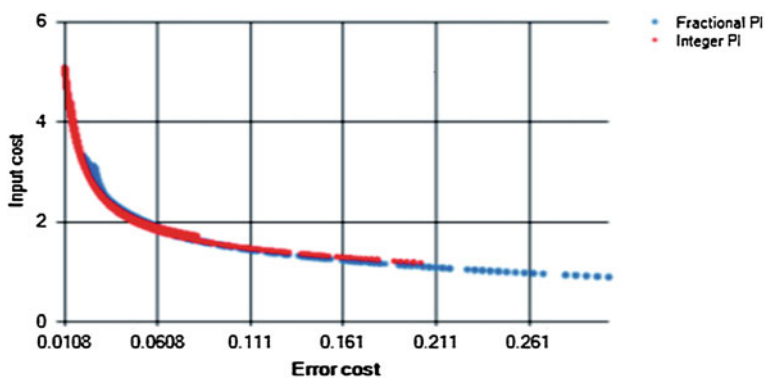


Fig. 64.7 Objective Space for FOPI controller (blue) and IOPI controller with lead circuit (red)

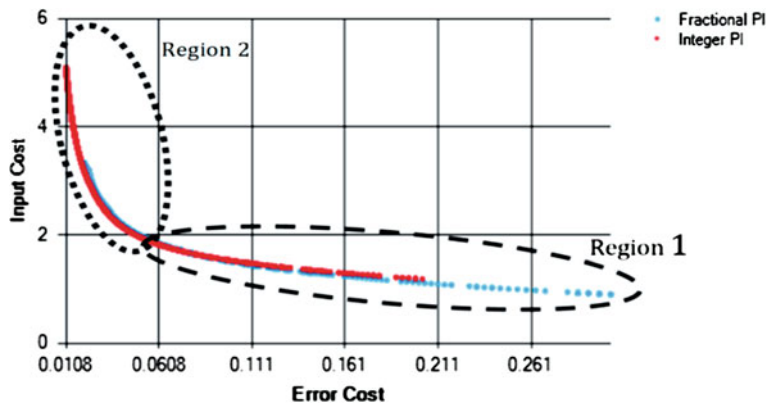


Fig. 64.8 POFs for FOPI controller (blue) and IOPI controller with lead circuit (red)

Table 64.2 Parameters of FOPI and IOPI controllers and the corresponding ISE and ISI values

| Gain (K) | Zero(I) | Order (α) | ISE | ISI |
|----------|---------|--------------------|-------|-------|
| 1.65 | 3 | – | 0.056 | 2.617 |
| 1.65 | 3 | 0.2521 | 0.066 | 2.510 |

restricted other performance features, but the aim was to see how the performance of the IOPI changes when a lead compensator is added to the open loop.

The results show that the performance features that the FOPI has can be closely matched with an IOPI controller with a lead circuit.

64.8 Implementation of Controllers

The parameters and cost function values of the IOPI and FOPI controllers, that were implemented are shown in Table 64.2. The results of the FOPI controller is shown in Fig. 64.9, the IOPI controller is shown in Fig. 64.10 and IOPI with lead circuit is shown in Fig. 64.11.

The implementation of the controllers deviated from the theoretical closed loop response because the motor is non-linear at start-up.

The FOPI controller produces a steady state error which was expected because the controller was not modified as in Ref. [3]. The other two controllers produced closed loop responses that tracked the setpoint. The FOPI settled after 10 s which is because the FOPI is controller is fractionally integrating [3].

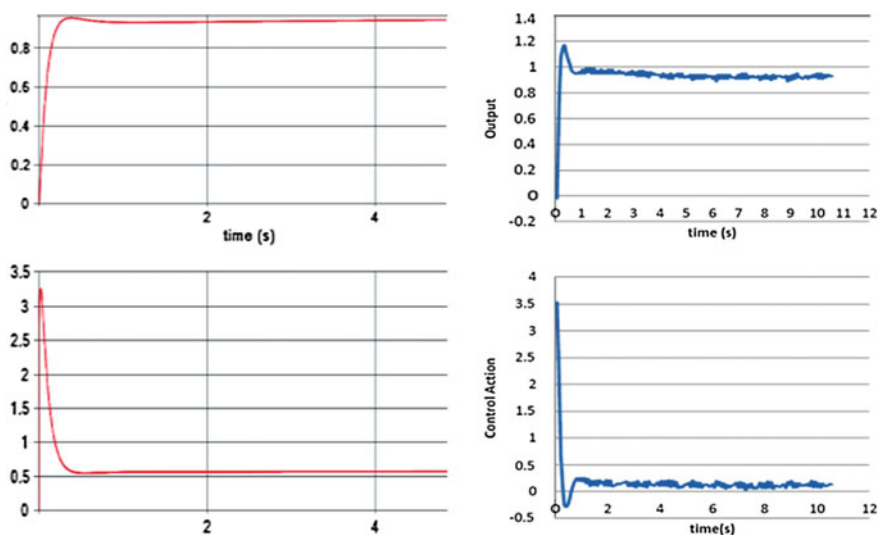


Fig. 64.9 Simulation results (*left column*) and implementation results (*right column*) for FOPI controller

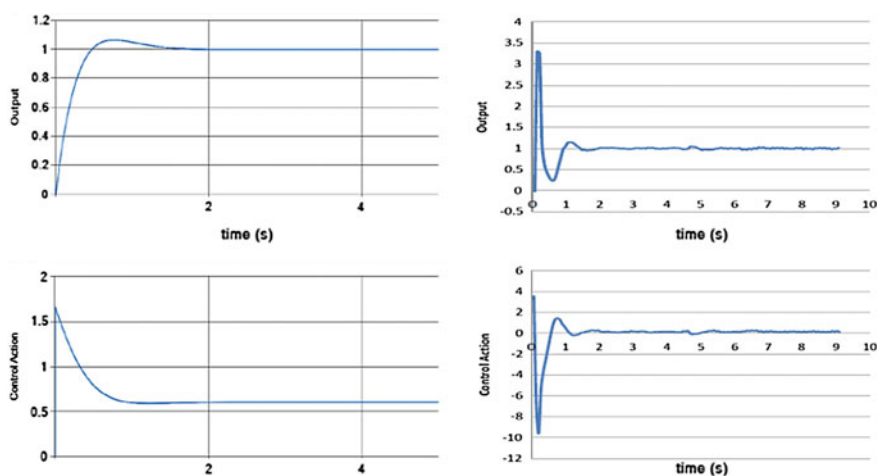


Fig. 64.10 Simulation results (*left column*) and implementation results (*right column*) for IOPI controller

64.9 Conclusions

A comparison of the FOPI and IOPI controller was conducted using Pareto Optimal Fronts. It was shown that FOPI controller is not always better than IOPI controller, under the considered cost functions. The FOPI controller was further

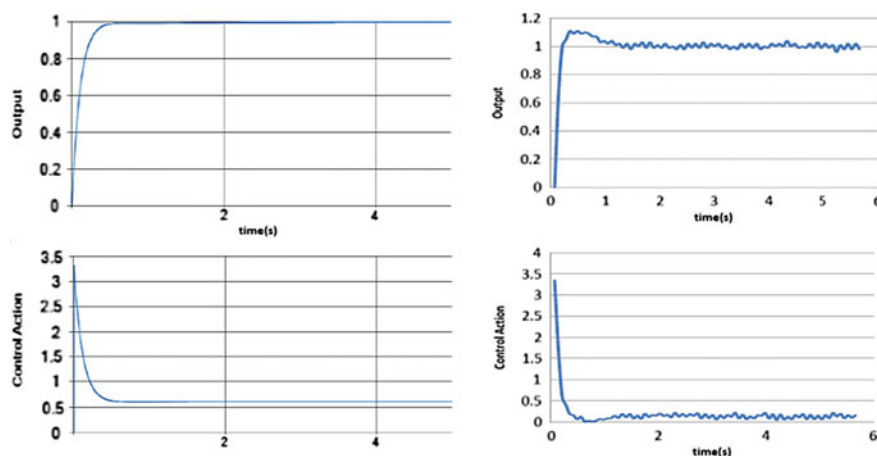


Fig. 64.11 Simulation results (*left column*) and implementation results (*right column*) for IOPI controller with lead circuit

compared to the IOPI controller with lead circuit added. The IOPI controller with the lead circuit is better than the FOPI controller in most regions of the POFs than the IOPI controller alone. Thus the lead circuit can provide additional parameters that can be tuned to meet specifications, without having to consider fractional-orders.

The approximation of the phase-frequency for the ideal FOPI controller with an integer-order transfer function cannot be achieved for positive values of the phase. The implemented FOPI controller produced a steady state error in the response whilst the IOPI and IOPI controller with lead circuit did not.

References

1. Maciejowski JM (1989) Multivariable feedback design. Addison-Wesley, Wokingham
2. Podlubny I (1999) Fractional-order systems and PID-controllers. *Autom Control* 1(44):208–214
3. Jesus S, Barbosa S, Tenreiro JA (2010) Effect on fractional orders in the velocity control of servo system. *Comput Math* 59:1679–1686
4. Zhao C, Chen YQ, Xue D (2006) Fractional PID of DC-motor with elastic shift: a case study. In: American control conference, Minneapolis, 2006
5. Onat C, Tan N, Yeroglu C (2009) A new tuning method for fractional PID controller, in electrical and electronics engineering international conference, Bursa, 2009, pp 312–316
6. Oustaloup A et al Frequency-band complex noninteger differentiator: characterization and synthesis
7. Tehrani KA et al (2010) Design of fractional order PID controller for boost converter based on multi-objective optimisation, in power electronics and motion control conference, Ohrid, 2010. pp 179–185
8. Maciejowski JM (1989) Multivariable feedback design. Addison-Wesley, Wokingham

Chapter 65

A Pattern-Based Approach for Representing Condition-Action Clinical Rules into DSSs

A. Minutolo, M. Esposito and G. De Pietro

Abstract The integration of clinical recommendations into clinical Decision Support Systems (DSSs), aims at increasing the consistent, effective, and efficient of the daily medical practice. The actual application of those systems to medical settings depends on the severity of the upgradability and maintainability they require. With this respect, this paper proposes a pattern-based approach to guide and assist physicians in the process of editing and formalizing clinical recommendations, formalized as if-then rules.

65.1 Introduction

Clinical recommendations have shown to be able to improve the efficiency of medical practices, and their outcomes, when followed [1]. This is true, mostly, when clinical practice guidelines are provided as clinical decision support [2].

Recently, several studies [3, 4] have proposed the DSSs as the most profitable way for integrating clinical practice guidelines and improving patient safety or

A. Minutolo (✉) · M. Esposito · G. De Pietro
Institute for High Performance Computing and Networking, ICAR-CNR,
Via P. Castellino 111, 80131 Naples, Italy
e-mail: aniello.minutolo@na.icar.cnr.it

M. Esposito
e-mail: massimo.esposito@na.icar.cnr.it

G. De Pietro
e-mail: giuseppe.depietro@na.icar.cnr.it

A. Minutolo
University of Naples “Parthenope”, Centro Direzionale, Isola C4, 80143 Naples, Italy

quality of care. In particular, knowledge-based DSSs are able to assist physicians in their decision making activities, by providing clear explanations of why actions and suggestions are concluded. This feature is essential and allows doctors to confide and understand the outcomes of the clinical DSS [5].

In order to formalize clinical recommendations into a Knowledge Base (KB), the physicians' knowledge needs to be refined into a machine-readable formalism which will be processed by the system for simulating experts' behaviors or for providing suggestions and explanations.

Several knowledge representation methods are used at present, even if the symbolic approach which combines ontologies and rules has recently appeared as the more appropriate to describe medical knowledge, since easily understandable by a non-technical audience, e.g. clinicians.

Specifically, if-then rules provide an efficient and intelligible way to model experts' behaviors and clinical recommendations in order to support physicians in their decision making activities. Furthermore, if-then rules are considered essential to provide clear explanations to the user of why actions and suggestions are provided when determined input data have been gathered.

Doctors are not usually supported to directly manage the clinical guidelines formalized into a KB, but they can only take advantage of the outcomes provided by the DSS. Since the available knowledge acquisition tools are commonly designed for a knowledge engineer audience, they are typically too complex for clinicians, and only an intervention made by technicians can alter the existing KB.

By providing more intuitive mechanisms to manage and update a KB, clinicians are motivated to accept and use clinical DSSs in daily medical settings, since they are mostly entrusted with the suggestions generated starting from their expertise, especially if inserted by them.

With the goal of providing user-friendly solutions for managing and editing the KB of a clinical DSS, this paper proposes a pattern-based approach for modeling condition-action clinical rules expressed in the form of if-then rules built on the top of ontological vocabularies, and presents an intuitive rule editing tool which implements such patterns.

The approach has been devised: (i) to offer editing patterns closer to non-technical users, e.g. clinicians, in order to support the process of clinical knowledge representation in a simple fashion; (ii) to reduce the complexity of the formalization process at the cost of functionality, by enabling the creation of only the specific types of ontologies and rules that are needed and functional in the context of clinical DSSs.

The rest of the paper is organized as follows. [Section 65.2](#) introduces an overview of the existing general-purpose solutions for building KBs and addresses the motivations underlying the approach proposed. [Section 65.3](#) depicts the pattern-based approach, while, in [Sects. 65.4](#) and [65.5](#), design considerations and the rule editing tool implemented are described, respectively. [Section 65.6](#) concludes the work.

65.2 Related Work and Motivations

Existing tools for editing and managing a KB are usually able to provide many general-purpose capabilities by adopting knowledge representation formalisms, such as ontologies and rules [6]. Due to their powerful abilities, they often result very complex and high oriented to the supported knowledge representation formalisms which typically make them unusable for non-expert users [7].

Several studies [8–10] have shown that domain experts are usually not used to model their knowledge in terms of ontologies and rules. Concepts such as the definition of classes and sub-classes of entities are often very intuitive, but more technical concepts, such as logic axioms and rules, can be managed only through a knowledge engineer.

We speculated that to reduce the gap between domain experts and knowledge editing tools, user interfaces should be designed with the aim of speaking the language of the specific domain for which a knowledge base, composed in terms of ontologies and rules, has to be developed.

To overcome these issues, we have focused on KB editing functionalities mostly concerned to the managing of condition-action clinical guidelines, since, to the best of our knowledge, none of the existing tools has been developed with the aim of being mainly oriented to that direction. As a result, we have focused on developing intuitive KB's editing functionalities with the goal of simplifying the provided editing features so to reduce the required technical abilities for managing both the ontological terms and the procedures involved in clinical DSSs.

Specifically, we have analyzed some typical condition-action clinical rules, in order to identify common, and repetitive, elements in their structures with the goal of bringing back the creation and editing process of complex clinical rules, to simpler and handier objects.

65.3 The Pattern-Based Approach

Typically, clinical recommendations contain statements referred to attribute values owned by one or more entities, such as a patient, a clinical exam, and so on. Other kinds of statements, instead, can refer to the existence of a relation between entities. In the following, an example of the typical if-then structure, proper of condition-action clinical rules, is showed:

if the age of a patient is more than 18 and the measured heart rate of the patient is more than 100 bpm (beats per minute), then the suspected physiological status of the patient is tachycardia;

In a clinical knowledge-based DSS, an if-then rule is usually modeled by a set of statements to be verified (called the antecedent part of the rule) related to the acquired data, and a set of actions (called the consequent part) to be performed when the antecedent part of the rule is true. For instance, the above rule contains

two statements to verify and one action (in this case, an assertion about the patient status) derived when both the two rule statements are true.

Analyzing the above clinical recommendations, it is possible to reveal that, while some statements are referred to generic classes of entities, some other statements regard, instead, specific instances of such entities. For instance, the statement “*the age of a patient is more than 18*” aims at testing the existence of an entity, a patient, having a determined attribute value. There could be one, more than one or no patient satisfying that statement, it depends on the current known world. It has to be highlighted that the evaluation of the above statement is completely free from any other eventual statements in the same rule. The statement “*the measured heart rate of the patient is more than 100 bpm*”, instead, contains the meaning that the patient to consider has been already identified elsewhere. Then, once discovered a patient satisfying the first statement, the measured heart rate of that specified patient has to be evaluated. Specifically, the second statement contains a connection to another statement in the rule, and it will be not achievable until the first one is evaluated. Similar considerations could be drawn for the rule actions, which are usually referred to entities already determined in the antecedent part of the rule.

After introducing the reported preliminary notions, a more formal description is given in the following. We defined **self-contained** term, a rule term (statement or action) that can be treated by only analyzing it and the current known world, freely from any other eventual terms in the same rule. Contrariwise, we defined **external-connected** term, a rule term referred to a particular instance of entity. Then, to be evaluated, it requires information coming from other statements in the rule.

In this respect, the example rule reported above is composed by a self-contained statement which let to determine from available patients the specific one having age more than 18, and an external-connected statement which is referred to an adult patient, determined from the previous statement. Finally, also the rule action is an external-connected term, and it is referred to the same patient considered in the rule statements.

According to the above definitions, the terms of a complex clinical rule can be subdivided into self-contained and external-connected terms. Moreover, analyzing the clinical rules terms, it is possible to reveal further common elements, that can be distinguished by the type of referred entities and, the particular connections established between the terms of the same rule. Specifically, since rule actions could be usually described as special rule statements having determined restrictions, we have focused on the analysis of the common and repetitive elements in generic clinical rule statements. Starting from these considerations, we have identified a set of statement patterns which let to describe the generic structure of a clinical recommendation.

We have defined as **relational statement pattern** a rule condition that aims at testing the existence of relations between two entities. It is defined by a relation to satisfy, and two entities which can be associated through the relation. For example, the statement “*a patient is suffering from a disease*” will verify the existence of the relation “*is suffering from*” between “*a patient*” and “*a disease*”.

Table 65.1 Relation statement pattern

| Pattern | Example | Symbol |
|--------------------|---|-----------------------|
| self-contained | “a patient is suffering from a disease” | R |
| external-connected | “a patient has completed a clinical exam” and “the patient is suffering from a disease” | $\swarrow R$ |
| external-connected | “a disease is hereditary” and “a patient is suffering from the disease” | $R \swarrow$ |
| external-connected | “a patient has completed a clinical exam” and “a disease is hereditary” and “the patient is suffering from the disease” | $\swarrow R \swarrow$ |

In the above example, the entities involved are generic and only their typology is indicated (patient and disease), then the considered statement is self-contained and it can be evaluated freely from other statements results.

A relational statement can also be defined external-connected by specifying either one or both entities of the statement, in order to verify only if the relation involves the specified entities. The statements “*the patient is suffering from a disease*”, “*a patient is suffering from the disease*”, and “*the patient is suffering from the disease*” are example of relational statements in which some entities have been already identified elsewhere.

In Table 65.1, a brief summary of the relational statements identified is reported.

Moreover, some other kinds of statements, which can be found in typical clinical rules, are referred to attribute values of entities, in order to compare them either with constant threshold values or between themselves.

A **Single Attribute Statement Pattern** involves an entity, an attribute of the entity, a constant value comparable with the attribute value, and a logical operator to be evaluated between the attribute value and the constant value. For example, the statement “*the age of a patient is more than 18*” is a self-contained statement that will evaluate the existence of “*a patient*” having the attribute “*age*” whose value is “*more than*” the constant value “18”.

The external-connected version of this pattern let to model statements referred to a specific entity owner of the attribute. For example, the statement “*the age of the patient is more than 18*” would mean that a specific patient has already been determined elsewhere.

Moreover, because this pattern needs only an entity to be defined, differently from the relation statement pattern, only one kind of external-connected version can be defined. The defined single attribute patterns are showed in Table 65.2.

In order to model attribute statements which aim at comparing attribute values between themselves, we defined the **Double Attribute Statement Pattern**: namely, it involves two entities, two comparable attributes of them, and a logical operator to be evaluated between the attribute values. For example, the statement “*the stature of a patient is less than the length of a bed*” will verify the existence

65.4 Design Considerations

In this section, we provide some design considerations about the development of an editing interface implementing the defined patterns, and able to provide user-friendly facilities for encoding clinical recommendations in the knowledge base of a clinical DSS.

In order to encode clinical guidelines into a clinical knowledge-based DSS, the physicians' knowledge needs to be refined into a machine-readable formalism which will be processed by the system for simulating experts' behaviors or for providing suggestions and explanations.

As shown in the previous section, the compositional elements of clinical recommendations are usually entities related to a specific clinical domain, with their inter-relations and attributes. In this respect, we chose to adopt ontologies and semantic web technologies to formalize a knowledge base, since they provide standard and high expressive formalisms to model and share the knowledge about a specific domain.

Specifically, the entities operated by the rules, such as a *patient*, a *clinical parameter*, a *disease*, are concepts of interest in a specific clinical domain. The knowledge about the raw and abstract concepts, their instances, attributes, and inter-relations, namely the *declarative knowledge* of the KB, can be formalized in form of ontology by using the standard Web Ontology Language (OWL) [11].

Moreover, to compose the knowledge about the procedures of the decision making activity, namely the *procedural knowledge*, we have adopted the Jena rule language [12], since it is the most appropriate language for writing if-then rules operating on OWL ontologies.

In Jena, a rule is composed by a conjunction of statements to be satisfied, and a set of actions to be performed when the antecedent part is true. The rule language enables to refer in the rule bodies not only to classes and properties defined by a given ontology but also to variables explicitly defined in the rules.

With the goal of developing an editing interface supporting the above defined patterns, each of them has been redefined in terms of ontological concepts, attributes, and relations in order to determine information required to compose each particular rule term.

Specifically, given the ontology \mathbf{O} , the set \mathbf{C} of the concepts defined in \mathbf{O} , the set \mathbf{R} of the relations defined in \mathbf{O} and indicated with $R_{\text{domain}} (R_{\text{range}})$ the domain (range) restriction of a relation R , and with I_C the set of instances of a concept C , the relational patterns can be built through the tuples defined in the Table 65.4.

In the same way, given the ontology \mathbf{O} , the set \mathbf{C} of the concepts defined in \mathbf{O} , the set \mathbf{A} of the attributes defined in \mathbf{O} , a constant value v having range v_{range} , and indicated with $A_{\text{domain}} (A_{\text{range}})$ the domain (range) restriction of an attribute A , and $\mathbf{L}(\text{range}_1, \text{range}_2)$ the set of logical operators suitable between the range of two attributes, and with I_C the set of instances of a concept C , the attribute patterns can be built by the tuples showed in Table 65.5.

Table 65.4 Ontological descriptions of relational patterns

| Definition | Pattern |
|--|-----------------------|
| $RP(C_1, R, C_2) / C_1 \subseteq R_{\text{domain}}, C_2 \subseteq R_{\text{range}},$ where $C_1 \in \mathbf{C}, C_2 \in \mathbf{C}, R \in \mathbf{R}$ | R |
| $RP(C_1, i_1, R, C_2) / C_1 \subseteq R_{\text{domain}}, C_2 \subseteq R_{\text{range}},$ where $i_1 \in I_{C_1}, C_1 \in \mathbf{C}, C_2 \in \mathbf{C}, R \in \mathbf{R}$ | $\swarrow R$ |
| $RP(C_1, R, i_2, C_2) / C_1 \subseteq R_{\text{domain}}, C_2 \subseteq R_{\text{range}},$ where $i_2 \in I_{C_2}, C_1 \in \mathbf{C}, C_2 \in \mathbf{C}, R \in \mathbf{R}$ | $R \swarrow$ |
| $RP(C_1, i_1, R, i_2, C_2) / C_1 \subseteq R_{\text{domain}}, C_2 \subseteq R_{\text{range}},$ where $i_1 \in I_{C_1}, i_2 \in I_{C_2}, C_1 \in \mathbf{C}, C_2 \in \mathbf{C}, R \in \mathbf{R}$ | $\swarrow R \swarrow$ |

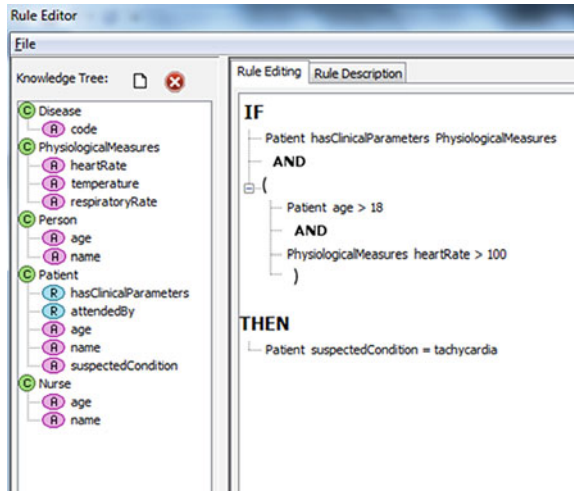
Table 65.5 Ontological descriptions of attribute patterns

| Definition | Pattern |
|---|-------------------------|
| $AP(C, A, op, v) / C \subseteq A_{\text{domain}}, v_{\text{range}} \subseteq A_{\text{range}},$ where $C \in \mathbf{C}, A \in \mathbf{A}, op \in \mathbf{L} (A_{\text{range}}, v_{\text{range}})$ | A_1 |
| $AP(C, i, A, op, v) / C \subseteq A_{\text{domain}}, v_{\text{range}} \subseteq A_{\text{range}},$ where $i \in I_C, C \in \mathbf{C}, A \in \mathbf{A}, op \in \mathbf{L} (A_{\text{range}}, v_{\text{range}})$ | $\swarrow A_1$ |
| $AP(C_1, A_1, op, A_2, C_2) / C_1 \subseteq A_{1\text{domain}}, C_2 \subseteq A_{2\text{domain}},$ where $C_1, C_2 \in \mathbf{C}, A \in \mathbf{A}, op \in \mathbf{L} (A_{1\text{range}}, A_{2\text{range}})$ | A_2 |
| $AP(C_1, i_1, A_1, op, A_2, C_2) / C_1 \subseteq A_{1\text{domain}}, C_2 \subseteq A_{2\text{domain}},$ where $i_1 \in I_{C_1}, C_1, C_2 \in \mathbf{C}, A \in \mathbf{A}, op \in \mathbf{L} (A_{1\text{range}}, A_{2\text{range}})$ | $\swarrow A_2$ |
| $AP(C_1, A_1, op, A_2, i_2, C_2) / C_1 \subseteq A_{1\text{domain}}, C_2 \subseteq A_{2\text{domain}},$ where $C_1, C_2 \in \mathbf{C}, i_2 \in I_{C_2}, A \in \mathbf{A}, op \in \mathbf{L} (A_{1\text{range}}, A_{2\text{range}})$ | $A_2 \swarrow$ |
| $AP(C_1, i_1, A_1, op, A_2, i_2, C_2) / C_1 \subseteq A_{1\text{domain}}, C_2 \subseteq A_{2\text{domain}},$ where $C_1, C_2 \in \mathbf{C}, i_1 \in I_{C_1}, i_2 \in I_{C_2}, A \in \mathbf{A}, op \in \mathbf{L} (A_{1\text{range}}, A_{2\text{range}})$ | $\swarrow A_2 \swarrow$ |
| $AP(C, i, A_1, op, A_2) / C \subseteq A_{1\text{domain}}, C \subseteq A_{2\text{domain}},$ where $i \in I_C, C \in \mathbf{C}, A_1 \in \mathbf{A}, A_2 \in \mathbf{A}, op \in \mathbf{L} (A_{1\text{range}}, A_{2\text{range}})$ | \mathcal{A}_2 |

Each pattern requires a specific set of information to be defined. For instance, to compose a relational pattern the user should select the three elements involved (the relation R , and the concepts C_1 and C_2).

Moreover, by taking advantage of the underlying ontology, many of the required information can be automatically computed letting users just to select the entities to insert in a rule, and determining the statement to create just analyzing the particular selected entities.

Fig. 65.1 The rule editing interface



65.5 The Rule Editing Tool

Starting from these design considerations, we have implemented a knowledge editing tool embedding the above defined patterns with the goal of providing a simple and intuitive interface to the clinicians who do not have a deep technical expertise about ontologies, OWL syntax, and Jena if-then rules.

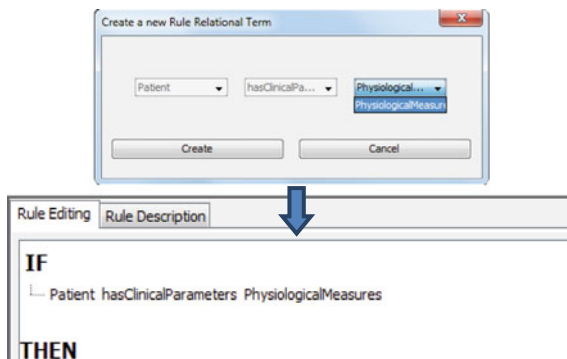
The overall editing interface is shown in Fig. 65.1, and it is organized as a Knowledge Tree (the left area) and a Rule Editing Interface (the right area). First of all, the Knowledge Tree contains concepts, attributes, and relations to be used in the rules. It has been designed with two main goals: (i) to show all the concepts defined in the underlying ontology, highlighting, for each concept, its attributes and inter-relations; (ii) to lead the user in the process of composing well-formed rule statements by hiding formalisms and constraints. Specifically, a rule statement can involve instances, classes, relations and attributes, but it will be congruent and verifiable only if domain and range restriction are observed.

Users can compose rule terms by selecting either a relation or an attribute in the Knowledge Tree. Each available relation (attribute) is visualized under a concept only if the concept itself is contained in the domain restriction of the considered relation (attribute). In this way, the same relation (attribute) can be visualized under several concepts in the Knowledge Tree.

On the other hand, the Rule Editing Interface provides two graphical areas, corresponding to the antecedent and consequent parts of a rule, where it is possible to drop items from the Knowledge Tree in order to compose a new rule term. As a result, the Rule Editing Interface is graphically arranged in two areas.

The "IF" icon localizes the root node of the antecedents tree where each child node can be respectively a statement, a logical conjunctive/disjunctive connector or a couple of circle brackets. Statements included into a couple of circle brackets,

Fig. 65.2 The interface for creating a self-contained relational pattern



which are associated to a higher evaluation priority, are arranged and visualized as nodes placed in a nested level with respect to the brackets including them.

Similarly, the “*THEN*” icon is the root node of the consequents tree where each child node can be respectively a statement or a logical conjunctive connector. Since the consequent part of a rule consists in a conjunction of statements due to the choice of the Jena rule formalism, as previously described, neither opened/closed circle brackets nor logical disjunctive connectors are admitted.

Dragging a relation (attribute) from the Knowledge Tree, and dropping it in the Rule Tree interface, the user will start the process of composing a rule term involving both the concept above the selected relation (attribute) and the relation (attribute) itself.

For example, to apply a self-contained relational pattern, $RP(C_1, R, C_2)$, the relation R and the concepts C_1 and C_2 have to be specified.

Because a relation R in the Knowledge Tree brings the information about both the relation itself and a specific concept C_1 contained in the domain restriction of R , the selected relation R can be used to compose a relational pattern, and the user has only to select the specific concept C_2 by choosing between the concepts included in the range restriction of R .

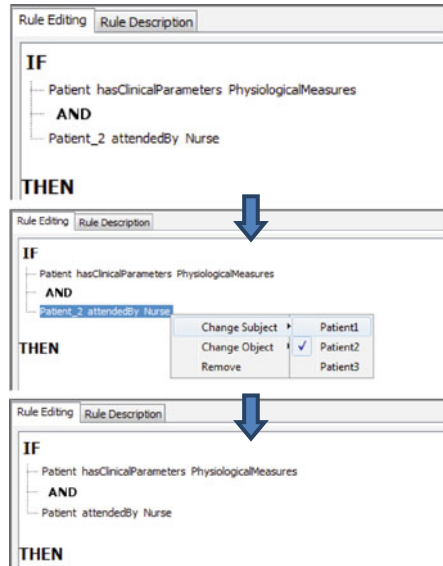
Then, for example, to create a new rule statement $RP(Patient, has\ Clinical\ Parameters, Physiological\ Measures)$ user has to drag the relation *has Clinical Parameters* visualized under the concept *Patient* in the Knowledge Tree, and to drop it into the “*IF*” area of the Rule Editing Interface. This latter will ask the user to select the concept C_2 , as showed in Fig. 65.2, and the statement will be created.

Moreover, to apply an external-connected relational pattern, two-steps are required, namely the creation of a self-contained version and the specification of one or more instances to refer to.

It is important to note that when two statements in the rule are referred to the same concept, by default, the Rule Editing Interface labels them with two different ids in order to indicate that the statements are referred to two different generic instances.

Then, applying, for instance, the pattern $RP(C_1, i_1, R, C_2)$ to express the statement “the *Patient attended By a Nurse*” linked to the patient inserted in the

Fig. 65.3 Using an external-connected relational pattern



previous example, the self-contained pattern $RP(Patient, attended\ By, Nurse)$ is first created and, then, by right-clicking on the statement, the *Patient* mentioned in it is set as the same mentioned in the other statement previously created (see Fig. 65.3).

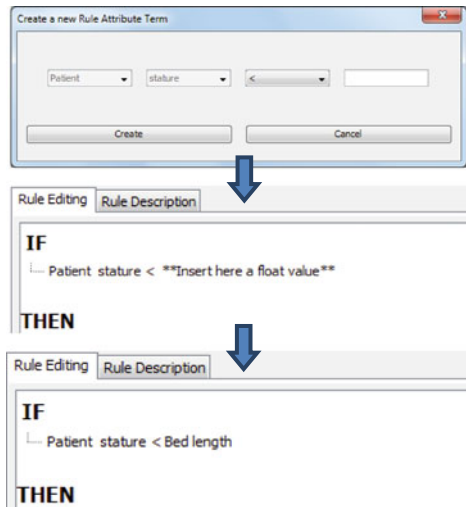
Both the self-contained single attribute pattern $AP(C, A, op, v)$ and its external-connected version $AP(C, i, A, op, v)$, work in a very similar fashion with respect to the self-contained and external-connected relational patterns. In particular, to apply $AP(C, A, op, v)$, only a specific logical operator op and a constant input value v , to which apply op , have to be specified according to the attribute A and its range restriction. Moreover, to use $AP(C, i, A, op, v)$, a previously defined instance i of the concept C has to be specified by right-clicking on the statement inserted.

Furthermore, to use the self-contained double attribute pattern $AP(C_1, A_1, op, A_2, C_2)$, two steps are required, namely the creation of a single attribute pattern and the specification of the second attribute involved.

For instance, to create the statement $AP(Patient, stature, <, length, Bed)$ first a statement $AP(Patient, stature, <, \text{" "})$ is defined according to the self-contained single attribute pattern. Differently from this latter pattern, the self-contained double attribute pattern requires to not specify any input value for the created statement (see Fig. 65.4). Indeed, the attribute *length* of the concept *Bed* is dragged and dropped, in place of the input value not inserted, into the statement previously created (in particular where the string “Insert here a float value” is reported).

Finally, to use the external-connected double attribute pattern $AP(C_1, i_1, A_1, op, A_2, i_2, C_2)$, previously defined instances i_1 and i_2 of the concept C_1 and C_2 have to be specified by right-clicking on the statement inserted.

Fig. 65.4 Using a self-contained double attribute pattern



In particular, to realize a self-connected double attribute pattern, the user has simply to specify that both attributes involved into the pattern belong to the same instance.

65.6 Conclusions

The paper described a pattern-based approach to guide and assist the creation and formalization of condition-action clinical recommendations. The approach has been implemented into a Rule Editing Tool which offers graphical facilities for easily inserting and editing clinical recommendations expressed in the form of if-then rules built on the top of ontological vocabularies.

The approach was devised: (i) to offer editing patterns closer to non-technical users, e.g. clinicians, in order to support the process of clinical knowledge representation in a simple fashion; (ii) to reduce the complexity of the formalization process at the cost of functionality, by enabling the creation of only the specific types of ontologies and rules that are needed and functional in the context of clinical DSSs.

Future work will regard the definition of new patterns to handle vagueness in condition-action clinical guidelines by using the Fuzzy Logic Theory [13].

In particular, next steps in this direction will regard the definition of patterns for supporting the building of fuzzy rules on the top of ontological concepts and properties.

References

1. Woolf S (1992) Practice guidelines, a new reality in medicine: II. methods of developing guidelines. *Arch Intern Med* 152(5):946–952
2. Scott I (2009) What are the most effective strategies for improving quality and safety of health care? *Intern Med J* 39:6
3. Niemi K, Geary S, Quinn B, Larrabee M, Brown K (2009) Implementation and evaluation of electronic clinical decision support for compliance with pneumonia and heart failure quality indicators. *Am J Health Syst Pharm* 66:4
4. Brokel J, Shaw M, Nicholson C (2006) Expert clinical rules automate steps in delivering evidence-based care in the electronic health record. *Comput Inform Nurs* 24(4):196–205
5. Shortliffe E, Cimino J (2006) *Biomedical informatics: computer applications in health care and biomedicine*. Springer, New York
6. Protégé <http://protege.stanford.edu/>
7. García-Barriocanal E, Sicilia MA, Sánchez-Alonso S (2005) Usability evaluation of ontology editors. *Knowl Organ* 32:1
8. Escórcio L, Cardoso J (2007) Editing tools for ontology construction. *Semantic web services: theory, tools and applications*. Idea Group, USA
9. Denny M (2004) Ontology tools survey <http://www.xml.com/pub/a/2004/07/14/onto.html>
10. Gómez-Pérez A, Ortiz-Rodríguez F, Villazón-Terrazas B (2004) *Ontological engineering*. Springer, Berlin
11. OWL2 (Web Ontology Language) <http://www.w3.org/TR/owl2-overview/>
12. Carroll JJ, Dickinson I, Dollin C, Reynolds D, Seaborne A, Wilkinson K (2004) Jena: implementing the semantic web recommendations. In: *Proceedings of the 13th international World Wide Web conference on Alternate track papers and posters*. ACM, USA
13. Zadeh L (1965) Fuzzy sets. *Inf Control* 8:3

Chapter 66

Authorization of Proxy Digital Signature in Workflow Systems

Samir Fazlagic and Narcis Behlilovic

Abstract In public key infrastructure systems the digital signature provides a mechanism to validate the identity of a signer and verify the authenticity and integrity of a signed document. However, the standard procedure for verifying digital signature cannot prove that the signer is authorized to sign the document. With proxy signature it is particularly important to prove whether the delegator is the authorized signer. In an organization environment with more document signers, such as in workflow processes in document management systems and electronic archive systems, a control mechanism should be provided throughout the document lifecycle to check whether the signer was actually authorized to sign the document. This paper proposes a mechanism for validating the authorization of the proxy signer during the whole document life cycle.

66.1 Introduction

In document management systems that use workflow processes for user collaboration, an electronic document will usually be signed by multiple signers in sequential tasks. Before they sign a document, some of the signers may need to verify the previous signatures. In certain cases it is necessary to check whether the previous signer was actually the authorized signer. The order of signing is of utmost importance, because it illustrates the decision-making procedures and as such is subject to validation of signature. The mechanism for verifying a digital signature can determine the validity of the signatures in integrity terms, but the

S. Fazlagic (✉) · N. Behlilovic

Faculty of Electrical Engineering, University of Sarajevo, Sarajevo, Bosnia and Herzegovina
e-mail: minutolo.a@na.icar.cnr.it

digital signature itself does not provide proof that the signer in question is entitled to sign the document.

The current proxy digital signature schemes do not provide for signature authorization verification [1]. The digital signature technology is used to provide the authenticity of the signer and the integrity of the electronic document, where a digital signature cannot exist without the digital content and is always closely linked to a specific electronic document or its segments. However, the verification of digital signatures in a standard signature scheme does not provide information whether the signer of the document is the authorized signer [2]. Such is the case of signature validation in electronic archive system. The authorization validation problem is particularly evident in cases of validation of signature rights delegation. For example, there is a case where a user has delegated his document signing right to another user whereby he himself did not have the right to sign the document. Using standard mechanisms for verifying the proxy signature will not show that the signature is not valid. This shortcoming will not be so emphasized in signature validation while the document is still in the workflow process, because the signers are expected to recognize the order of the signers and those deputizing them. However, the validation of the authorization of the proxy signature may become a challenge after some time. For example, tenure of a board member has recently terminated. The issuance of a new certificate will cause the previous certificate to be placed on the revocation list, which will make all previous signatures invalid. Therefore, while he retains his digital certificate he will no longer be authorized to sign any future documents. To overcome security threats and weaknesses in the existing schemes, a new proxy authorization model is presented.

In smaller organizations the proving of the authorization of the signers may not be a challenge, since employees are expected to be familiar with the hierarchy of responsibility and the authorization of the signer. However, in large organizations e.g. government institutions this may present a technical challenge due to their size and business complexity. This paper proposes a solution for validation of the authorization for signing rights.

66.2 Related Works

The signer uses a private key to sign documents while signature verifiers use a corresponding public key. To ensure the authenticity of the received public key before validating signatures, verifiers have to verify that public key belongs to the person that signed the document. There are three feasible ways of providing the authenticity of the public key: certificate-based public key system, identity-based public key system and self-certified public key system [3]. In the certificate-based public key system the issuing authority sign each certificate, which guarantees the connection between the identification data and the public key. In the identity-based public key system the authenticity of the public key is implicitly provided. Therefore, the authenticity of the public key must be validated before proceeding

to the signature verification [4, 5]. Checking the validity of signatures is not complete without authorization of the signer.

There are several proxy schemes that meet security requirements, but they offer no solution for controlling the delegation of privileges [1]. The problem is inherited from the standard digital schemes. In the standard signature scheme environment it is understood that the signer is the authorized person. Similarly, in the proxy scheme, the delivery of a proxy key authorizes the proxy signer to sign. Such an implicit authorization scheme cannot meet the requirements, because the delegator often has a need to split privileges between more delegates.

Attribute Certificates (AC) has become a standard way in Public Key Infrastructure (PKI) systems for proving that the entity has the right to perform certain actions. AC, signed by trusted Attribute Authority (AA), binds an identity in the public key certificate with a particular set of attributes. The authorization decision can then be made by verifying the connection between that identity to the attributes. This approach has a disadvantage because the lifespan of the privileges in AC is usually much shorter than the validity of PKI certificate.

Another method is a combination of the AC and proxy certificate. The proxy certificate contains a public and private key pair that is signed by the original certificate. This is a standardized way of transferring privileges to the X.509 systems. A proxy certificate has a short lifetime. Unlike the public key certificate, certified by the Certificate Authorities (CA), the proxy certificate is identified by another public key certificate, allowing or the proxy certificates to be created dynamically without intervention from the certificate authorities. Solutions based on combining information from proxy certificate and attribute certificates has a number of drawbacks with respect to the more complex binding to attribute certificate [6].

In terms of security, any solution based on proxies need to be treated carefully. Since the private key is sent with it and there is no mechanism for revoking proxies, anyone who steals it can impersonate the owner. Even if someone knows that proxy private key has been stolen, there is no standard way to stop it being used. On the other hand, since the proxy certificate usually have a short lifetime, so the potential damage is rather limited. In literature there are many works dealing with authentication of the rights to sign but this still remains an open problem [4].

As shown in [7] the proxy digital signature could be implemented using workflow technology and standard digital signature scheme, showing the necessity of switching a signer dynamically. For example, the process defined in a way that three signatories should sign an electronic document sequentially, should be modified in the event that one of the signers has the right to delegate the signing right to another person and to decide to do so. According to [8], the modification of the process can be done either through changes of the instance of the process definition or of the workflow process model. In the event of a change of the workflow model, all running instances need to be restarted after the rollback. This approach with the rollback is not acceptable because the signatures will be declared invalid, and the user should re-sign a document that has already been signed. So, the solution must support the change process during its execution. Re-routing mechanism for process instances is not the subject of this paper.

The problem of resource authorization in workflow management systems including access right for document signing has become a research topic in the area of workflow processes. In terms of resource control authorization in workflow systems, there are generally several models for authorization. In order to provide the specific task performed by the responsible person, an appropriate mechanism for the authorization must be embedded in a Workflow Management Systems (WfMS). A role-based authorization model is based on dynamic allocation of resources. According to the context workflow activities, allowing the assignment of users and permissions to roles will not properly assign tasks to the right participants. The results of some studies in this area have been reported in [9, 10]. However, a role-based model alone is not sufficient to meet all the requirements of security policies of an organization [11]. To solve issues related to part-time role Bertino et al. have proposed the Temporal-RBAC model [10]. However, periodic enabling of roles and temporal dependencies among the roles cannot handle temporal constraints related to the enabling of the user-role and role-permission assignments. Also TRBAC does not make clear distinctions between well-defined notations of *role enabling* and *roles of activation* [10]. These solutions are related to the authorization of resources, including authorization to perform activities of workflow processes. Although the same can be used for the management of rights to sign the document, these solutions do not themselves provide at a later stage the mechanism for validity of the authorization. This can be overcome by extending signature schemes with a license structure for digital signature [4]. The work done by Ugur and Sogukpinar [4] is closely related to the work being reported in this paper.

66.3 Proposed Solution

The key aspect of the proposed model is that the workflow system provides information to validate the authorization of signatures that become part of the document. To integrate the proxy scheme with the conventional WfMS, the data model of the WfMS needs to be extended. By carrying out the process definition according to the principle of enabling roles and roles of activation [11], WfMS ensures that the document is submitted for signature to the person in the order as provided by the process definition. The workflow engine should authorize task performers at the time of signing in such a way that documents can only be signed by authorized signers. After the user signs a document and submits a task, the latter would be finished and information about delegation recorded in the document and the authorization cancelled [1].

The order of the signers is determined according to the organization's specific business process and the corresponding roles of its employees. The information on the roles of the signers would be contained in an electronic document along with the digital signature.

Each signer is distributed a CA certified key pair, including the WfMS, which has the role of the trusted third party signer. A list of signers and roles is

initialized. The user role assignment list (UA) contains information on the persons and the associated roles delegated. The active role assignment (AA) list contains information on active roles, while the signer role assignment (SA) list includes information on the required roles for authorized signature. For instance in UA list, roles vice governor and governor are linked to the actual names, while in the SA list those roles are linked to signer1 and signer2 respectively.

66.3.1 Design Issues

Generally, in the proxy signature scheme, the principal signer delegates his signing rights to a proxy signer, thereby enabling the proxy signer to sign documents on behalf of the original signer.

In our opinion there are two instances when signing rights need to be delegated:

(1) Planned delegation

The principal signer intends to delegate his/her rights to sign a certain document, for a certain period of time, to another signer before the instance of workflow process definition is initiated.

This design issue will be solved through separating of the signing roles from the signer and binding the workflow process instance to the roles.

(2) Delegation by intervention

The workflow process is stopped due to the absence of the signer and needs re-routing. Workflow processes for signing the document are a human-centric task. In situations where a signer is unable to sign an electronic document, the workflow management system (WfMS) should be able to re-route a document to another person (automatically or by administrators) [8, 12]. Therefore, the workflow process should be defined in a manner that the tasks are linked to the roles and not actual persons. This approach provides a way to switch a real signer dynamically and to continue the workflow process execution [9].

This design issue will be solved by:

- (a) Using predefined alternative order in the process definition
- (b) Repeating delegation and starting process instance from the beginning with new parameters

This solution assumes that the signing of the document is linked to business roles, and not for specific signatories. Process definitions should be organized according to the organizational function, not the actual individuals, using only role from UA list, while workflow execution task should follow information from process definitions and AA list. Electronic document should be prepared for signing according to the information from process definition and SA list. This approach allows the problem of delegation of rights between signatories is

replaced with the problem of granting the proper role to a particular user for a specific period of time.

66.3.2 Data Model

WfMS data model should be extended with data structure for handling signature rights and delegation. This will provide that all the required information can be captured and modeled in the workflow definition.

Data structure *userRoleList*, *activeRoleList* and *roleSignerList* are defined to store, map and update information about the *user role assignment* (UA), the *active role assignment* (AA), and the *role signer assignment* (SA). Information how to verify and authorize delegation to proxy signer, should be contained in the document.

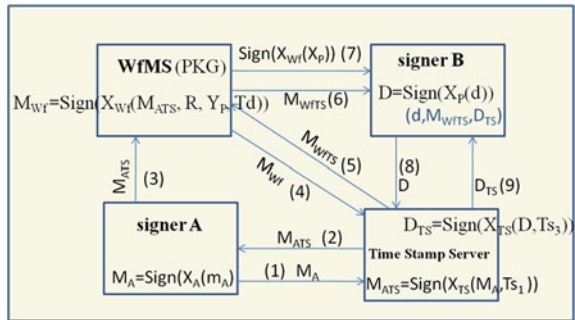
66.3.3 Proxy Signature Scheme with Authorization

The basic idea behind the construction of such proxy scheme is to incorporate warrant inside another warrant which is related to the proxy private and public key. The keys have a short term and are always combined with a time stamp. The following notations are used:

- (X_A, Y_A) : A signer's key pair
- (X_{Wf}, Y_{Wf}) : WfMS signer's key pair
- (X_B, Y_B) : B signer's key pair
- (X_P, Y_P) : proxy signer's key pair
- (X_{TS}, Y_{TS}) : Time Stamping Authority's key pair
- Td: the expire date/time of the delegation
- m_W : the warrant issued by signer A
- M_A : delegation message to WfMS signed by A
- M_{ATS} : delegation message M_A with time stamp
- M_{Wf} : delegation message to signer B signed by WfMS
- M_{WfTS} : delegation message M_{Wf} with time stamp
- Ts: time stamp
- D: signature of document d with X_P key
- D_{TS} : signature D with time stamp
- ID_A : Identity of A signer
- ID_B : Identity of B signer

The process of starting delegation and proxy signature generation is shown in Fig. 66.1. When process of signing is finished, M_{WfTS} and D_{TS} become a part of document d.

Fig. 66.1 Signer A delegates signing right to signer B



Step 1. Initial setup

Each signatory is distributed a certified pair of keys, including the WfMS, which has the role of trusted third party signer. Initialize the UA, PA and SA the list of signers, which contains information about the required roles for authorized signature.

Step 2. Starting delegation process

The signer A signs a delegation message $m_A = (ID_A, ID_B, m_w)$ with his private key $M_A = \text{Sign}(X_A(m_A))$, and delivers it to the time stamp server TS for time stamp and sends the resulting $M_{ATS} = \text{Sign}(X_{TS}(M_A, Ts_1))$ to the WfMS. The WfMS then updates the list that contains signers and their respective responsibilities.

Step 3. Proxy generation and proxy signature generation

The WfMS always has a role of the principal signer, so it generates a short-term key pair (X_P, Y_P) for the proxy signer, designated by A. Before initiating the generation of keys WfMS must consult the authorization list to check whether A is entitled to delegate these privileges to the signer B. The WfMS generates a key pair just before the signature, or when signed by the workflow will start the action of signing the document. It composes a delegation message $M_{Wf} = \text{Sign}(X_{Wf}(M_{ATS}, R, Y_P, Td))$. Td is the expiry time and date of the delegation signing capability for a particular digital signature. Time Td is very close to time stamp, because the generated key pair has very short life time. R contains data about the role for binding the proxy signer and the role-signer. Now the WfMS sends M_{Wf} to TS for time stamp, $M_{WfTS} = \text{Sign}(X_{TS}(M_{Wf}, Ts_2))$. The WfMS then sign short-term proxy private key X_P as $\text{Sign}(X_{Wf}(X_P))$, and send to proxy signer B together with delegation message M_{WfTS} . The proxy signer B first verify $\text{Ver}(\text{Sign}(X_{Wf}(X_P)))$, then authenticates proxy key X_P with Y_P . If it holds then signer B with the key X_P sign the document $D = \text{Sign}(X_P(d))$, and delivers it to the time stamp server TS for time stamp $X_P = \text{Sign}(X_{TS}(D, Ts_3))$. The proxy signer B with the proxy key X_P , for the signer A within the scope of authority, generates proxy signature of the document as tuple (d, D_{TS}, M_{WfTS}) . To prove that is proxy signer B with identity ID_B , he can sign M_{WfTS} with private key X_B $M_{WfTSB} = \text{Sign}(X_B(M_{WfTS}))$.

Step 4. *Proxy signature verification*

To verify the proxy signature (d, D_{TS}, M_{WfTSB}) on document d , given as:

$(d, \text{Sign}(X_{TS}(D, Ts_3)), \text{Sign}(X_B(\text{Sign}(X_{TS}(M_{Wf}, Ts_2))))))$ where $D = \text{Sign}(X_P(d))$
and $M_{Wf} = \text{Sign}(X_{Wf}(M_{ATS}, R, Y_P, Td))$,

by executing the standard verification algorithm verifier first authenticates Y_P

$(M_{WfTS}) = \text{Ver}(Y_B(M_{WfTSB}))$

$(M_{Wf}, Ts_2) = \text{Ver}(Y_{TS}(M_{WfTS}))$

$(M_{ATS}, R, Y_P, Td) = \text{Ver}(Y_{Wf}(M_{Wf}))$

then verify delegation

$(M_A, Ts_1) = \text{Ver}(Y_{TS}(M_{ATS}))$ and

$(ID_A, ID_B, m_W) = \text{Ver}(Y_A(M_A))$

and checks statement $Td > Ts_3 > Ts_2 > Ts_1$ then a verifier should verify

$(D, Ts_3) = \text{Ver}(Y_{TS}(D_{TS}))$

$d = \text{Ver}(Y_P(D))$

If role given with R match the role of role-signer authorization for document signing is proved, and signature is valid.

66.3.4 Implementation Issues

The proposed solution can be implemented using X.509 attribute certificate combined with the proxy certificate taking account of the overlap between their capabilities. The AC certificate is issued by the principal signer, and the proxy certificate is issued by WfMS. They both become part of an electronic document and are logically bound to the digital signature. It should be noted that the actual signer never signs electronic documents with the standard digital signature instead it always signs as the proxy signer. Since very few proxy schemes deal with revocation mechanism [6] that will be addressed in further works.

66.4 Conclusions

In this paper we have introduced generic proxy scheme with authorization. First, we have shown the necessity of signature authorization control, and proposed solution integrating standard digital scheme with workflow management system, in such way that information on the signers' roles from the WfMS be part of the electronic document and serve for later validation of the authorization of the signing rights. Primarily discussed were two delegation types: planned delegation and one by intervention. The former would be based on the separation of the signing roles from the signer and binding the workflow process instance to the role signers. The latter would include either the use of predefined alternative order in the process definition or repeating delegation and starting the process instance with new parameters.

References

1. Xiao Y, Zhou Z (2008) A new kind of self-certified proxy digital signature scheme. In: International conference on computer science and software engineering, vol 3, 12–14 Dec 2008, pp 766–769
2. Leung KRPH, Hui LCK (2000) Multiple signature handling in workflow systems. In: Proceedings of the 33rd Hawaii international conference on system science (HICSS'2000), Hawaii, USA, Jan 2000. IEEE Computer Society Press
3. Hsu CL, Wu TS (2004) Efficient proxy signature schemes using self-certified public keys. *Appl Math Comput* 152:807–820
4. Ugur A, Sogukpinar I (2009) A framework for licensed digital signatures. In: First international conference on networks and communications NETCOM '09, 27–29 Dec 2009, pp 428–432
5. Lee B, Kim K (2002) Self-certified signatures. In: Progress in cryptology: INDOCRYPT 2002: third international conference on cryptology in India. Lecture notes in computer science, vol 2551/2002, pp 199–214. doi: [10.1007/3-540-36231-2_17](https://doi.org/10.1007/3-540-36231-2_17)
6. Internet X.509 Public Key Infrastructure (PKI) (2011) Proxy certificate profile. Available at: <http://www.ietf.org/rfc/rfc3820.txt>, visited, Oct 26, 2011
7. Fazlagic S (2010) Delegating signing capability in workflow systems. In: International conference on computer engineering and technology (ICCET 2010), vol 4, April 2010, Chengdu, China, pp 324–327
8. Qi S, Chuanliang C (2011) The design of dynamic workflow based on document flow system. In: International conference on computer and management (CAMAN), 19–21 May 2011, pp 1–3
9. Tang D, Guo J, Zhang Q (2011) A dynamic workflow authorization method based on participant expression rules. In: IEEE international conference on computer science and automation engineering (CSAE), vol 4, 10–12 June 2011, pp 345–349
10. Joshi JBD, Bertino E, Latif U, Ghafoor A (2005) A generalized temporal role-based access control model. *IEEE Trans Knowl Data Eng* 17(1):4–23
11. Tan K, Crampton J, Gunter CA (2004) The consistency of task-based authorization constraints in workflow. In: Proceedings of 17th IEEE computer security foundations workshop, 2004, 28–30 June, pp 155–169
12. Suyan W, Wenbo L (2011) A flexible official document system based on authority management. In: International conference on business management and electronic information (BMEI), 13–15 May 2011, pp 529–532

Chapter 67

Semi-Agile Approach to Software Development Process

Deniss Kumlander

Abstract Most modern companies realize that the best way to improve stability in the global, rapidly changing world is to be innovating and produce software that will be fully used and appreciated by customers. The key aspect on this road is personnel and processes. One of the best approaches to software development process into follow modern agility techniques. Unfortunately sometimes requirements to execute such are too high although the benefits are also sufficient. The paper proposes that the relaxing of those restrictions in many key aspects formulated as semi-agile technique can improve the level of possible adoption of it without any decrease of acquired benefits. The approach sufficiently differs from the hybrid agile proposals.

67.1 Introduction

The number of projects failures is very high in modern software development despite all newest approaches. Those failed projects' costs are carried by customers and add a lot of extra cost to successful projects. Unfortunately the situation with “successful” projects is not very different as well: just one fifth of the developed functionality is used “often” or “always” and 16 % more “sometimes”. The remaining functionality represents improperly spent development resources as it is either never used or used extremely occasionally [1]. The global market increasing competition between software vendors and much more demanding markets force companies to stabilize their productivity and improve their development process in all possible ways [2]. The key factor on this road is personnel and the process, which is recognized as a very important aspect of company success [3].

D. Kumlander (✉)
Department of Informatics, TTU, Tallinn, Estonia
e-mail: kumlander@gmail.ee

Unfortunately the software industry is a highly technological sector [4] with a shortness of the personnel resource in many countries. Therefore we need both to motivate company employees creating for them challenging and comfortable environment to work in and to be able to overcome the restrictions of resources shortness building up successful teams from the available individuals. This task is far from been trivial and therefore it is not the easiest to solve [5–7]. Many modern software development methodologies rely on the advanced teamwork, which is stimulated by a high level of freedom in several types of decisions. The freedom in decisions is usually described by using an “autonomous team” term, which is defined as a team performing its’ tasks independently and therefore such team granted with a significant respect within the organisation [8].

Generally, the software development process quality depends on many aspects and the team performance is just one of them. The next, but not least one is the process itself. The ultimate goal of software engineering process is to provide customers with tools that will help them to automate their activities or achieve other desired goals. The modern software development faces new challenges as customers demand much higher quality of the released software, shorter development cycle and increased flexibility of defining requirements. Adding to the previously said also external parameters like quickly changing business environment we get a process having a lot of uncertainties where a task of matching expectations and the released software become quite a challenging task especially after many months of development. This huge pressure on software vendors produces a relatively high level of software projects failures. Research show that up to 27 % of all projects fail because customers are not satisfied with the delivered software [9] and a lot of other projects fail since those do not fit into budgets. Sometimes it happens since those projects are having difficulties with meeting customers’ requirements during final stages and are rebuilding the software again and again. This clearly demonstrates existence of gaps between developed software (features, budget) and customer expectations. Therefore it is crucial to follow techniques like agile in order to bridge the gap like that by using reviews, demos, shortened development cycle etc. Unfortunately the reality is not this simple and the agile technique sometimes is unable to bridge the gap or companies are not able to follow them due too high restrictions applied to teams and companies willing to follow them. The aim of this paper is to show a relaxed agility techniques which could both help to eliminate the remaining gaps and make agile techniques more suitable for ordinal software vendors.

67.2 Agile Core Techniques

Agile core techniques include, but are not restricted by the following techniques:

1. Self-organized teams
2. Sprints and backlogs
3. Continues integration
4. Clean code approach

67.3 Self Organised Teams

Some authors see either the autonomous team or self-management team to be a direct synonym for a self-organised team [10, 11], which is generally defined as a team able to act autonomously without the supervision.

There are different types of autonomy and those include external autonomy, internal or individual. Alternatively the autonomy could be applied to different subjects (or set of subjects) like people, planning, goals, products decisions and so forth. External autonomy is defined [12] as the level influence of management and other individuals (outside the team) on the team's activities and the smaller the influence is the higher external autonomy is owned by the team. The external autonomy can be:

1. Obtained by the team as a compromise between organizational (i.e. hierarchical) management style and a need to run projects effectively;
2. Granted to the team by the management deliberately in order to force team independency and stimulate innovating and creative thinking;
3. Occur due any gaps in the management hierarchy.

The internal autonomy defines in what degree all members of the team are involved into making decisions, i.e. where decisions are made jointly or by a very restricted set of chosen members of the team. The individual autonomy refers to the level of independency or each individual member of the group defining what control s/he has over his/her duties/tasks and freedom to re-organise those.

Classically the main requirement for converting a team into a self-organising team is a shared understanding that the constant communication between different people and roles within the team is the only way to achieve the required result in software development process. This should be combined with a clear understanding that each member should do her best in order to complete the project in the fastest and the most efficient way and a team internal trust that each member does his best within his skills field. Although this requirement is generally sufficient to make a team to be successful, there are still a lot of examples where teams failed on this road [10, 13].

Before we introduce semi-self organised teams we should define requirements of the standard self-organised teams approach. The first one is defined for persons included into such teams—they should correspond to the following criteria:

- Has the ability to think independently;
- Has the high level of knowledge or education;
- Has an ability to learn;

The second one promotes the need for an effective collaboration, i.e. a quick and efficient exchange of the information within the team, and is formulated as “The team should be relatively small”. Moreover it is advisable to include into the team individuals having the same level of cross-respect. The main idea of this is of course already formulated by the earlier requirement of having the high level of

knowledge, but it is not always a matter of just knowledge or ability to learn, so it is separated into a standalone requirement. If any team person will not correspond to this, then others will tend to skip him in internal discussions breaking the rule of efficient internal collaboration and so making dead ends where the “team” spent time inefficiently. Moreover other team members will keep such person on second roles managing his/her work and this could potentially lead to conflicts as all developers officially are on the same level. In result this sufficient difference in knowledge and involvement could unbalance the team and blow it up [14]. For example there is a natural conflict between required autonomy levels to move effectively toward the team goal. On one hand the team requires group autonomy, on another hand individual autonomy and those could conflict sufficiently decreasing the team effectiveness if some individuals dominate in their individual autonomy over others.

67.4 Semi Self Organised Teams

The main idea of the semi-self organised teams is to relax strict requirements described above balancing the team by several approaches described below including taking away certain percentage of autonomies.

First of all earlier described internal conflicts are not always that much guaranteed to happen. This greatly depends on the low skills’ members’ ego and approach to the work—have they come to learn or they do have high ambitions and tend to teach others? If they are not egoistic and ready to accept the smaller set of responsibilities and right then the team could become externally self-organised and internally partly hierarchical. Notice that self-organisation principle always assumed to have a flat hierarchic in other words monotone team. In our approach we can build the team combining persons ready to cooperate so that the self-managing team is clustered in the end result into standard members and guided members. This kind of teams is much easier to build up, execute and more efficient to run since experts within the team will make decisions.

Next we would like to define the “semi-self-organised” team—it is a team that is able to act as semi-organised during a short (restricted) period of time and so is able to survive a temporary lack of management without decreasing its performance. Moreover in the long term such teams could require quite a minimal management efforts and the strength of “semi” part could vary basing on independency level given to the team versus the amount of management efforts required to keep the team moving toward milestones defined by the organisation. In other words such team can be guided on monthly base and will not require constant management efforts.

So, the semi self-organised team requires less in order to be produced and is the same efficient and powerful in the short term as the standard self-organised team and is inexpensive to keep as still requires quite minimal management attention.

Especially well those suit for critical or pilot projects, which are normally not long although in the ordinal scenario those work nicely as well.

The additional technique to stabilise the semi-self organised team include the following:

1. Rotate the manager role within the team to extend the knowledge of problem the team as the whole should solve and knowledge regarding the developed product
2. Value individual autonomy and allow to select tasks to keep the tasks challenging and stimulating
3. Promote co-work including high transparency of tasks and progress (i.e. nobody should refuse to report what he is doing at the moment to anybody), code reviews motivating the knowledge transfer including the experience transfer and pair programming to motivate lower level employees to get involved into “high matters”.

67.5 Sprints and Backlogs

Majority of agile teams organize the documentation in a form of backlogs where you list up all the features you are going to implement basing on the current knowledge of the future release. The last part of the sentence is important since the goal of the product can be shifted during the project since, unlike most historical approaches, the product development process is organised into sprints, i.e. into development increments lasting several weeks (typically 3–4). Each increment should produce, in most cases, ready to release software (addition). Such sprints always end up with a demo to product owner and other involved parties during which the software is either accepted or rejected or small modifications are required. These demos also clarify the current position and allow evaluating the goals basing on visualisation of current status of the project during the demo and so help to reroute the product evolution in the correct direction. Notice that this approach is based on constant evaluation of the product and help us to correct the implementation process on early stages. Moreover the management team always is able to see both the progress and have visualized result versus management/customers’ expectations shifting the product from virtual world to reality step by step.

Besides, the backlog is built for each sprint dividing the functionality, included into it, into small pieces as well as defining a responsible person for each task. Beside it introduces a time line drawing the optimum line from the beginning and the current progress line by decreasing the amount of work remaining on the daily basis. This way we do achieve

1. Visualisation of the progress which is used by both team developing the software and management to detect potential problems early
2. Combining personal responsibility and team one. So the team has enough information to help owner of critical tasks, but no task has a collective ownership leaving some to have none to work on.

67.6 Semi Sprints

The overall idea of sprint and backlog is superior not only to progress the work but also to resolve uncertainties in the project in case of corresponding planning [15] and practically needs no changes except a note that there are a small set of project types where this technique cannot be employed [16], which mostly does not overlap at all with the commercial software production cases.

The only problem of this part is related to the main goal of executing sprints: provide early the progress and receive the feedback from the product owner regarding both correctness of the current sprint result and future sprints (directions of product development). Although it sounds like exactly what product management wants from the development team it is not always the case especially if the company have a long experience consuming older approaches the software development. This produces problems first of all because this feedback cycle is not one directional, but is nothing else than a collaboration between product owner and development, testing teams. In other words in some companies development team can become much more interested in getting feedback than the management in reviewing or providing such. The reason is simple—although development team can carry on traditionally by goals formulated in the beginning of the project, consuming the agile approaches they quickly become fond of it basing on experience and will try to remove the overload peaks by ensuring early correctness of delivered iterations. This matter can be complicated in distributed teams where the product owners teams located far from development center and so cannot be easily accessed by team leaders.

This problem can be resolved by establishing the ambassadors' roles in the overseas teams, who should represent the team needs within another team, motivate to provide feedback, remind, push and act in other possible ways in advantage of the team they represent. Here we can see the possible shift of the R&D manager roles which will be less involved into direct management of the teams due self-organisation of those and more into organisation there work including their representative part defending development team interests.

67.7 Remarks on Continues Integration

The continues integration process is a power mechanism to monitor the status of the project from the build, compilation and integration of different system modules points of view. We will not describe this process here in more details as this is a well-established and widely used technique, but instead will proceed to proposed improvements.

The continues integration process is not a standalone one, but should be extended by the following:

1. Continues awareness of testers on completed elements and partly developed in order to avoid producing misleading error reports on parts in development and concentrating efforts on completed elements;

2. Continues testing even if no test driven development is used—automating the testing process as much as possible on all levels (concentrating first of all on high or meta levels) to identify broken parts;
3. Continues performance testing to detect both trends and arising performance problems related to the latest changes in the code on early stages

67.8 Semi-Clean Code Approach

The clean code approach requires avoiding solution complexity increasing in advance. In other words if there is no direct use to build an advanced solution you should restrict yourself to the simple solution sufficient enough for the current problem or problems that will arise in the nearest future. The approach is expanded also to documents and comments. Documenting (in form of specifications) should go to minimum or be avoided completely. The same applies to writing comments, which is partly replaced, in compare to the old style software development, by the test driven development in agile practices.

Unfortunately the company personnel doesn't always have enough experience and education to act as software architect and correctly

1. Determine on the fly the correct granularity of the architecture—does the current addition require more general architecture shift or can be local;
2. Identify right moment to look at the system from the upper level in order to refactor globally;
3. Build up the architecture knowing all the features presented in the software including current functionality and planned (here the planned is used for cases when you have two similar alternatives and selecting one of them based on planned features will not mean increasing the complexity in advance, but will mean decreased probability of having to refactor in the nearest future).

Considering all said above we can conclude that the ad hoc approach cannot be applied in most projects and the help of software architecture is needed. Can it be provided on the fly during the system building as each sprint defines what will or will not be developed? Unfortunately some software developers not only unable to define the classes and architecture by themselves but also unable to ask right questions during the implementation phase, which leads to constant believe that nothing global should be build. This will end up in a huge mess (of classes and interfaces) without any layers or granularity, i.e. results in a system, which is both complex to maintain and develop further as any new development produces problems in many other software areas making it extremely hard to balance.

Therefore we propose that software architecture role should be seen as equal UI designer role etc., in other words it should be a single responsibility for the person to determine correct software architecture, plan its evolution basing on the sprints initial plans and produce architecture documentation which should be delivered to all

team developers. This will not only centralize the process of building architecture, but also will give guidelines to developers sufficiently simplifying their tasks.

From our point of view basing on experience in many projects the question of commenting code is also an essential one in order not to lose the knowledge within the software companies. The problem here is none constant list of team members—unfortunately all motivational techniques we do employ [3–5] do not always help us to keep workforce stable and sometimes the understanding of the code without the author become practically impossible. Therefore, together with naming conventions, the most sophisticated code pieces should be commented and evaluated (both code and sufficiency of comments) during the code review.

Finally, we would like to discuss the question of eliminating the documentation from the agile development. There are a lot of good reasons to follow this advice:

1. The initial full documentation cannot be correctly formulated from the beginning since we don't know how sprints, demos and changing business requirements will change the project end result;
2. The documentation covering the current sprint in most cases contains huge gaps since during the implementation a lot of crucial questions are stated by developers, which obviously was hard to foresee writing the initial specification draft. Those questions in most cases are solved ad hoc without revising the specification. In the result the specification is dangerously incorrect and misleading and cannot be correctly used by the testing department to evaluate the software

At the same time the fixation of what is done should be still made in some way to provide testers with enough knowledge on the built features. In order to overcome problems we do propose the following: although the agile team is normally small enough to generate constant knowledge transfer, there are two groups of people knowledge transfer among which is crucial—analysts and testers, since those represent the beginning and completion stages of the product development process and the cycle should be closed in order to ensure correctness of development. Therefore we should insist on putting their work tables as close as possible. Moreover the idea of pair programming should also be evolved beyond the development team and applied to testers and analysts. Those could form a team both planning the functionality (so testers could start planning test cases with analysts) and testing, so analysts will educate testers and help them to understand the context (the reason) of developed features.

67.9 Conclusion

In the paper we have revised base approaches of the agile software development and argued how it can be redefined as a semi-agile approach for cases and teams, which cannot execute the pure technique. Following the semi-agile principles those teams will be able to derive majority of agile process' benefits staying within

the restrictions they are unable to overcome either temporary or constantly. In the first part we have reviewed self-organised teams. First of all the paper proposed changes organising teams by relaxing restrictions and loosening certain percentage of autonomy that doesn't result in any sufficient decrease of performance in short term and is balanced by simply additions in the long term projects. The paper also introduced semi-sprints and semi-clean code approaches in order to overcome typical problems of the agile process bridging remaining gaps between theoretical vision and practical result of project organisation following modern techniques affected by uncertainties and lack of stability in documentation and personnel.

References

1. Khan AA (2004) Tale of two methodologies for web development: heavyweight vs agile, Postgraduate minor research project, 2004, pp 619–690
2. Kumlander D (2006) Software design by uncertain requirements, Proceedings of the IASTED international conference on software engineering, 2006, pp 224–2296
3. Armstrong M (1991) A handbook of personnel management practice. Kogan Page, London
4. Rauterberg M, Strohm O (1992) Work organisation and software development, Annual review of automatic programming, vol. 16, pp 121–128
5. Daly D, Kleiner BH (1995) How to motivate problem employees, Work study, vol. 44(2), pp 5–7
6. Gerhart B (1987) How important are dispositional factors as determinants of job satisfaction? Implications for job design and other personnel programs. J Appl Psychol 72(3):366–373
7. Herzberg F (1987) One more time: how do you motivate employees? Harv Bus Rev 65(5):109–120
8. Guzzo RA, Dickson MW (1997) Teams in organizations: recent research on performance and effectiveness. Annu Rev Psychol 47:307–338
9. Kumlander D (2009) Uncertainties management framework—foundational principles, Proceedings of the 11th international conference on enterprise information systems, pp 103–108
10. Moe NB, Dingsoyr T, Dyba T (2008) Understanding self-organizing teams in agile software development, Australian software engineering conference, 2008, pp 76–85
11. Fenton-O'Creevy M (1998) Employee involvement and the middle manager: evidence from a survey of organizations. J Organ Behavior 19(1):67–84
12. Hoegl M, Parboteeah KP (2006) Autonomy and teamwork in innovative projects. Hum Resour Manag 45(1):67–79
13. Kumlander D (2006) On using software engineering projects as an additional personnel motivating factor. WSEAS Trans Bus Econ 3(4):261–267
14. Langfred CW (2000) The paradox of self-management: Individual and group autonomy in work groups. J Organ Behavior 21(5):563–585
15. Bennatan EN, Emam KE (2005) Software project success and failure, Cutter Consortium, <http://www.cutter.com/press/050824.html>
16. Kumlander D (2006) Hybrid software engineering. WSEAS Trans Bus Econ 10(3):686–691

Chapter 68

The Influence of Student Body-Talk Reaction in Formulating Effective Teaching Strategy

Ahmad Sofian Shminan and Runhe Huang

Abstract This research attempts to explore and elucidate the potential applications of educational technologies such as ICT and ubiquitous computing technologies in solving core domain problems that exist in lectures of higher education. This article emphasizes the importance of student-centric awareness in an effective teaching system. Apart from the continuously integrated and updated personal profile, non-verbal communication such as students' body-talk reaction is one of the important elements of reading and understanding students in the classroom. In the proposed effective teaching system, student facial expression and sitting posture are continuously monitored and analyzed, based on the applicable set of effective teaching support agents are employed to support a teacher to make a conclusion, take an action, or automatically involve in the teaching process.

68.1 Introduction

Students are the most important individuals in the classroom learning environment. Passivity of students to engage in any activity and superficial learning to interact with the instructor is an indication of the failure to create a positive atmosphere of learning required in classroom. A teacher should not regard the task of creating a positive classroom atmosphere to be easy. In fact, it is a challenge.

A. S. Shminan (✉)

Graduate School of Computer and Information Sciences, Hosei University,
3-7-2, Kajino-Cho, Koganei-Shi, Tokyo, 184-8584 Japan
e-mail: wakahara@hosei.ac.jp

R. Huang

Faculty of Computer and Information Sciences, Hosei University,
3-7-2, Kajino-Cho, Koganei-Shi, Tokyo, 184-8584 Japan

There is a variety of effective teaching strategies, one psychological approach addressed to create a positive atmosphere in the learning process is to apply non verbal communication [1, 2]. It requires a teacher to read and understand the instinct of the heart, personal needs and reactions through observation of student reactions during real-time learning process. A teacher can use students' facial expressions such as smiling, frowning, nodding his head, biting lips, and sitting posture for instance, slanting on the chair or lying on the table, as a source of feedback while delivering the content. A teacher can use this information as input in the decision whether to accelerate or stunt the speed of lectures delivered. Review by Jolly [3], shows that about 35 of verbal communication and 65 % of non-verbal communication needs to be done by a teacher when teaching to ensure effective teaching and learning process. Having these skills and understanding is not easy though it is considered as an advantage for a teacher who owns it. Even though a teacher possesses such teaching skills, it is certainly not very practical when it comes to a room the size of a huge lecture hall.

With the rapid developments of information, communication technologies and the advance researches on ubiquitous computing, the convenience of monitoring students' activities and behaviors with wired or wireless surveillance cameras and other sensing devices in classroom learning becomes possible. Student-centric effective teaching strategies [4] that blend effective, dedicated and systematic such as group work, role play, paraphrasing, debate and etc. can be applied accordingly upon the conclusions of being aware of students' learning situations. Success in implementing educational technology-based teaching strategies has been reported in various field studies [5], [6] and [7]. Thus, the direction of this research is towards progress on how to help teachers create effective elements and active teaching-learning process which is conducive in the classroom atmosphere. This article is focused on describing a student-centric effective teaching system in which students' facial expressions and sitting postures are continuously monitored and analyzed at every specified interval. A set of effective teaching support agents consisted of body-talk reaction context awareness agent, teaching strategy agent, learning assessment agent, and teacher support agent, etc. are employed to support a teacher to improve the teaching effectiveness and motivate students to actively involved in the learning process.

68.2 Related Case Studies

Effective teaching is not a new topic. There has been a long history and intensive studies had been conducted by many educators, educational organizations, ICT and ubiquitous computing based education system developers and researchers. Although those studies are for different subjects, different groups of students, different learning objectives, we detect the existence of diversity in effective teaching methods, strategies, and systems.

The use of interactive technology alongside traditional lecture delivery is in fact seen to contribute toward creating an active learning environment. An interactive

lecture application has, in fact, been developed before. Its existence differ in variance and type [8]. Meanwhile in 2005, a group of researchers from the University of Wake Forest succeeded in developing the second generation of PRS; it was named the Class in Hand application, which operated using a PDA. The developed application is able to add specific additional abilities to assist interactive communication between lecturer and student [9]. The development of this interactive lecture application continued when an application named Concert Studio was produced by Fraunhofer Institute for Integrated Publications and Information (IPSI) Germany [10].

In another research dimensions, two researchers from Japan has developed an application called uClassroom [11]. This application is used as a facilitator in the CMS application. They describes the concept of ubiquitous classroom and its implementation that enables us to expand the awareness among faculty and students in classroom. One-to-one technology is getting more and more attention, and it will start to make changes to education [12]. The researcher apply self-paced learning as the pedagogy in one-to-one classroom, and develop a monitoring system to help teacher to handle the classroom. The focus of this research is to develop the tools that can be used by lecturers to monitor teaching in the classroom learning. Technologies when embedded in teaching strategies to support the cognitive and social process of learning, can provide unique opportunities for teachers [13]. Under the face-to-face learning environment, the Classroom Response System based on affective computing can effectively capture the learning outcomes of students immediately then send it to teachers as feedback [14]. According to this researchers, with this teacher support tool, interactive learning as questions and answers (Q&A) can be much easier and fun by using handheld transmitters. This researcher proposes a classroom response system, and this system will be used to achieve the effectiveness of learning through the test; the same time when using webcams to record learner s facial expressions.

Through previously mentioned analysis of studies, these domain problems can be resolved by employing an active and interactive approach in the teaching and learning process. However, based on study results analyzed by the researcher, previous studies have been focussed on the improvement of lecture delivery. Moreover, any study involving the teaching and learning aspect that attempts to explore student learning needs based on body-talk reactions and learner profiles was not clearly defined. And so, it is the researcher's view that this study may be highlighted because it contributes greatly to the field of education.

68.3 Proposed Approach

The grass root framework for the development of such personalized ubiquitous education system consists of several main components, namely Student Context Resources, Human-centric Information Processing, Agents Service Platform, and Effective Lecturing Framework. Figure 68.1 gives an overview of the proposed

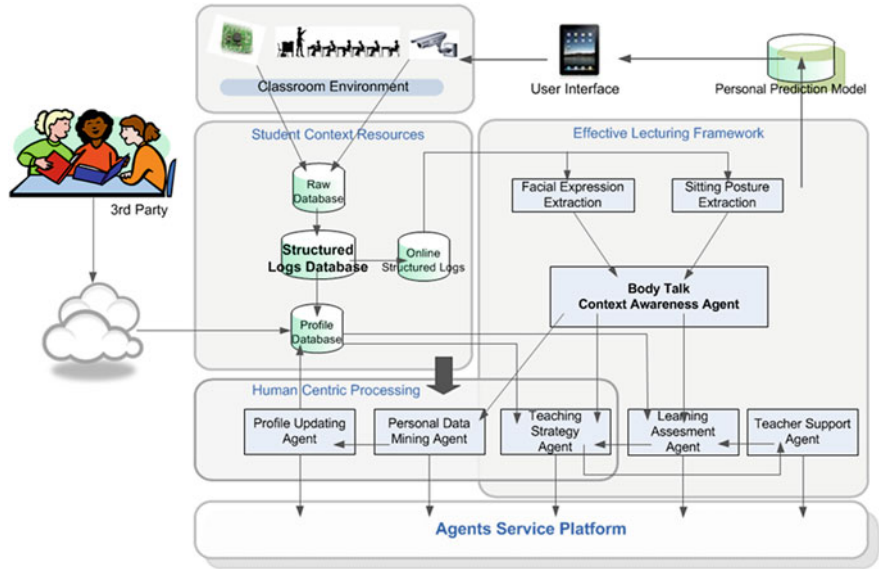


Fig. 68.1 Overview of the proposed system

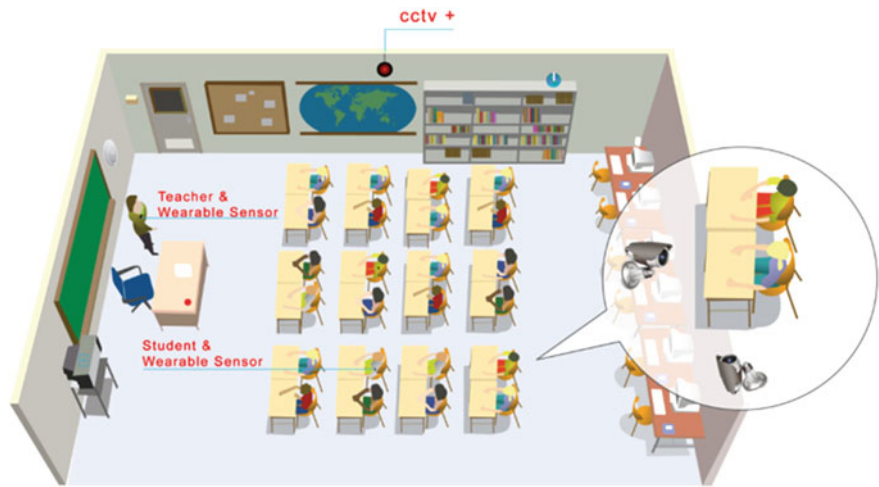


Fig. 68.2 A scenario of learning in class

system framework. As shown in Fig. 68.2, some commonly used ubiquitous devices and sensors are equipped in the classroom.

Some kinds of students’ activities and behaviour are monitored and recorded. In particular, web cameras and Kinect sensor are set up to catch each student’s face image from the front direction and sitting posture from the side.

All raw data from all devices and sensors are collected and stored in Student Context Resources. The facial images and sitting posture images of students in the classroom are stored in Online Structured Log but processed in Interactive Lecturing System. From the results of the facial expression extraction and the sitting posture extraction of students, the awareness agent concludes each student's body-talk reaction, which is one of the sources that contributes to teaching strategy generation. Other two sources for teaching strategy generation are from the online learning assessment and the personal profile. A personal profile is generated and updated offline. The personal profile contains personal information and as well as personalized features such as personality, interests, grade records, etc. The combination of static offline information from the personal profile and dynamic online conclusions from the body-talk reaction awareness, and also the learning assessment is supposed to produce effective teaching strategies for supporting a teacher to conduct interactive lecturing in the classroom. All agents that play different roles are running on Agent Service Platform. They communicate and collaborate each other towards a same goal: providing an interactive teaching environment for students in a lesson or a learning activity.

68.4 The Student Context Resources

Student context information are the most important resources for identifying students' situations. Data monitored by sensors and recorded by devices in the classroom makes up a huge amount of raw data. There should be a reasonable structure for organising and managing the raw data.

Here, we introduce the concept of log; a certain data structured database, such as space log, student log, teacher log, device log. It is supposed to be easier for retrieval and the use than the raw data or the conventional database in a particular application such as student facial expression extraction and sitting posture extraction. It is worthwhile to point out that the hierarchical log system enables a log database be dynamically rearranged and restructured according to an objective or an application. Figure 68.3 gives an example of the hierarchical structured log data.

68.5 Body-Talk Context Awareness

According to an article 2006 written by Kajita and Mase [11], the process of learning and teaching in a higher learning institution, the most common and main activity that takes place is the interaction between teachers and students, the classroom on the other hand is the location that permits this. With this interaction process, a teacher will be able to identify each student's wellbeing. To acquire and measure a student's level of acceptance and understanding during a lesson can be

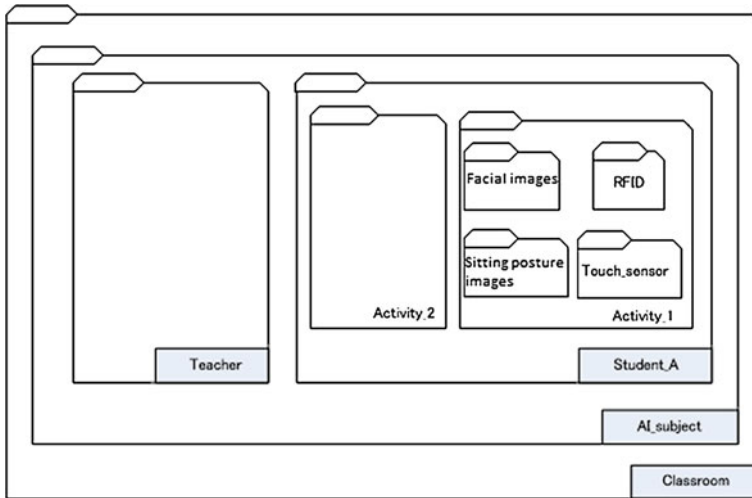


Fig. 68.3 An example of the hierarchical structured logs

seen by identifying the spontaneous reaction resulting from their facial expressions and movements. If the lesson is interesting and meaningful, students will react positively with a high level of interest as oppose to behaving indifferently, not paying full attention when the lecture is conducted bland and boring. As a good teacher, he/she is able to read students' reactions from their facial expressions and sitting postures and to understand students' learning states accordingly to take appropriate teaching measures so as to reach the maximum teaching effectiveness. Therefore, the facial expression extraction and sitting posture expression of students in a lesson or a teaching activity are two key components in the Interactive Lecturing System.

68.5.1 Facial Expression Recognition

Mehrabain in article by [15] states in her research, verbal communication contribute only 7 of total communication, vocal tone effects 38 whilst facial expression constitutes 55 %. This research shows that students' facial expression during class is paramount to learning and teaching strategy. A commonly used physiological framework by researchers as a guide to explain specifically facial expression is the FACS (Facial Action Coding System). Paul Ekman designed the FACS framework in 1970 [16].

To identify students' facial expression during class, a researcher need to develop a facial expression recognition. Usually, facial expression recognition goes through three process sequence beginning with input of facial data that is

recorded using web camera, feature extraction by deploying compatible algorithms and lastly, facial expression classification according to categories set in FACS framework.

The proposed approach consists of 3 main stages namely image pre-processing, facial expression feature extraction and facial expression classification. 2The description of each process is explained below:

- a Image Pre-processing: The image intensity was normalized using the histogram equalization. The face area of an image was detected using the Viola-Jones method based on the Haar-like features and the AdaBoost learning algorithm [17].
- b Facial expression extraction: The facial region is then converted to a column vector which forms the feature vector. The feature vectors suffer from high dimensionality, which can cause over-fitting during classification. One approach to reduce the dimension of the feature vectors is to apply principal component analysis. According to Turk and Pentland [18] article, the Eigen face is efficient approach to determine human face space. In this technique the image frontal face image can be constructed based on the small collection of weights for each new face.
- c Facial Expression Classification: The output from the stage above is analyzed by classifiers. In this paper, researcher implemented feed forward propagation network to classify two expressions (sad and happy). This Network has contained of two layers, the mid layer has 10 neurons and the output layer has 4 neurons [19].

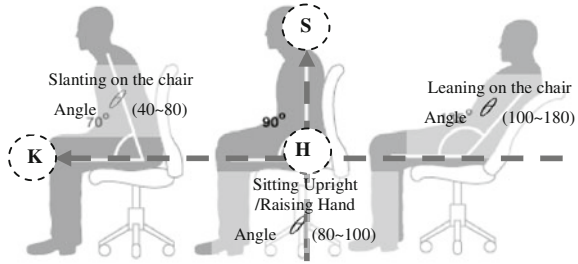
68.5.2 *Body Postures*

Apart from facial expressions, body positioning and movement are also considered as a source of non verbal feedback that can help an teacher assess and determine the level of acceptance and understanding of students [20]. In a learning environment, students are usually in a sitting position when listening to lectures [21], raise hands to ask questions and stand up upon answering a question. By observing the student's sitting position styles, such as sitting up straight, sitting slanting or sitting with crossed legs allows the teacher to value the level of concentration or focus of students to hear lectures delivered. In addition, the position and movement of the body is also said to have a strong relationship with the personality of a student [22].

(1) Associated with personality

A psychologist by the name of Carl Jung [23] said that students who actively ask questions during lecture sessions were classified as students with extrovert personalities, while students who sat silent, passive and do not interact directly

Fig. 68.4 Tracking of the shoulder, the hip, and the knee



belong to the introvert personality. As a course instructor you would expect an active interactive learning environment, however there will be a challenge of creating this environment since majority of the students remain passive and do not want to interact. Understanding a student's personality is a relatively long-term process and should be done from multiple angles. However, as one of non-verbal communications, a student sitting posture habits can reflect some aspects of his personality. In real learning situations, a continuous collection of a student sitting postures can somehow contribute to the personal profile's generation and updating.

(2) Associated with student learning state

From student's sitting postures, we can understand his certain learning states, such as concentrating (sitting upright), active (raising hand), non-focusing (slanting on the chair), and sleeping (lying on the table), and dynamically change teaching strategy. For example, if there are many students who are asleep during lecturing, the teacher agent may take a certain action, such waiting up those students, changing lecturing style like making a joke or telling a story, etc. If many students are in unfocused state, the teacher agent may invoke a strategy to draw their attentions. Therefore, the student sitting posture is also one of the factors to know students' learning state.

(3) Sitting posture recognition

The sitting posture recognition is based on the Kinect tracking system. A Kinect sensor [24], a horizontal bar, featured with an "RGB camera, depth sensor and multi-array microphone running proprietary software is placed by the side of each student in class. The Kinect SDK framework provides the interface API for writing applications utilizing natural interaction. This research uses the API for a student's sitting posture vision and producing sensory data. Together with Ms Kinect SDK which supports skeleton tracking, this system can perform the student's skeleton tracking and obtain 3 coordinates (x_s, y_s, z_s) , (x_h, y_h, z_h) , and (x_k, y_k, z_k) of the positions of the student's Shoulder, Hip, and Knee as shown in Fig. 68.4. By observation, it is noticed that a sitting posture is related to the angle between the line \overrightarrow{HK} (from hip to knee) and the line \overrightarrow{HS} (from hip to shoulder). The formula for calculating the angle is given below.

Table 68.1 A student’s body-talk reactions

| Facial expression | Sitting posture | Body-talk reaction |
|-------------------|-------------------|--------------------|
| Not understood | Non-concentrating | No interests? |
| | Sleeping | |
| | Concentrating | Hard working? |
| Confused | Raising hand | Active? |
| | Non-concentrating | Inactive? |
| | Sleeping | |
| Understood | Concentrating | Hard working? |
| | Raising hand | Active? |
| | Non-concentrating | Smart? |
| | Sleeping | No interests? |
| | Concentrating | Expected |
| | Raising hand | Expected |

$$\theta = \cos^{-1}(\frac{\overrightarrow{HS} \cdot \overrightarrow{HK}}{\left|\overrightarrow{HS}\right|\left|\overrightarrow{HK}\right|})$$

(68.1)

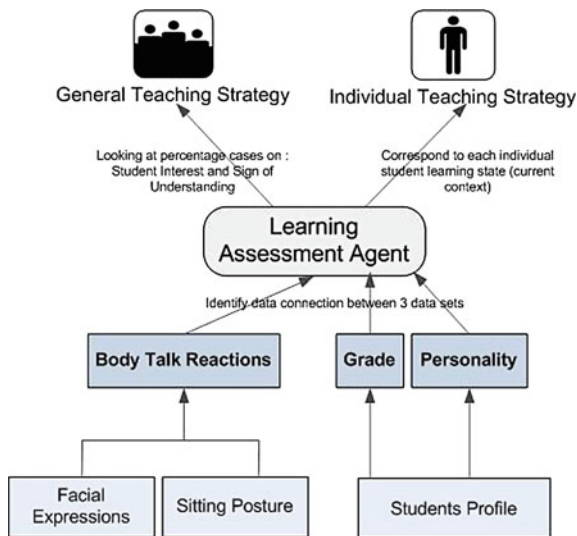
68.5.3 *Body-Talk Reaction*

Based on the student’s facial expression and sitting posture, the system has a mechanism to conclude a student body-talk reaction, which the system is used to further to understand a student’s current learning state. For example, if a student shows an understood face and concentrates on the lecture, it can be concluded that the student may be a good student. However, the important thing for the system is to check out those students who may have some problems and give the teacher some hints from their body-talk and to pay more attentions on them. For example, if a student shows a non-understood face and does not concentrate on the lecture, it can be concluded that the student may lose motivation or has no interests in this lecture. For confirmation, the system should further check the student’s profile and access his personal information or data such as his personality and grade.

In some cases, it is difficult to draw a conclusion from a student’s body talk. For example, if a student shows an understood face but does not concentrate on the lecture, it is difficult to know whether the student is smart or has no any interests in this course and does not care about if he understands the lecture or not. In this case, the system will check the student’s record or get the lecture’s opinion.

As shown in Table 68.1, the items attached with the question mark, ‘?’, require further input from personal profile or or the teacher so that it can be concluded. For those students who have no special problem, the system leaves them alone. For those students who seem to be having some problems, the system further checks their personality and grade in their personal profile. As for a final conclusion, a

Fig. 68.5 Learning assessment agent process



corresponding teaching strategy or a set of strategies in class is dynamically generated or selected from teaching strategy database.

68.6 Learning Assessment Analysis Agent

Situational data context pertaining to facial expression, body language and student's profile which have been collected through student context resources log' will undergo data assessment analysis process. Through this agent assessment learning process, knowledge can be extracted from body talk reactions and students' profiles data and interesting relationship among the two data sets can be recognized, automatically. Figure 68.5 shows a visual representation of how this process should occur. To illustrate the above visualization process in more details, the following describes three main processes that support the process.

- Extracting student profile information from the personal profile including the 3rd party data resource. Student profile information is needed to determine the individual-based teaching strategies. At this point, learning assessment agent will identify and analyze data relationship between body talk reactions (facial expression and the sitting position) with the student profile data (grade and personality).
- For the general strategy, the determination of strategies based on the total percentage of cases showed positive or negative reaction of students during the learning session. For example, if the percentage of the majority of students understand the lectures presented, it would not require a teacher to change teaching strategies. However, a teacher needs to make modifications to the

teaching strategies by adding some interactive elements if the majority percentage of cases shows students facial expression reaction (not understand) and position seating 'not concentrate' status.

- Meanwhile, the determination of individual-based teaching strategies are based on the current context of each student's situation. In this situation, a variety of contexts might happen in real time classroom learning session, such as students sleeps, students with a passive personality, or student achievement is a very poor during the test

68.7 Effective Teaching Strategy Agent

Upon recognizing student's learning states in a class, the interactive teaching strategy agent will generate a new or propose an existing teaching method to apply to a certain situation. The teaching strategies are roughly classified into general strategies and individual oriented strategies.

68.7.1 General Strategies

The general teaching and learning atmosphere can be concluded upon the analysis and statistic results of dynamical data such as students' facial expression and sitting postures, offline data from students' profiles. Although the teaching and learning atmosphere can be classified into many categories, this research is aimed at handling only the following a number of significant cases in terms of how many percentage of students show their interests in the lecture and signs of understanding the content in class.

- (1) There are more than 70 % of students who show their interests in the lecture and signs of understanding the content in class. The points of the strategy:
 - Basic teaching plan remains unchanged, identifying the problems of each of those students who often shows a negative facial expression and uncomfortable posture and applying individual oriented strategies.
- (2) There are more than 70 % of students who show negative facial expression and uncomfortable sitting positions in class. The points of the strategy:
 - Conducting online mini test, mini quiz, and/or mini questionnaire;
 - Revising teaching plan by using one level up teaching elements if the lecture is too simple, using one level down teaching elements if the lecture is too difficult, or adding some interesting teaching elements.

68.7.2 *Individual Oriented Strategies*

In order to achieve this, it is necessary to be aware of individual students, which require not only instantaneous data like facial expression and sitting posture at time t , but also continuously recorded data for extracting student's personal features like personality. This system is a self-circulation system in which recorded data in class and concluded body-talk awareness continuously contribute to the personal profile via the personal data mining. The personal profile in turn contributes to the process of generating individual oriented strategies.

(1) Data source from student body-talk reaction

- In-class data log: In class, the body-talk reaction context awareness agent collects each student's facial expression and sitting posture at every specified interval in class, concludes student's body-talk reaction, and sends all three kinds of data sources, facial expression extraction result, sitting posture extraction result, and body-talk reaction result to the personal data mining agent. The personal data mining agent store them into the in-class data log.

- Student study related feature analysis:

A personal profile's form is a long-term accumulation process and result from many aspects of data sources. In this research, it is assumed that each student has a personal profile in the system. Of course, the data resources come from the external, the 3rd parties as well as from the system itself, the self-circulation system, namely, the personal data mining and personal profile updating mechanisms. The personal profile is an important information source for strategy decision-making. At the current stage, this research takes two personal factors: personality and grade, from the personal profile as the input attributes of the decision-making process.

(2) Information from the personal profile

A personal profile's form is a long-term accumulation process and result from many aspects of data sources. In this research, it is assumed that each student has a personal profile in the system. Of course, the data resources come from the external, the 3rd parties as well as from the system itself, the self-circulation system, namely, the personal data mining and personal profile updating mechanisms. The personal profile is an important information source for strategy decision-making. At the current stage, this research takes two personal factors: personality and grade, from the personal profile as the input attributes of the decision-making process (Fig. 68.6).

(3) ID tree and strategies:

Together with another two input attributes, student facial expression and sitting posture, two identification trees based on a number of sets of training data are generated, one is for concluding if a student is active or passive, and another is for concluding if a student is interactive or not. Figure 68.7 shows an ID trees resulted from using the training data partially given in the top left and information gain based identification tree algorithm.

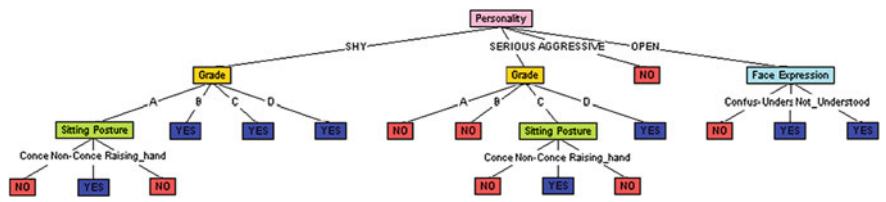


Fig. 68.6 ID tree for concluding if a student is passive

With the resulting ID tree, we can input a student’s online data (facial expression extraction and sitting position extraction results) and offline data (personality and grade retrieved from the profile) and output the conclusions if the student is active or passive in learning. Seemly, we can use training data to generate other ID trees and use them to make conclusions, such as if a student is interactive or not, if a student is interested the lecture and etc.

Corresponding strategies

- For those students who are actively involved in learning and have good grade, the system will appraise them and give one level up learning content for their self-study to encourage their challenging spirit.
- For those students who are interactively involved in learning and of open character, the system will encourage them by appraising them and give them chances to play important roles in interactive teaching, such as leading group discussion.
- For those students who are non-interactive or inactive type, the system will employ a mechanism to increase their confident and encourage them actively involve in interactive teaching, such as raising simple rather than difficult questions to those students. In a group discussion, if possible, it is better to arrange them in the group with some classmates or a team leader who are of mild personality and can take care of and encourage them.
- If the output is Null, it means that the training data is not sufficient. The system will accumulate more student data and use them for training ID trees. The ID trees are continuously refined towards better and better quality of trees. With better quality of trees, the conclusions from the trees will be more precise.

The recommended teaching strategies are sent to the teacher support agent for assisting a teacher. The teacher support agent may recommend the teacher to conduct another teaching plan or mini- quiz/test/questionnaire according to a general strategy. Possibly, it automatically send a message to a student for encouraging and motivating the student according to an individual strategy, and at the same time inform the teacher. It is also likely to automatically provide a student a set of easier or more challenging learning materials and exercises according to the individual’s learning state and level.

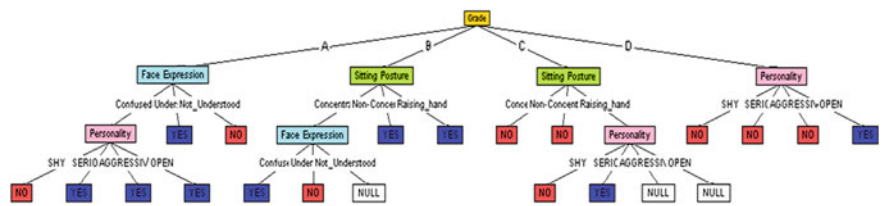


Fig. 68.7 ID tree for concluding if a student is interactive

68.8 Scenarios

Let us simulate a teaching scenario in which there are 100 students, a teacher, and 2 teaching assistants (TA). The teacher is conducting a course, Java programming. Since it is a programming course, it is composed of lecturing and laboratory session and has two 90-min units.

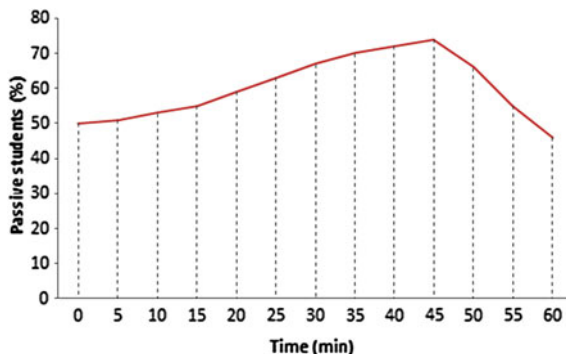
(1) To enhanced the proposed system

- There is a webcam in front of each student for catching a student’s face expression. The system performs the facial expression extraction and outputs each student’s facial expression type every 5 min.
- There is a Kinect motion sensor installed by the side of each student for tracking a student’s sitting posture. The system performs the sitting posture extraction and outputs each student sitting posture type every 5 min.
- Getting ready of 5 sets of teaching plans for corresponding to different levels, A + , A, B, C, D of students.
- Getting ready of 5 sets of mini- tests, quizzes in different levels, A + , A, B, C, D.
- Getting ready of a set of hints, source program outlines, solutions with explanations for corresponding to different levels of students.
- Conducting questionnaire at each of 5 stages for collecting students’ feedback

(2) Learning state scenarios

- Scenario-1: handling non-interactive students
It is thought a very normal case that in class a teacher raises a question and expects some students to answer. However, it is not the case in Japan. To the contrary, it is often case that there is none to answer the raised question. Why? Do not they know the answer to the question? No, it is not the cases. It is because they do not want to stand out and to be different from others. Environmental and cultural factors are more accurate as a contributor to why this situation occurred. With the conventional teaching means, it seems not easy to conduct interactive teaching in Japan. With our proposal, the system keeps recording students’ facial expression and sitting posture and processes them

Fig. 68.8 Number of non-interactive students versus time



every 5 min. Together with each student's inputs (students' personality and grade) from his/her profile, the system can conclude each student's one of learning states, if he/she is interactive or not at every noted moment using the ID tree.

The system continues to monitor the statistical data every 5 min interval and it invokes the strategy mechanism if the number of non-interactive students versus the total student number is over 70 %. As shown in Fig. 68.8, at the time coordinated as 45 min, the number of non-interactive students increases to 70 %, the system reminds the teacher whether he/she takes action by selecting a strategy to improve the situation or compels the system to invoke one of the strategy. In this case, the teacher leaves the decision to the system. The teaching strategy agent proposes another teaching plan and the teacher support agent invokes the teaching plan. The teaching plan is as follows:

- At first, switching to the lecture content with one level down. Meanwhile inserting the following teaching elements into the lecture content:
- Asking all students to do a mini-test that is matched with the first 45 min lecture contents. The mini-test result for each student is processed online by the system.
- Raising different questions to different students by referring to each student's learning state and the min-test result. For those interactive students, they stand up to answer the question. For those non-interactive students, they are asked simple questions and they can answer questions via the terminal without standing up.
- Giving students some entertainment Java applets to work on to make them interested in what they are leaning and practicing.

It is obvious that the system keeps monitoring students' learning state. The result shows that the number of non-interactive students is decreased to a certain stable level and the student learning state is getting better by the invoked teaching strategies in the proposed system. It is believed that the whole class learning situation will be improved in a long term.

68.9 Conclusion and Remarks

This research is mainly emphasized on body-talk reaction and profile based student-centric awareness approach for supporting effective teaching. This article described built-in-agents of learning to understand students' needs from their body-talk reactions and some elements from personal profile so as to support teacher in creating adaptive yet interactive learning atmosphere and effective teaching strategies to students in different levels and needs.

This research is still in its infancy and much work are to be pursued further. For our future studies, we would like to mainly focus on the following four aspects.

- Moving the proposed strategy to comprehensive implementation
- Improving and tuning the system along with real classroom experiments and analysis of the experiments
- Defining and extracting more facial expression features instead of using only four feature templates,
- Defining and extracting various body movement features instead of using only sitting posture,
- Taking into account of more kinds of student individual differences instead of using non verbal and personality

References

1. Simpson AW, Erickson MT (1983) Teachers' verbal and nonverbal communication patterns as a function of teacher race, student gender, and student race. *Am Educational Res J* 20(2):183–198
2. Ruth BC, Saba A-N, Shahrzad M (2004) The role of gesture in Bilingual education: does Gesture enhance learning? *Int J Bilingual Edu Bilingualism* 7(4):303–319
3. Jolly S (2000) Understanding body language: Birdwhistell's theory of kinesics. *Corp Comm Int J* 5(3):133–139
4. Philip G (2007) Five factors for effective teaching. *New Zealand J Teach Work* 4(2):89–98
5. Verdejo MF, Celorrio C, Lorenzo E, Sastre T (2006) An educational networking infrastructure supporting ubiquitous learning for school students. In: *Advanced learning technologies, 2006. Sixth international conference on, 2006*, pp 174–178
6. Huang R, Ma J, Jin Q (2010) An agent based approach for promoting interactive teaching and active learning. In: *Information technology based higher education and training (ITHET), 2010 9th international conference on, 2010*, pp 349–354
7. El-Bishouty MM, Ogata H, Yano Y (2008) A model of personalized collaborative computer support ubiquitous learning environment. In: *Advanced learning technologies, 2008. ICALT'08. Eighth IEEE international conference on, 2008*, pp 97–101
8. Scheele NK (2006) The interactive lecture: a new teaching paradigm based on pervasive computing. University of Mannheim, Germany
9. Kopf S, Effelsberg W (2007) New teaching and learning technologies for interactive lectures. *Adv Technol Learn* 4:60–67

10. Herreid CF (2006) 'Clicker' cases: introducing case study teaching into large classrooms. *J College Sci Teach* 36(2):43–47
11. Kajita S, Mase K (2006) uClassroom: expanding awareness in classroom to ubiquitous teaching and learning. In: Fourth IEEE international workshop on wireless, mobile and ubiquitous technology in education, 2006. WMUTE'06, 2006, pp 161–163
12. Ku OY, Huang OW, Chan T-W (2008) Teacher monitoring system in one-to-one self-paced learning classroom. In: Fifth IEEE international conference on wireless, mobile, and ubiquitous technology in education, 2008. WMUTE 2008, pp 196–198
13. Roblyer MD, Wiencke WR (2003) Design and use of a rubric to assess and encourage interactive qualities in distance courses. *Am J Distance Edu* 17(2):77–98
14. Lin K-C, Lin R-W, Chen S-J, You C-R, Chai J-L (2010) The classroom response system based on affective computing, in 2010 3rd IEEE international conference on ubi-media computing (U-Media), 2010, pp 190–197
15. Kasiran Z, Ibrahim Z, Yahya S (2008) Facial expression recognition as an implicit customers feedback, *Advances in human-computer interaction*, Oct 2008
16. Ekman P, Friesen WV, Hager JC (2011) Facial action coding system—the manual. [Online]. Available: <http://face-and-emotion.com/dataface/facs/manual/TitlePage.html>. Accessed: 03-Jul-2011
17. Bradski G, Kaehler A (2008) *Learning OpenCV: computer vision with the OpenCV library*, 1st edn. O'Reilly Media, 2008
18. Turk M, Pentland A (1991) Eigenfaces for recognition. *J Cognit Neurosci* 3(1):71–86
19. Kobayashi H, Hara F (1992) Recognition of Six basic facial expression and their strength byneural network. In: IEEE international workshop on robot and human communication, Proceedings, 1992 pp 381–386
20. Tammy SG (2007) Language learning beyond words: incorporating body language into classroom activities. *J Reflect Engl Lang Teach* 6(1):51–64
21. Stires L (1980) Classroom seating location, student grades, and attitudes. *Environ Behav* 12(2):241–254
22. Neff M, Wang Y, Abbott R, Walker M (2010) Evaluating the effect of gesture and language on personality perception in conversational agents. In: Proceedings of the 10th international conference on intelligent virtual agents. Heidelberg, Berlin, pp 222–235
23. Classics in the history of psychology—Jung (1921/1923) Chapter 10. [Online]. Available: <http://psychclassics.yorku.ca/Jung/types.htm>. Accessed: 25-Jul-2011
24. Microsoft Kinect SDK for developers | develop for the Kinect | Kinect for windows. [Online]. Available: <http://kinectforwindows.org/>. Accessed: 14-Nov-2011

Chapter 69

Interactive Mind Map Desktop Widget: A Proposed Concept

Tan Wei Xuan, Shakirah Mohd Taib and Saipunidzam Mahamad

Abstract Electronic mail (Email) system has become a popular method of communication in sharing and growing the knowledge. Selected documents that are shared through email attachments will be downloaded to the personal storage of the receivers. In order to ease the searching of the retrieved files, a basic taxonomy (classification in hierarchical structure) has been used as a common method to organize the storage directories. However, the directory classification need to be improved because by referring to a directory's name only, it is not enough to briefly understand the content of the files. This paper proposes a mind map of tags concept to be an additional element to enhance the documents classification and retrieval. The mind map is applied as a classification widget on a desktop that represents links of documents in the respective cloud of tags. The method and the proposed framework are discussed.

T. W. Xuan (✉) · S. M. Taib · S. Mahamad
Department of Computer and Information Sciences, Faculty of Science
and Information Technology, Universiti Teknologi Petronas, Bandar Seri Iskandar,
31550, Tronoh Perak, Malaysia
e-mail: tweixuan@gmail.com

S. M. Taib
e-mail: shakita@petronas.com

S. Mahamad
e-mail: saipunidzam_mahamad@petronas.com

69.1 Introduction

Within a community, a company, communication is vital: to exchange and share information, to store, to retrieve, to collect, to gather, to process, to distribute information. A decade ago, offices used to be full of paper files, binders, storage boxes: it contained the information accumulated project after project. The digital era get rid of the heavy and space consuming paper material. Gigabytes of digital files have replaced many paper files. However, the same problem remains; files labeling, files storage and retrieval [1]. The digital information management becomes very complex, as there are several ways to distribute and store information such as files, emails and web links exchanges. The more information that people receives, the more difficult it becomes to handle; read and understand the content, store and retrieve for sharing, distributing or processing.

People are used to similar tools, which contribute in daily tasks. The technique, method and process are still the same for decades that cause the growth in email box as well as data storage. As time goes by, people has discovered that the world is wider so as the knowledge. Thus, the information received is not managed in an effective way.

With the great importance of email, it is not surprising that it should receive attention in the literature exploring the filing habits of users. According to Fisher et al. [2], email was now used for a variety of functions, for which it was not originally intended and that this produced what they called email overload with the inbox acting as a task manager. The volume of email message received and the unintended task manager function made the process of filing either difficult or unrewarding for users. It has identified three principal ways in which people handled email overload depending on whether they use folders and whether they clean their inboxes daily.

Given a scenario where a user distributed a piece of information to the entire community in an organization, approximately has 300 people. The user sent five different emails, each of them containing a single attachment of minor size. The email series sent cost a lot to the overall community, on top of the digital resources used. Assume that it takes 14 s to extract a single file from one email. Thus, it takes 70 s for five files and approximately 5.8 h spent by 300 people to reach that email. Furthermore, If the average salary within the community is estimated at RM 8,000/month with 21 working days per month, the typical cost for the company, for simply extracting the files (not even reading through them) is RM 276 (about USD 78). The example, though based on a real case, is however certainly inaccurate as many employees might not have extracted the files. However, even with a ratio of 50 %, the figure remains high (USD 39). Thus, with proper mechanism such interactive mind map is needed to cater the above scenario.

Meanwhile, Janssen and Poot [3] found that the following three largest clusters of information-overloaded incidents are related to emails system:

- Ambiguous email—the email has uncertain content or action that requires more time to extract relevant messages and actions.

- Email cascades & multiple receipt of the same message via different people.
- Email workload—the size of email is huge. Thus, it needs more time and effort for processing.

Hence, it is believed that a proper mechanism would need to be built to cater above problem. This paper presents the analysis on mind map, as desktop widget with document classification features will improve communication standard within a community. It integrates with tag cloud to sort document and file received on email as medium to assist users on presenting digital information in a better way. This will indirectly enhance personal management of digital contents and information flows as well as reduce time consuming in sorting the right files for the right category. The Interactive Mind Map Desktop application will be able to arrange files received from e-mail to the respective topic types based on the tags provided together with the files. Users define own topic types and the topics may be classified into projects, notes, articles, contacts and more.

69.2 Concept and Theory

69.2.1 Email Management

Today, people are relying on email to send and to receive information. Email plays an important role of communication among knowledge workers [4, 5]. However, managing email is not an easy job where there are plenty of emails messages received in a day. Retrieving an attachment becomes difficult, and managing email box to meet quotas is the next burden has been faced, and the users are forced to free space to be able to send emails [6]. Emails were intended to make lives easier, but sometimes it just does the opposite [3]. Moreover, while the number of emails grows we are still using the same techniques.

69.2.2 Mind Map Application

Knowledge is being organized so as to facilitate understanding and problem solving ability. A concept of map organizes knowledge into categories and sub-categories so that it can be easily been remembered and retrieved. The hierarchical structure of a mind map conforms to the general assumption that the cognitive representation of knowledge is hierarchically structured [7]. Mind Mapping is used in several areas such as educations [8, 9], presentation, information retrieval [10, 11]. On a technical level, mind map can be collapsed or expanded to quickly increase or decrease the level of detail. It can be filtered based on priorities, keywords and colors to identify each of the clouds under the mind map. Mind map can be enriched by texts, graphics,

spreadsheet info, links to files, websites as well as Really Simple Syndication (RSS) feeds.

According to Buzan [12], the concept of mind map can be applied in various aspects of life to improve learning and thinking that will enhance the capability of an individual. In order to create a good mind map, several aspects need to be clearly identified. Each of the branches is connected to the central image and keyword for each topic is clearly defined. People from various industries have been starting to use mind map in their daily work, as they understand the power of mind map, which can enhance the productivity.

69.2.3 File Organization on Desktop and Tags

In the spring of 1993, Barreau conducted a study of seven managers to observe how they organized and retrieved information from their electronic workplace [13]. The goal of the research was to identify the types of documents used and to determine the factors affecting individual decisions to acquire, organize, maintain and retrieve information. People are using the location-based finding to find files. The user takes a guess at the directory or folder where she thinks a file might be located, goes to that location and then browses the list of files in the location until she finds the file she is looking for. The process is iterated as needed.

Documents received from various sources are hard to be identified. Somehow, people are facing difficulties to locate their files into the correct folders. The more information received, the more difficult it becomes to handle them: read and understand the content, store them and later on retrieve them for sharing, distributing or processing. Helping computer users rapidly locate files in their folder hierarchies has become an important research topic in today's intelligent user interface design [14].

Tagging is fast becoming one of the primary ways people organize and manage digital information [15]. Tags have been used widely for the organizational tools such as folders and search on user's desktops as well as on the web. The development of tagging has become a broad implications for information management, information architecture and interface design. By using tags, browsing and finding information will certainly be more productive. Websites such as Delicious and Flickr is a good example that has been using tags to indicate the information. The keywords, which are referred to as tags on the site, allow users to describe and organize their content that become a new term in the technology which have been greatly enhance the way people find information in a web.

Other example is DeepaMehta. It is an open source semantic desktop application based on the topic maps standard. The conceptualization and innovative graph based user interface have been guided through findings in cognitive psychology in order to provide a cognitively adequate working environment [16]. The system aims to evolve desktop applications into an integrated workspace enabling the user to organize, describe and relate information objects like text

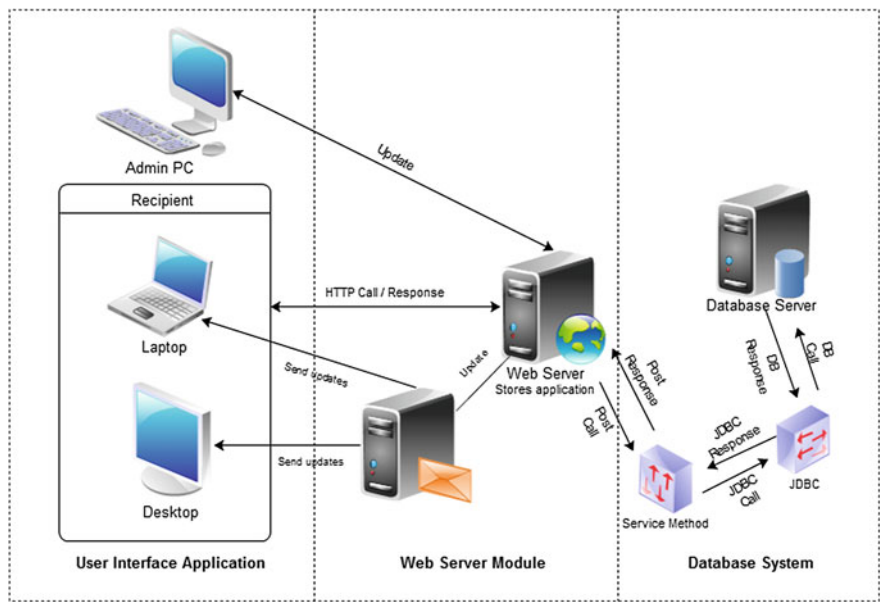


Fig. 69.1 Interactive mind map desktop application with tags framework

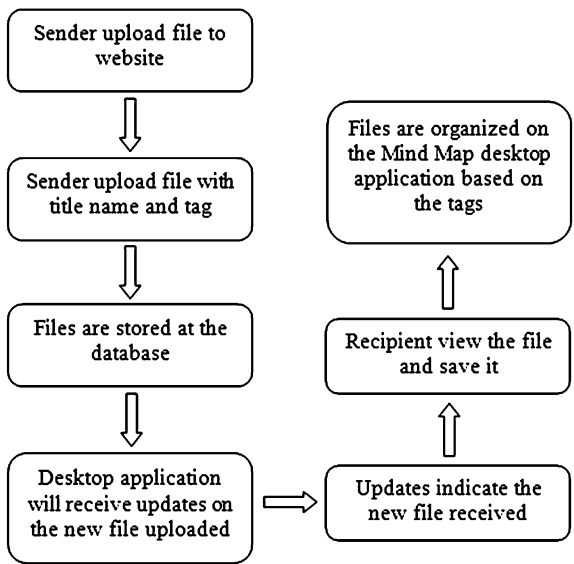
notes, external documents and media [17]. From the concept of the semantic desktop, the node and link type of DeepaMehta’s visualization is stimulated by the concept mapping approach. There has been research on a semi-formalized derivative of concept mapping, call knowledge mapping, that uses fixed sets of typed relations.

69.3 System Design Discussion

The proposed concept of the system is shown in Fig. 69.1. The design uses document classification widget based on the concept of mind map on desktop to provide an alternative way in sorting the email attachment files based on the topics. The sender can upload the files through the website. All the data will be uploaded into a central database. The JDBC (Java Database Connectivity) connects SQL databases with the java application (widget) which is the mind map desktop application. Thus, the recipient will receive updates from the mind map desktop application and view the files on desktop application based on the tags.

The idea of mind map as background, include “clouds” represent topics and are described by tags. The user can also set tags to the file and it will be displayed on the mind map. In windows environment, organizing file geographically on the desktop might be scattered by windows whenever the screen resolution changes or a crash occurs. Nevertheless, an application document classification is very useful

Fig. 69.2 Overview of the proposed system flow



in organizing files in a reliable and smart way. The concept of this mind map desktop creates a collaborative environment whereby a user may send files to other users based on the tags. Tags describe in brief the file without the need to open and read the contents.

The overview of the system flow is shown in Fig. 69.2. The more information received, the more difficult it becomes to handle them. However, the mind map desktop application will come to a solution whereby users able to store the files according to the respective tags group for easy retrieval in future. After choosing the right file to be sent to the recipient, the sender will need to name the title and select the right tag for the file. This is to ensure that the recipient will know the content of the file by the tag selection. Besides, the tag also indicates the file to be stored in which category in the mind map desktop application at the recipient's place.

The mind map based on tags is developed using an open source Java platform. Figure 69.3 shows an example of mapping after several files had been uploaded to the server. The title of the files is being sorted according to the respective categories and tags. Specific color scheme will be used to indicate the new entry of files to the desktop application. Recipients can easily differentiate the new files and when a click action is being done, the file will be opened. Recipient can also retrieve and save the file if necessary.

Fig. 69.3 Mind map desktop application with tags

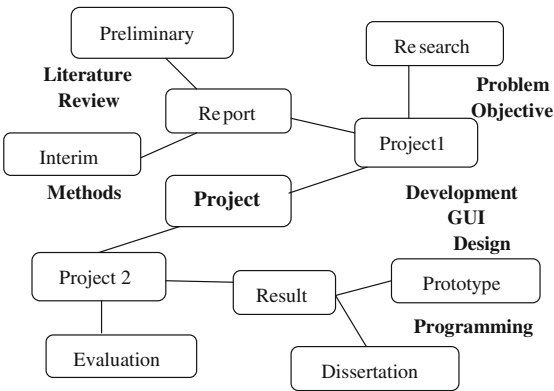
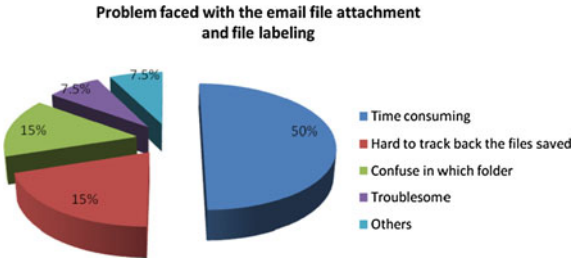


Fig. 69.4 Interactive mind map desktop application with tags framework



69.4 Discussion

A data gathering was conducted with 200 participants in an organization, were participated in the questionnaires. Its aims to gather the behavior of the participant in handling email with attachment. Figure 69.4 show the number of respondents towards the problem faced with the email files attachment and file labeling. Fifty percentage of the respondents feel that downloading the email with file attachment and label the files are time consuming thus reducing the productivity of doing their tasks in the office. The other two of 15 % respondents shows that it is hard to track back the files saved and it will be confusing for them to know which file to be save in which folder. Meanwhile, 35 % respondents found it is quite troublesome and other problems stated were too many files to be saved and read, thus creating overflow of files.

The participant mostly received three or more email messages per day and about 2 % of them are able sort the files received accordingly and able to track back their files because they received very few emails with file attachments per week. The participants were agreed with the concept of document classification widget that would help them to organize the files effectively. However, some of participant felt that the concept might be hard for them to learn in order to use the tool. However, most of them believe that a document classification widget is

essential in order for them to work productively without to bother on organizing the files on their desktop.

Mind Map application is a platform for knowledge management. Knowledge is represented in a semantic network and is handled collaboratively. It combines interdisciplinary research with the idea of Open Source to generate a true benefit for workflow as well as for social processes. There are many different ways people are organizing document, to locate files and folders. Document classifications seem to be important in order to have an easy track back whenever users need to use the files in an effective way. The use of tags is useful to display metadata about an item. Tags used as a short term to describe the specific information or files in this case in order for users to easy identify the content of the files without having to open the files needed. The integration of these concepts as desktop application will provide a virtual tools, which could help in managing a lot of document and information. The more information received, the more difficult it becomes to handle them. However, the Mind Map desktop application will come to the alternative solution whereby users able to store the files according to the respective tags group for easy retrieval in future.

69.5 Conclusion

Data and information overloading becomes an issue that emerged a rapid advances in computer and telecommunication technology. People can share knowledge and collaborate through various medium of communication. Given the great importance of email, it is not surprising that it could receive attention in the literature exploring the retrieving habits of users. The volume of email received and the unintended task manager function made the process of filing either difficult or unrewarding for users. The proposed concept applies the cognitive knowledge representation in files retrieval desktop widget. Another strategy has been of sharing files explored to replace the sharing practice through email attachments. It seems possible to adopt this concept in any organization to increase cost and time efficiency.

References

1. Khoo CSG, Luyt B, Ee C, Osman J, Lim H-H, Yong S (2007) How users organize electronic files on their workstations in the office environment: a preliminary study of personal information organization behaviour. *Inform Res* 11(2):293
2. Fisher D, Bush AJ, Gleave E, Smith MA (2006) Revisiting whittaker & sidner's "email overload" ten years later, microsoft research
3. Janssen R, de Poot H (2006) Information overloaded: why some people seem to suffer more than others. *NordiCHI* 2006, 14–18 Oct 2006

4. ProofPoint Inc., (2004) Best practices in messaging security: managing email security policy and regulatory compliance requirements in financial services, health care and public sector, White paper, Ziff Davis Publishing Holdings
5. Bellotti V, Ducheneaut N, Howard M, Smith I (2003) Taking email to task: the design and evaluation of a task management centered email tool. Proceedings of the CHI 2003 conference on human factors in computing systems, ACM, New York
6. Bergman O, Tucker S, Beyth-Marom R, Cutrell E, Whittaker S (2009) It's not that important: demoting personal information of low subjective importance using grayarea, Proceedings of the 27th international conference on human factors in computing systems
7. Brinkmann A (2003) Graphical knowledge display—mind mapping and concept mapping as efficient tools, Education Review, University of Duisburg, Germany
8. Chen J (2008) The using of mind map in concept design computer-aided industrial design and conceptual design, 9th international conference on CAID/CD 2008, pp 1034–1037
9. Lin C-C, Shih D-H (2009) Mind mapping: a creative development in industrial engineering education, wireless communications, networking and mobile computing, 2009. 5th international conference on WiCom '09
10. Beel J, Gipp B, Stiller J-O (2009) Information retrieval on mind maps—what could it be good for? 5th international conference on collaborative computing: networking, applications and worksharing, 2009, Washington
11. Zualkernan IA, AbuJayyab MA, Ghanam YA (2006) Sixth international conference on advanced learning technologies, 2006, Kerkrade
12. Buzan T (2008) Mind map application,
13. Barreau D, Nardi BA (1995) Finding and reminding—file organization from the desktop
14. Bao X, Herlocker JL, Dietterich TG (2006) Fewer clicks and less frustration: reducing the cost of reaching the right folder. Proceedings of the 11th international conference on intelligent user interfaces, ACM, p 185
15. Smith G (2007) Tagging: people—powered metadata for the social web, New Riders
16. Richter J, Völkel M, Haller H (2005) Deepamehta—a semantic desktop. (2005) In: Decker S, Park J, Quan D, Sauermann L (eds) Proceedings of the 1st workshop on the semantic desktop. 4th international semantic web conference, vol. 175, Galway, Ireland
17. Staps M, Richter J (2000) DeepaMehta—a semantic desktop. University of Karlsruhe, Germany

Chapter 70

An Algorithm for Replication in Distributed Databases

Adrian Runceanu and Marian Popescu

Abstract Distributed databases are an optimal solution to manage important amounts of data. We present an algorithm for replication in distributed databases, in which we improve queries. This algorithm can move information between databases, replicating pieces of data through databases.

70.1 Distributed Databases

70.1.1 *Distributed System*

We consider a distributed system composed of a set of database sites S which are fully connected. The sites communicate through message passing. The system is asynchronous because there is no bound on process relative speeds, clock drifts, or communication delays. Sites can only fail by crashing and we do not rely on site recovery for correctness. However, we assume that when a site recovers, it does so with the state that it had before the failure. Furthermore, we assume that our asynchronous model is augmented with a failure detector Oracle so that, consensus is solvable [10].

A. Runceanu · M. Popescu (✉)
University Constantin Brâncuși of Târgu-Jiu, Targu-Jiu, Romania
e-mail: marian@utgjiu.ro

A. Runceanu
e-mail: adrian_r@utgjiu.ro

70.1.2 Distributed Database

We consider a distributed relational database as a relational database whose relations are distributed, i.e. fragmented, among the set S of database sites. This distributed database is given by $DDB \subseteq DB \times S$.

The relations of the database can be fragmented horizontally, using a selection operation from the relational algebra, or vertically, using a projection operation. To avoid semantic changes to the relational database as a consequence of the fragmentation process, the following properties must be enforced [9], where the function $frags(R)$ gives the fragments of a relation R :

Completeness. The fragmentation cannot generate any loss of information:

$$R = \cup R_i, \forall R_i \in frags(R)$$

Reconstruction. It must be possible by using a relational algebra operation ∇ to rebuild the original relation R as follows:

$$R = \nabla R_i, \forall R_i \in frags(R)$$

This operation is a union in case of horizontal fragmentation and a join in case of vertical fragmentation.

Disjointness. If a relation R is horizontally fragmented, then every two distinct fragments cannot have a single common tuple

$$\forall R_i, R_j \in frags(R), R_i \cap R_j = \phi \quad \text{where } i \neq j$$

If a relation R is vertically fragmented, then every two fragments must have the same keys,

$$i.e. \forall R_i, R_j \in frags(R)$$

$$R_i \cap R_j = \{set \text{ of primary key attributes of } R\}$$

where $i \neq j$

An important assumption we make is that for every fragment of a relation there is a correct site that replicates it.

70.2 Data Fragmentation

We propose one software application which integrates different management tools, communication, evaluation, monitoring etc. together on a common platform. The aim is to provide technological support for teachers and students to optimize the phases of the teaching and learning process through e-learning/e-work.

Based on the above requirements, it is necessary to design an application that can improve performance of data access. As we know, every database is built to

provide users high availability of data, however, a major cost in executing queries in a distributed database is the data transfer cost incurred in transferring multiple database objects (fragments).

70.2.1 Fragmentation

The objective of fragmentation is to determine fragments to be allocated on different sites so to be minimized the total data transfer cost on executing a set of queries [2–5].

Three kinds of fragmentation can be applied to a relation in a distributed database: Vertical Fragmentation, Horizontal Fragmentation and Mixed Fragmentation.

70.2.1.1 Vertical Fragmentation

Vertical fragments are created by dividing a global relation \mathbf{R} on its attributes by applying the project operator:

$$R^j = \prod_{\{A_j\}, key} R, \quad \text{where } 1 \leq j \leq m \quad (1.1)$$

where $\{A_j\}$ is a set of attributes not in the primary key, upon which the vertical fragment is defined and m is the maximum number of fragments.

A vertical fragmentation schema is complete when every attribute in the original global relation can be found in some vertical fragment defined on that relation. Then the reconstruction rule is satisfied by a join on the primary key(s):

$$\forall R^j \in \{R^1, R^2, \dots, R^m\} : R = \bowtie_{key} R^j \quad (1.2)$$

The disjoint ness rule does not apply in a strict sense to vertical fragmentation as the reconstruction rule can only be satisfied when the primary key is included in each fragment. So excluding the primary key, no data item should occur in more than one vertical fragment [1, 6].

70.2.1.2 Horizontal Fragmentation

Horizontal fragmentation divides a global relation \mathbf{R} on its tuples by using the selection operator:

$$R^j = \sigma_{P_j}(R), \quad \text{where } 1 \leq j \leq m \quad (1.3)$$

Where P_j is the selection condition as a simple predicate and m is the maximum number of fragments.

The horizontal fragmentation schema satisfies the completeness rule if the selection predicates are complete. Furthermore, if a horizontal fragmentation schema is complete, the reconstruction rule is satisfied by a union operation over all the fragments:

$$\forall R^j \in \{R^1, R^2, \dots, R^m\} : R = \cup R^j \quad (1.4)$$

Finally, disjointness is ensured when the selection predicates defining the fragments are mutually exclusive [1, 6].

The derived horizontal fragmentation occurs when a member relation inherits the horizontal fragmentation of its owner. If the completeness and disjointness rules are satisfied for the owner fragments, they are intrinsically satisfied for the child fragments. The global relation can be reconstructed by the application of the union operator, as for primary horizontal fragmentation [7]. In this paper [8] is presented a general approach of the data fragmentation in a distributed database. Using this tool for obtained the best partitioning scheme with implementation of several classic algorithms is one solution in designing phase of a distributed database.

70.2.1.3 Mixed Fragmentation

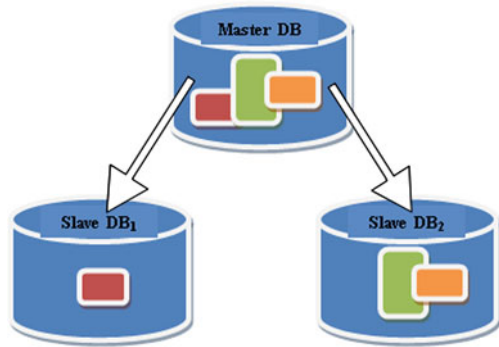
An hybrid fragmentation schema is a combination of horizontal and vertical fragments. If the correctness and disjointness rules are satisfied for the comprising fragments, they are implicitly satisfied for the entire hybrid schema. The reconstruction is achieved by applying the reconstruction operators in reverse order of fragment definition. That is, if a global relation underwent horizontal fragmentation followed by vertical fragmentation, the reconstruction would consist of a join followed by a union [7, 8].

70.2.2 Replication

Assuming that the database is fragmented, we have to decide what data will be allocated on different fragments of the network stations. When data are allocated, they are either replicated or kept in a single copy. The reasons for reliability and efficiency replication are for read-only queries. If there are multiple children, even if the system is falling, some children are likely to be accessible, so no data is lost. In addition, read-only queries that access the same items can be executed in parallel because there are children on several stations.

On the other hand, the execution of queries that are updates may cause problems because the system must ensure that all copies are updated correctly. Therefore, the decision on replication is a compromise that depends on the ratio of read-only queries and queries perform updates. This decision affects almost all control functions and almost all algorithms DDBMS. Creating multiple copies of the same information is justified in economic terms only if the next inequation is satisfied:

Fig. 70.1 Structure of replication schema



$$CMC + \sum_{i=1}^p CAL_i + CA < \sum_{i=1}^p CAD_i \quad (1.5)$$

where the parameters involved have the following meaning:

CMC—the cost of memory copy

CAL_i—local access to the user cost

CA—the cost of updating the copy

CAD_i—the cost of remote access to the primary copy to the user

p—number of users

Multiple copies is the responsibility of management's DDBMS. The updates can be performed either simultaneously or on secondary copies and they assume a delay cost.

70.3 Our Solution

We developed an application that uses horizontal fragmentation with partial replication. Figure 70.1 shows the structure used for replication in the system developed. Obviously, this solution is punctual; it means that is built for a special database schema. However, the algorithm used for distribution and also to keep consistency can be applied to others applications.

In this algorithm (MS—Master/Slave) we verify each query and keep a counter to know the place and frequency of the user's access, if the system detects an access pattern, then it copies a set of records into a slave database. This approach allows improving the performance of queries around database. Although this solution seems very simple, works fine with the requirements. It is necessary that every computer must have a slave database for executing the next algorithm:

Algorithm MS (Const Value)

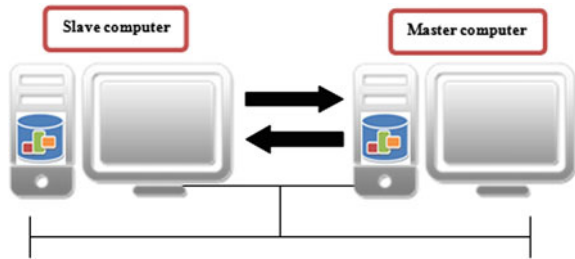
Input: a number of local database computers

Output: replications records at each local database

Fig. 70.2 Query on local database



Fig. 70.3 Sending query from slave database to master database



Begin

```
//local variable - counter = 0
for each requested query do
  counter = counter + 1//at slave computer
end_for
if counter = value then
  Request the set of records that the user is asking for and save this information
  into the slave database
  counter = 0
end_if
```

End.

In order to keep consistency into the distributed database it is necessary that both master and slave computers have the same information, but in this case, since the user will access their own information from the slave computer, it is not necessary to copy the information immediately, but later. This schema of information management allows database availability even when the connection between slave and master database is broken.

Another algorithm (Slave/Master Search) has been designed to provide access to local database before sending the query to the master computer. This algorithm tests for the information into the slave database (Fig. 70.2), if such records are founded, the information is given to the user, otherwise the query is sent to the master database as shown in Fig. 70.3.

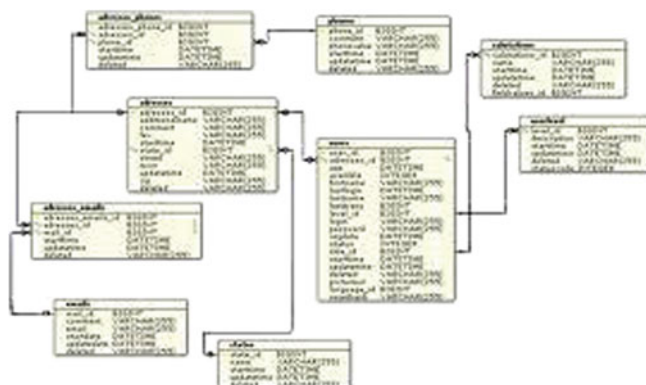
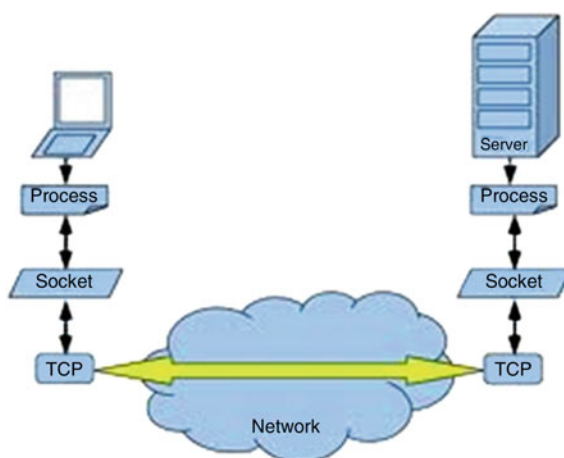


Fig. 70.4 Communication with sockets using TCP/IP protocol

Fig. 70.5 Database schema used for the application



Algorithm SMS

Input: local database computer

Output: master database computer

Begin

for each local requested query **do**

if requested query is found **then**

send to the local user

else

send to the master database computer

end if

end for

End.

In this case, all the information can be found at master database, but the system can detect when a user has an access pattern and it copies the information needed to a slave computer (partial replication), so the users have their information nearest, decreasing the search time. For the application developed, we used communication with sockets based on client–server model (Fig. 70.4).

We designed a basic database schema that is used to manage all the informations about users of this application. This schema is shown in Fig. 70.5.

The schema shown in Fig. 70.5 is only a subschema of the entire application. This because the real schema has around 45 tables and many fields.

70.4 Conclusions

We have presented an algorithm for a punctual solution of our application. This algorithm can move information between databases, replicating pieces of data through slave databases. The algorithm is based on two techniques: Master/Slave, to provide fast access to database on user queries and Slave/Master Search (two computers) in order to get availability.

We tested the algorithm in a database laboratory with 40 computers, and it seems to be a solution for the problem established by the university, since the results obtained in a local computer network have a good behavior.

References

1. Bellatreche L, Karlapalem K, Simonet A (1997) Horizontal class partitioning in object-oriented databases. In: Proceedings of the 8th international conference on database and expert systems applications (DEXA'97) Sept 1997 (Lecture Notes in Computer Science), vol 1308, pp 58–67
2. Ceri S, Negri M, Pelagatti G (1992) Horizontal data partitioning in database design. In: Proceedings of the ACM SIGMOD international conference on management of data, pp 128–136
3. Chakravarthy S, Muthuraj R, Varadarajan R, Navathe S (1994) An objective function for vertically partitioning relations in distributed databases and its analysis. Distributed and parallel databases. Kluwer Academic Publishers, New York, pp 183–207
4. Ezeife CI, Barker K (1995) A comprehensive approach to horizontal class fragmentation in distributed object based system. *Intl J Distrib Parallel Databases* 3(3):247–272
5. Navathe SB, Karlapalem K, Ra M (1995) A mixed partitioning methodology for distributed database design. *J Comp Softw Eng* 3(4):395–426
6. Navathe SB, Ra M (1989) Vertical partitioning for database design: a graphical algorithm. In: Proceeding of the international conference on management of data, ACM-SIGMOD'89, pp 440–450
7. Runceanu A (2007) Fragmentation in distributed databases. International joint conferences on computer, information, and systems sciences, and engineering (CISSE 2007) conference. Conference proceedings book, Dec 3–12, 2007, University of Brigeport, USA, publish in Innovations and advanced techniques in systems, computing sciences and software

- engineering, Springer Science. ISBN:978-1-4020-8734-9 (Print) 978-1-4020-8735-6 (Online). doi:[10.1007/978-1-4020-8735-6_12](https://doi.org/10.1007/978-1-4020-8735-6_12)
8. Runceanu A, Popescu M (2009) Eval tool for evaluate the partitioning scheme of distributed databases. International joint conferences on computer, information, and systems sciences, and engineering (CISSE 2009) conference. Conference proceedings book, Dec 4–12, 2009, University of Brigeport, USA, vol 1—Technological developments in networking, education and automation (Elleithy, Sobh, Iskander, Karim, Mahmood), SCSS 09. Published in Volume 1—CISSE 2009 Proceedings: technological developments in networking, education and automation. ISBN:978-90-481-9150-5
 9. Tamer O, Valduriez P (1999) Principles of distributed database systems, 2nd edn. Prentice Hall, Englewood Cliffs, p 07362
 10. Tushar DC, Toueg S (1996) Unreliable failure detectors for reliable distributed systems. J ACM 43(2):225–267

Chapter 71

General Dispatching of Lignite Mining Pit

Constantin Cercel and Florin Grofu

Abstract High automation level of mining pit machinery, they dispersion, high number of information required to control the whole technological process, are just few elements that highlight the complexity of activities from a lignite mining pit equipped with excavation machinery, continuous transport and dumping machinery. A good working of all implied machineries in technological process from a mining pit is provided by remote control from distance after some rules required by the technological, technical and working security restrictions control realized by a mining pit dispatcher [Popescu L, Cristinel R, Florin G (2006) Numeric system for energetic control feasible in NSLO pits. Revista Minelor No.5, ISSN 1220-2053]. This paper proposes a dispatching system on National Society of Lignite Oltenia level, which may integrate the whole technological process from each mining pit in Oltenia.

71.1 Introduction

National Society of Lignite Oltenia (NSLO) is a Romanian national society which has as main activity coal exploitation in Oltenia area. In this area are opened 19 mining pits.

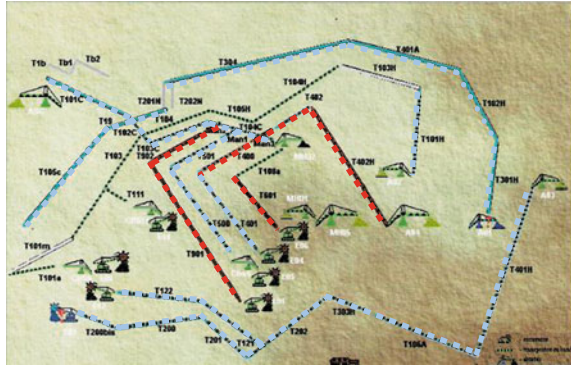
First dispatching in NSLO consisted in synoptic panels for technological flow with optical indication of the machineries status (working or stationary) and

C. Cercel (✉) · F. Grofu (✉)

Constantin Brancusi University, Calea Eroilor Street, No. 30, Tirgu Jiu, Romania
e-mail: costi@utgjiu.ro

F. Grofu
e-mail: florin@utgjiu.ro

Fig. 71.1 Synoptic diagram for the machineries



indicates also maximum three important defects. These allow the start and stop command of transporting belt circuits using radio or phone communications with local operator. Mining pit dispatch records manually working diagram for each equipment. All main links were made by cable [2–6].

New technical progress in informational domain made possible their use in mining, radio or optical transmission of commands and information, program-mable logical controllers, electronic transducers and process computers are just few examples in this way.

In general terms, mining pit dispatching includes two types of dispatching: technological and energy. An important factor in normal and optimal working of all technological flows is the weather parameters, so weather monitoring system is a new element that should be included in dispatching system.

71.1.1 Technological Dispatching

Technological dispatching of a mining pit must meet the follow requirements: monitoring machineries status and parameters, the status of technological and transport lines, indicating their information in a synoptic diagram as in Fig. 71.1.

Technological dispatching implies technological parameters monitoring (excavated mass, transported mass of coal and barren), operator's alerts in case parameters overcome the reference values. Modern dispatching creates databases with: number of working hours for each machinery/equipment to establish some maintenance cycles, number of stationary hours for each machinery/equipment, number of stationary hour by causes, quantities of excavated coal, quantities of barren excavated, quantities of coal transported, evaluation of coal stock in coal deposit at different time, quantities of coal delivered.

71.1.2 Energy Dispatching

Energy dispatching is about energy parameters measurement (current, voltage, active power, reactive power) on each equipment (excavator, belt transporter, dump

machinery) and send these data to mining pit dispatch, and receiving of commands for technological flow to ensure framing energy consumption in default values.

71.1.3 Weather Dispatching

If we analyze the influence of climatic parameters can see that a thorough knowledge of these can increase productivity and avoid unwanted situations. For example real-time insight into the outdoor temperature can result in negative temperatures at the start of anti-frost protection program to protect machinery from coal pits.

Also in the case of coal stored in piles, it is known that this passes through oxidation at low temperature, in presence of methane and other volatile material on its surface. This exothermic oxidation increases coal temperature and if the heating temperature is not removed, there is a stage to coal on fire. This phenomenon is called spontaneous combustion and may be amplified by certain conditions of temperature and humidity of the environment.

Another situation where it is necessary to know the climate is related to storm water discharges from the pit. In pit-coal from Oltenia special problems occur due to large quantities of water from both infiltration and precipitation. To evacuate such large amounts of water pumps units are used, whose number and capacity depends on the pit area and estimated quantity of water shall be discharged from this pit. The power consumption of these pump units is very high, about 10 % of total electricity consumption in pit.

To reduce electricity consumption, it is necessary first of all to correlate the power parameters (input current, power factor $\cos \varphi$, etc.), characterizing the electric motors acting pumps, with the technological parameters such as water level from collection basin, warning level of the water collection basin and the amount of precipitation. In case of unstarting on time a corresponding number of pumps, depending on the amount of precipitation, can lead to flooding such as that in Fig. 71.2 which may disrupt the smooth running of the technological process.

Other climatic parameters can lead to extreme situation in equipment functioning taking into account the big dimensions on it. In Fig. 71.3 it is presented an excavator used in coal exploitation.

In case of so big dimensions machinery the very important weather parameters are the wind speed and the wind direction. The operator must take into account these values to position the machinery to avoid dangerous situations.

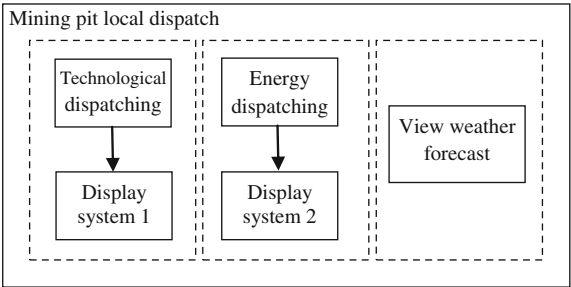
Fig. 71.2 Flooded pit



Fig. 71.3 Coal excavator



Fig. 71.4 Structure of actual dispatching system



71.2 Dispatching System Architecture

In present, in NSLO mining pits were implemented technological and energy dispatching systems for almost all used equipment. On the local dispatch center exists a weather forecast system, for weather evolution estimation in time.

Functioning structure for local dispatch (for a pit) is presented in Fig. 71.4.

As seen in Fig. 71.4, the dispatching system has three distinct components, without any interactions and use three distinct display systems:

- technological dispatching provides information about equipment status and about technological and transport lines;
- energy dispatching offers information about energy parameters and power consumptions for each equipment;
- weather forecast informs the dispatch operator about possible weather changes in future, it will take decisions to avoid unwanted situations.

This type of system is very difficult to follow, difficult to manage, and monitored process control depends on the dispatch operator experience and vigilance.

In terms of communication with NSLO central dispatching system, these is an asynchronous type, link were made with phone systems and within own communication networks.

This paper proposes a solution for implementation of a general and complete dispatching system on NSLO level, implying to create new structures on mining pit level.

The functions proposed for the dispatching system are the follow:

- automatic acquisition of parameters from technological process, through transducers with unified or pulse signals output;
- primary processing of measurement values in local stations (scaling, averages, limits framing, alarms, monitoring, etc.);
- reference prescription, calibration, maintenance;
- storage, archiving and delivering of measurement processed on a period of time, operating reports, statistics;
- communication in data flows with other system nodes and with system architectures levels, including production databases management system;
- sending data in pyramidal system to hierarchical high level;
- monitoring energy consumptions and balanced with status parameters of the equipment's and production;
- monitoring active and reactive power an following the direction of power flow;
- local control and signalization of circuits, interrupts, interlocks, alarm currents;
- position and control of machineries in working field;
- indicating the status of machineries and equipment, excavated mass debits, transported mass debits;
- providing information about weather changes in real time, and weather parameters for each pit.

General structure of proposed system is presented in Fig. 71.5.

WEATHER block from Fig. 71.5 represents the weather monitoring system which must be implemented into dispatching system on pit level to provide information in real time about temperature, precipitations, wind speed and direction and other important weather parameters for entire technological flow.

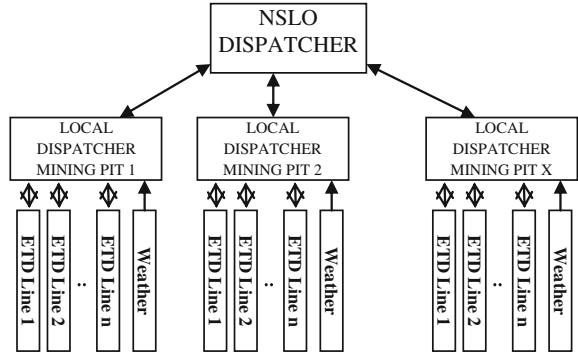


Fig. 71.5 General structure of proposed system

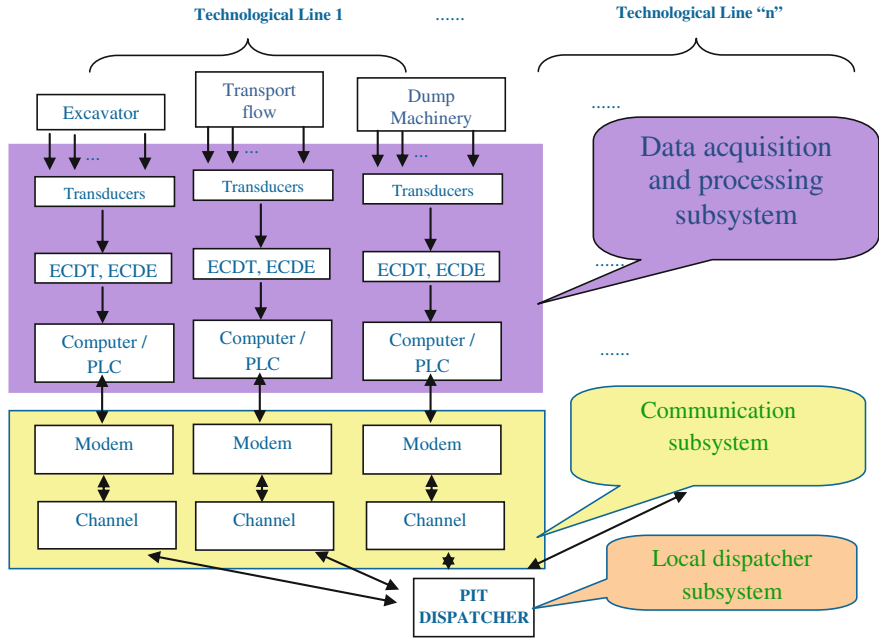


Fig. 71.6 ETD architecture

The blocks ETD Line x (Energy and Technological Dispatching for a technological Line x) represents a united monitoring system for both energy and technological parameters for a technological line. This system provides information about energy and technological values for each equipment of a technological line in a compact way.

A complete description of ETD block is presented in Fig. 71.6. Functionally it is composed by three subsystems:

- data acquisition and primary processing subsystem;

- communication subsystem;
- local dispatcher subsystem.

Data acquisition and processing subsystem include: transducers, equipment for energy and technological data acquisition, computer systems or PLC. The inputs from process are given by various command and detection elements used in a automation system: command buttons, selecting keys, limiters, transducers, protections etc.

Communication subsystem will ensure the handled of data from process through communication modems and using various communication channels (wires, radio, optical fiber, GSM etc.), to a communication server present in “local dispatcher subsystem”, and then transferred to dispatcher local network.

In case of proposed system the communication will be bidirectional, so communication subsystem must ensure data transfer from dispatcher to pointed equipment.

Local dispatcher subsystem is designed as a computer network and includes:

- communication server
- communication modems
- network server
- workstations
- printers
- displays

On this level will be implemented a dispatching application with following functions:

- viewing mining pit structure and technological flow;
- viewing equipment parameters and its position in pit;
- creating databases with specific information from system (pit structure, activities reports, indicators etc.);
- tracking of technological process indicators (working times, debits etc.);
- tracking power supply of pit (consumption);
- weather parameters monitoring;
- real time communication with central dispatcher from NSLO and delivering the information from process.

Central dispatcher will have an application which communicate in real time with local dispatchers of pits through internet or own communication networks. Through this application it will receive, as synoptic diagram, the main data from each pit dispatcher with possibilities to detail these information.

Central dispatcher will be able to send in real time energy and technological reference values for technological lines from a pit, can communicate and establish with local dispatcher operator solutions for some defects/problems of a technological line, and can establish the status of a machinery depending on technological flow parameters and weather parameters to obtain a maximum efficiency in maximum security conditions.

71.3 Conclusions

A central type monitoring of the parameters from a mining pit ensure an efficiency schedule exploitation activities. Thereby the central dispatcher can coordinate the activities from a pit to ensure all coal requirements in security conditions.

Although apparently weather parameters (temperature, rain, wind direction and speed) don't influence so much the technological processes from lignite mining pit, however a central knowledge in real time of environment parameters leads to an increasing of productivity and to taking decisions with regard to avoiding some dangerous situation in equipment exploitation.

Central dispatcher may require reduction of activities and even stop in mining pits with unfavorable working conditions simultaneously with increasing activities in mining pits with safety conditions.

On the central dispatcher level will be stored general databases with all activities for pits and system, general reports about working times and stop times, quantities of coal and barren, weather parameters. Also, it will generate maintenance schedules for each pit and equipment, according with received data.

References

1. Popescu L, Cristinel R, Florin G (2006) Numeric system for energetic control feasible in NSLO pits. *Revista Minelor* No.5, ISSN 1220-2053
2. Runceanu A, Marian P (2009) Evaltool for evaluate the partitioning scheme of distributed databases. In: *International joint conferences on computer, information, and systems sciences, and engineering (CISSE 2009) conference, Proceedings: technological developments in networking, education and automation*, ISBN: 978-90-481-9150-5
3. Grofu F, Cercel C (2011) Integrated system for weather monitoring in coal exploitation. *Annals of the Constantin Brancusi University of Targu Jiu. Engineering Series*, Issue 1/2011, pp 91–100, ISSN 1842-4856
4. Popescu LG (2008) Dispatch of mining pit's activities, necessity or globalization's effect?. *Annals of the Constantin Brancusi University of Targu Jiu. Engineering Series*, Issue 1/2008, pp 5–18, ISSN 1842-4856
5. Huidu E (1996) *Coal storing systems*. Tehnica Publishing House, Bucuresti
6. Popescu LG, Grofu F, Dinca A, Cercel C (2008) Dispecerizarea energetica si tehnologica a carierei Rosia. In: *Regional Energy Forum—FOREN 2008*, ISBN CD-ROM 1—FOREN 2008: 978973720177-5

Chapter 72

Towards Improving the StatscanTM X-Ray Image Quality through Sliding-Mode Control of the C-Arm

M. Esmail, M. Tsoeu and L. John

Abstract A method to improve image quality of the StatscanTM X-ray system is investigated. Transient errors in the trapezoidal motion profile of the scanning C-arm may cause mismatches between the detector and collimated beams from the X-ray source. This results in the partial degradation of image quality. The performances of two C-arm motion controllers were investigated using computer simulations and experiments: a standard PI and DMRISMC. The controllers are competitive at high sampling rate. However, the DMRISMC controller shows better performance motion profile tracking when slow sampling rate is used. There are strengths and weaknesses for both controllers and DMRISMC was expected to improve the image quality.

72.1 Introduction

StatscanTM is a full-body digital radiography X-ray machine developed for normal diagnostic as well as a high incidence of penetrating injuries [1, 2]. It works in a linear slit-scanning mode (LSSM) driven by a linear motor manufactured by Kollmorgen. Despite the many advantages of slit-scanning, the X-ray images may become degraded if the C-arm is not moving smoothly. This results in a mismatch between Detector CCD clocking and the C-arm linear velocity particularly during

M. Esmail (✉) · L. John

Department of Human Biology, University of Cape Town Private Bag,
Rondebosch 7701, South Africa
e-mail: esmailmoh19@gmail.com

M. Tsoeu

Department of Electrical Engineering, University of Cape Town Private Bag,
Rondebosch 7701, South Africa

trapezoidal motion profile phases. The StatscanTM image may thus be blurred for a very short section at the beginning or the end of each scan. The primary issue is the residual vibration caused by improper acceleration or deceleration during motion tracking which results in decreased precision and a longer settling time [3]. The challenge in such system is to improve motion profile tracking performance and rejection of disturbance resulting from frictions, ripple forces, uncertainties in the machine parameters, etc. [4].

In attempts to solve such problems, the literature highlights two directions of research, which have been conducted to date on precision motion control. One is motion profile planning and the second is the use of non-linear and optimal controllers. For example, practical efficient asymmetric velocity profiles were proposed by Rew, Ha and Kim [5]. This method allows the jerk magnitude to be manipulated during the deceleration phase by using a single parameter. Similarly an online smooth trajectory for industrial systems was presented by Zheng, Su and Müller [6]. By using two modified non-linear tracking differentiators in order to generate a smooth trajectory from ramp and step set point references. In addition several advanced non-linear control methods, optimal controllers and adaptive controllers have been proposed in order to deal with plant uncertainties and to improve maximum tracking accuracy [7].

The classical proportional–proportional integral cascaded controller (P-PI) with some modification has successfully been implemented in numerous motion control applications [8, 9]. However, it is not robust enough to deal with the variations caused by external disturbances and parameter changes during operation. In contrast, sliding mode control provides a robust controller [10]. It has many advantages, including insensitivity to parameter variations and model uncertainties, external disturbance rejection, fast dynamic responses and good transient performance. The one noted drawback of the sliding mode controller is the infinite-frequency magnitude oscillations around the sliding surface [11]. Such chattering has been reduced with boundary layer technique [12]. In recent years, the sliding mode control method has taken on a general design form and many applications are valid for linear and non-linear multi-input multi-output (MI/MO) systems are show promising developments that chattering will be entirely eliminated [13].

The conventional sliding mode controllers (SMC) have been applied to track motion profiles for example, Jamaludin, Brussel and Swevers [12] compared P-PI and classical SMC control tracking of an x–y milling process with disturbance rejection using reference set-point tracking and circular tests. They reported that the conventional sliding mode controller worked well but required a relatively high-frequency bandwidth to ensure better stiffness. Wu and Ding [14] controlled a high motion system using an iterative-learning variable-structure controller. A multi-segment SMC controller was also proposed to reduce trapezoidal motion profile tracking error [15].

The discrete model reference integral sliding mode controller (DMRISMC) has been applied to control of simple mechanical structures as well as to linear motor based precision servo systems. This type of controller was necessary because

difficulties in specifying design objectives in terms of a performance index and large variations in plant parameters resulted in conventional SMC and linear optimal controllers not being suitable. Another advantage of using DMRISMC is that the servo problem and the regulator problem can be separated [17].

In an DMRISMC system the desired response to a desired signal is specified by a reference signal which provides the desired signal to the feedback loop. The block diagram of DMRISMC is shown in Fig. 72.4. The task of the controller is simply to drive the error between the output of the process and the output of the reference model to zero [16, 17]. For this reason, the DMRISMC has been proposed to address partial image blurring in the Statscan™ X-ray machine owing to its robustness properties.

We therefore, compare the motion tracking ability of the DMRISMC and the PI controllers, in order to ultimately suggest a method to improve Statscan image quality.

72.2 Problem Overview

72.2.1 The System Description

Figure 72.1 shows the Statscan™ digital radiography X-ray mechanism. Its C-arm is driven along a linear length of 1800 mm by an ironless linear motor, model IL-240-50 AI, manufactured by Kollmorgen, and is equipped with a linear optical encoder (LS 603) by Heidenhain as the measurement feedback component. The machine's total mass is approximately 1219 kg, while the total moving mass of the C-arm (including the linear motor coil, the X-ray tube, heat exchanger and detector) is 515 kg. The control hardware systems include Cascaded P-PI controller and SERVOSTAR® (CE06250) drive produced by Kollmorgen. According to Ding and Wu [6] the nominal model of such system can be described by

$$g(s) = \frac{y(s)}{u(s)} = \frac{A_g}{s(Ts + 1)} \quad (72.1)$$

where $g(s)$ is the dynamic relation between the system position $y(s)$ and the input $u(s)$ and the system velocity is described by means of a first order system by differentiating $y(s)$.

It was not possible to obtain the system model parameters such as, the system gain (A_g) and time constant (T) using a step tests. Owing similarity, an alternative model was therefore, obtained by using step tests on a laboratory scale DC motor.

Fig. 72.1 Statscan™ X-ray machine

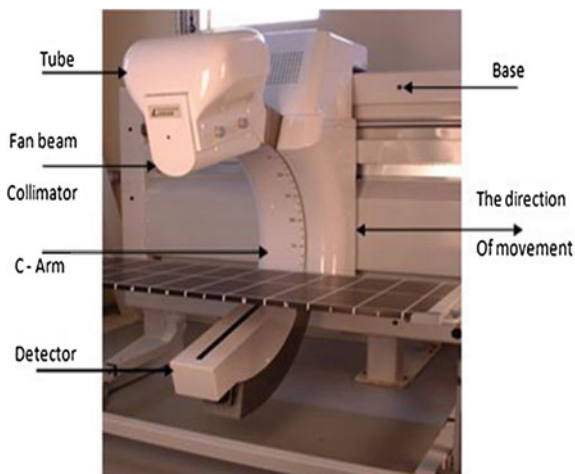
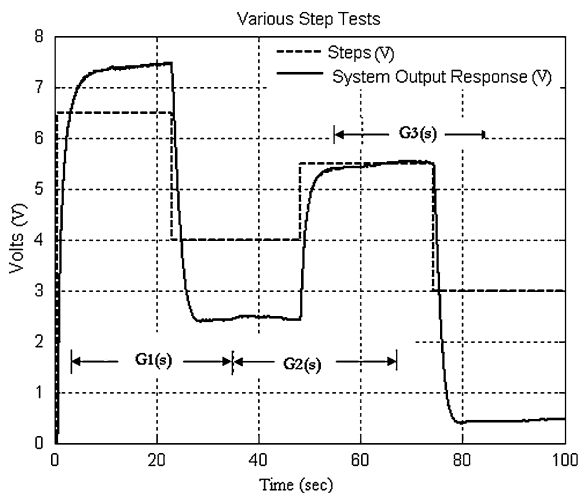


Fig. 72.2 Various step test



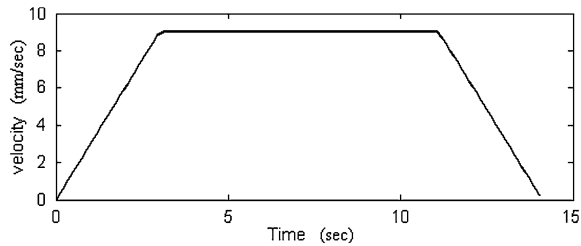
72.2.2 System Model Identification

To obtain a matched model that implements both control algorithms, step tests data were obtained as shown in Fig. 72.2. On the graph, the time is in seconds while the output velocity and input of the motor are measured in volt (V).

Simple graphical analysis techniques were applied to step tests data to identify the model [18]. The system transfer function model is given by as:

$$g(s) = \frac{2.52}{s(1.36s + 1)} \quad [V/V] \quad (72.2)$$

Fig. 72.3 Trapezoidal velocity motion profile



To improve DMRISMCM tracking performance, the position model in “(72.2),” will be used for velocity for the DMRISMCM implementation.

72.2.3 Trapezoidal Motion Profiles

In all motion control applications described earlier, the controller’s aim was to ensure that the system followed predesigned trajectory motion profile. Several of these motion profiles have been proposed with aims of suppressing transient vibration, reducing settling time, minimizing overshoot, and alleviating disturbances and uncertainties of the plant parameters [3, 4, 19]. The tradeoff between the speed of motion and vibration reduction complicates motion planning [6]. The faster the system, the higher the rapid increment or decrement of acceleration. This results in a mismatch between the detector’s clocking speed and the linear speed of the C-arm during the early part of the scan and hence the degradation in the image quality in the form of partial image blurring. Currently, the strategy implemented in the Statscan™ machine is to suppress the initial data that may be severely blurred. However, this issue needs to be addressed. The trapezoidal motion profile as illustrated in Fig. 72.3 were used in this simulation and experimental study.

It is expected that the robustness of sliding mode control method against disturbances and plant uncertainties will result in digital sliding mode control (DSMC) out-performing other control methodologies. This is specifically expected to be valid around rapid increments or decrements in acceleration (Fig. 72.4).

72.3 Controllers Design

72.3.1 Regular Form

Prior to the DMRISMCM’s design, the system transfer function model in “(72.2),” is transformed into a state space representation assuming that this state space model is further transformed into an appropriate regular form by choice of an orthogonal matrix [16].

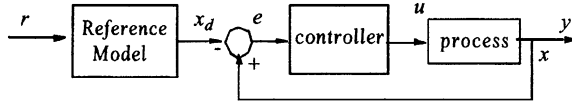


Fig. 72.4 The model reference control system as implemented in DMRISMC controller [17]

$$\dot{x}(t) = Ax(t) + Bu(t) + f(t, x, u) \quad (72.3)$$

$$y(t) = Cx(t)$$

where $x(t)$ is the state vector, $u(t)$ the system input, $y(t)$ the system output, system is full rank, (A, B) is controllable pair and, $f(t, x, u)$ presents the disturbances. The matrices $A \in \mathbb{R}^n \times \mathbb{R}^n$, $B \in \mathbb{R}^n \times \mathbb{R}^1$ are transformed to the discrete domain as given in [17], which are defined as:

$$x(k+1) = \phi x(k) + \Gamma u(k) + \Gamma f(k) \quad (72.4)$$

$$y(k) = Cx(k)$$

72.3.2 Reference Tracking Model

It is important to define pre-compensator reference model before designing the DMRISMC whose states correspond directly to states of the system [16], i.e.

$$x_d(k+1) = \phi_r x_d(k) + \Gamma_r r(k) \quad (72.5)$$

$$y_r(k) = C_r x_d(k)$$

Where

$$\Phi_r = \Phi - \Gamma L_f, \Gamma_r = \rho \Gamma \quad (72.6)$$

Where L_f is the feedback gain, ρ is a positive scalar and $r(k)$ is the command input at sampling time instant, Inserting Eq. 72.6 into “(72.5),” and substituting, $C = C_r$, and $e = x - x_d$, the error dynamics of the feedback control loop is obtained by subtracting the result of Eq. 72.4 from the results of inserting Eq. 72.6 into Eq. 72.5.

$$e(k+1) = \phi e(k) + \Gamma u(k) + \Gamma (f(k) - u_f(k)) \quad (72.7)$$

$$y_e(K) = Ce(K)$$

where $u_f(k) = -L_f x_r + \rho r(k)$

72.3.3 Control Law

For SISO tracking control, the integral is added to improve tracking performance and the sliding surface is defined as:

$$S(e) = Se = [\lambda \quad 1] e \quad (72.8)$$

where $S\Gamma$ is non-zero, Γ is the system input vector in “(72.4),”

λ is a positive scalar which determines the rate of converging to the sliding surface. The next step is to design the controller to satisfy the reachability and sliding phase conditions [17], the controller composed of discontinuous or switching controller (u_d) and equivalent controller (u_{eq}) which defined as:

$$u = u_{eq} + u_d \quad (72.9)$$

where $u_{eq} = (S\Gamma)^{-1} S \Phi e(k) + u_d(k) - f(k)$ and $u_d = k \operatorname{sgn}(s(e))$ where

$$\operatorname{sgn}(s(e)) = \begin{cases} 1 & \text{if } s(e) > 0 \\ -1 & \text{if } s(e) < 0 \end{cases} \quad (72.10)$$

A drawback of a defined controller structure is chattering. Chattering refers to finite-frequency magnitude oscillations around the sliding surface, due to the discontinuous nature of the control action that switches between two system structures. In order to overcome chattering, boundary layer techniques [16, 19] are used by replacing the signum-like function Eq. 72.10 with a continuous approximation function $v_\delta(s(e))$ in the thin boundary layer about the sliding surface which takes the following form:

$$v_\delta(s(e)) = \frac{s(e)}{|S(e)| + \delta} \quad (72.11)$$

where, δ is positive constant that represents the degree of the continuous approximation and $s(e)$ is a switching surface for error dynamics.

72.3.4 Formulation of PI Controller

For motion applications, the P-PI cascaded formulation still remains the dominant strategy in servo systems. In this configuration, a proportional controller (P) controls the outer position loop and a proportional integral controller (PI) controls the velocity inner loop [4]. A theoretical analysis of this form of controller is available [4, 9, 12]. For simplicity, the PI controller formula that was implemented in the velocity loop is given as:

$$PI(s) \frac{k_p(s + k_i)}{s} \quad (72.12)$$

Where k_p and k_i are PI controllers gains. In discrete domain the difference equation method is used to implement the PI controller.

72.4 Simulation and Experimental Results

72.4.1 Simulations Results

The computer simulation results of the DC motor discrete state space model of Eq. 72.4 are presented to investigate DMRISM C controller performance in tracking a trapezoidal motion profile. The simulation of both controllers was written in MATLAB. The model is transformed into the regular form as in Eq. 72.3 and then discretized with sampling time $T_s = 0.109$ s,

$$(\phi, \Gamma) = \left(\begin{bmatrix} 1 & 0.1047 \\ 0 & 0.9230 \end{bmatrix}, \begin{bmatrix} 0.0058 \\ 0.1047 \end{bmatrix} \right)$$

The reference model is obtained according to Eq. 72.6 results in

$$(\phi_r, \Gamma_r) = \left(\begin{bmatrix} 0.9781 & 0.0879 \\ -0.3967 & 0.6181 \end{bmatrix}, \begin{bmatrix} 0.0879 \\ 0.3967 \end{bmatrix} \right)$$

Where, $L_F = [3.7887, 2.912]$, $P = 3.7887$.

For robust performance, the DMRISM C parameters need to be tuned. The tuning parameters were chosen based on guidelines that were provided by Kaya [20]. The controller designer has to set the rate of converging to sliding surface (λ), switching amplitude (k), boundary layer width (δ). The guidelines required that all DMRISM C parameters must be positive, resulting system dynamics in switching surface must be stable, and that the discontinuous controller must be kept reasonable in order to avoid damaging the actuator. However, in DMRISM C these guidelines will not have much influence since the controller action is very small due to the Moore–Penrose Matrix inverse [16]. Therefore, the S vector was chosen to be [1, 2], k was chosen to be 5, δ was chosen to be 2 for chatter reduction and sampling times for DMRISM C simulation were chosen to be 10 and 109 ms and for PI simulation was chosen to be 109 ms.

For the PI controller, the two parameters that must be tuned are proportional gain (k_p) and rest integral time gain (k_i), These parameters were chosen to be 6.60 and 3.59 respectively based on the root locus tuning method for P-PI cascaded for servo systems, discussed by Żabiński and Trybus [21].

The simulation results of the DMRISM C and PI controllers with low tracking error are illustrated in Fig. 72.5 and Fig. 72.6 for DMRISM C and Fig. 72.7 for PI. A trapezoidal motion profile is applied. In the simulation, the high acceleration is

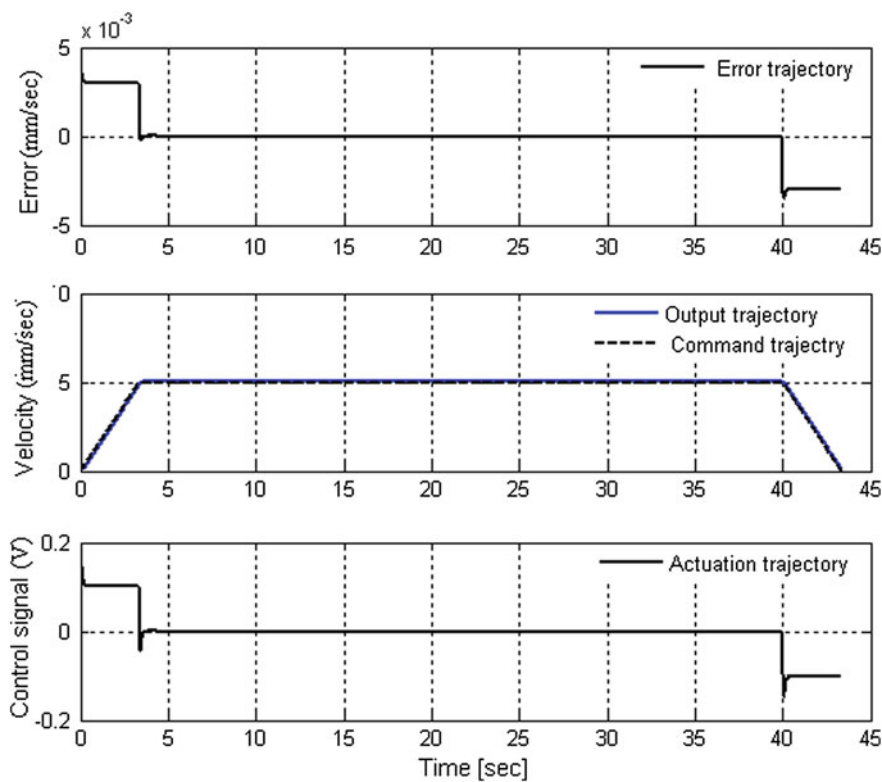


Fig. 72.5 DMRISMC tracking Simulation at sampling time 10 ms

set to 1.5 mm/sec^2 , the desired position is set to 200 mm and the maximum speed is set to 5 mm/sec.

Looking at trapezoidal motion profile tracking, the DMRISMC provides perfect performance, despite a motion tracking error which is around zero. However, it is noted that the tracking error is proportional to the sampling rate: The lower the sampling time the lower the tracking error. This is obtained at the cost of the low control action compared to PI controller, see Fig. 72.7.

Table 72.1 shows the average final position errors obtained for both controllers when the simulation tests run at speeds of 3, 6 and 9 mm/sec with 100, 200 and 300 mm positions. The results show that DMRISM achieves almost the same position error at various velocities while PI increases the position error when the velocity increases. This observation may illustrate why there is more distortion on image at high velocity motions.

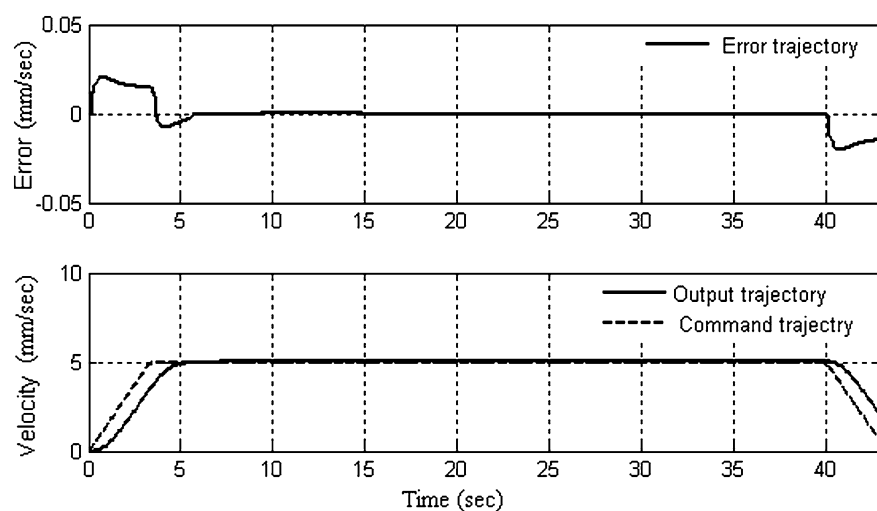


Fig. 72.6 DMISM tracking simulation at sampling time 0.109 s

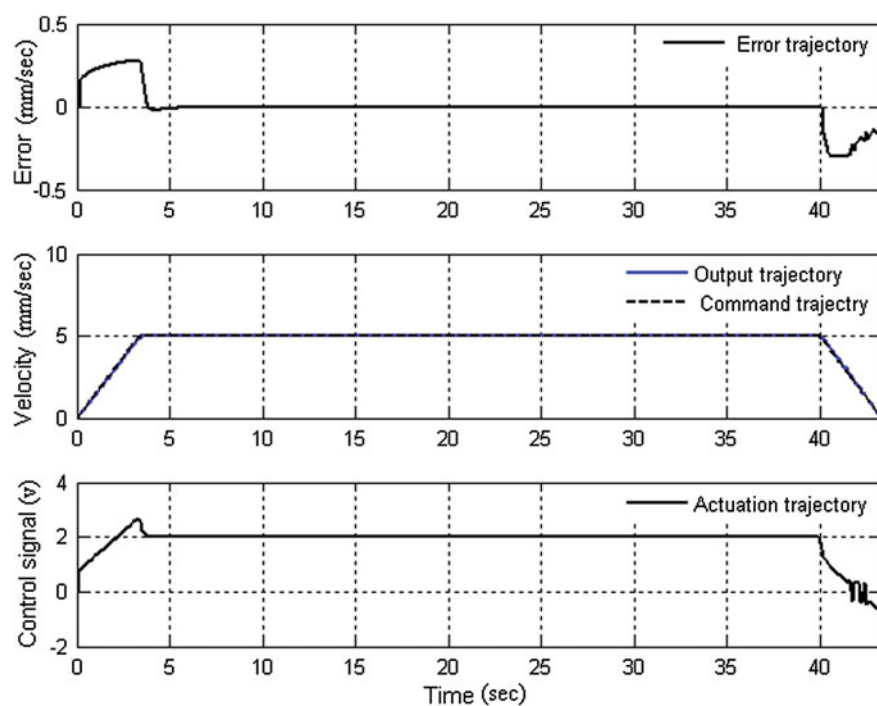


Fig. 72.7 Trapezoidal motion profile simulation using PI

Table 72.1 the Average Position Error

| Velocity | PI Average position error | DMRISM Average position error |
|----------|------------------------------|----------------------------------|
| 3 mm/sec | 0.268 mm | 0.0157 mm |
| 6 mm/sec | 0.5927 mm | 0.0382 mm |
| 9 mm/sec | 0.927 mm | 0.0238 mm |

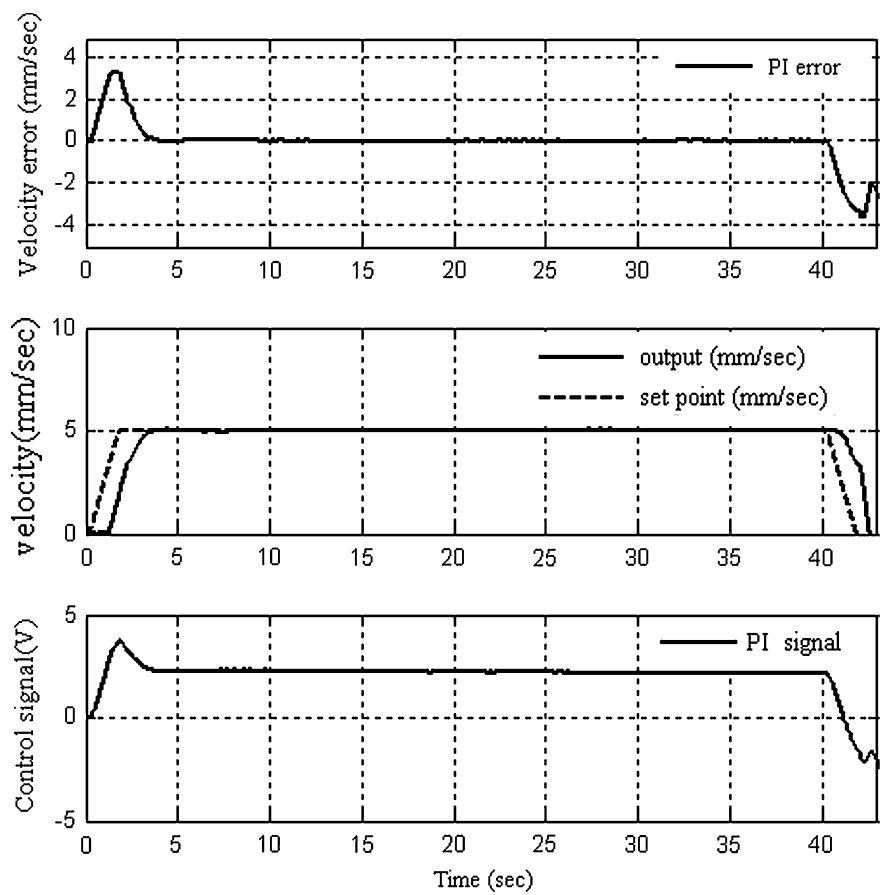


Fig. 72.8 PI experimental test results

72.4.2 Experimental Results

The following plots present the experimental time responses for simulated controllers in Sect. 72.4.1. The implementations were coded in C++, linked with Newmat matrix library and compiled on a Borland C++ compiler [22]. The-PI

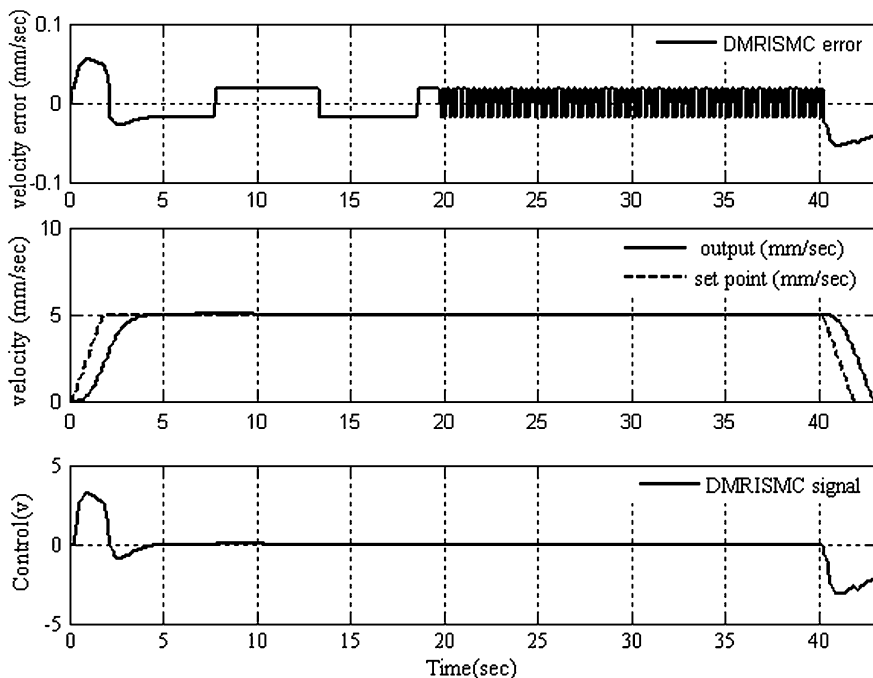


Fig. 72.9 DMRISMIC experimental test results

controller gains were chosen as $k_p = 0.964$ and $k_i = 0.65$ while for the DMRISMIC the same simulation parameters were applied with observer feedback gain matrix $[-3.077, 21367]^T$ when system poles assigned at -2 and -3 on the s plane. In the experiment one of the limiting factors is sampling period, therefore, the sampling rate was chosen to be 0.109 s for both controllers.

Relating in Fig. 72.8, during the acceleration and deceleration phases the PI controller requires ± 3.5 V as control effort during whilst an effort of 2.3 V is necessary when the system reaches the steady state. By looking at tracking errors, the PI's velocity error is 0.02808 . The system reaches a steady state after 3 s and there is an overshoot of 0.008 which causes the system to settle at 5 s.

The velocity error, trapezoidal motion profile tracking and DMRISMIC signal responses are shown in Fig. 72.9. The DMRISMIC controller produces a control signal at ± 3.2 V magnitude during the acceleration and deceleration phases, whilst at steady state gives 0 V.

The system reaches a steady state after 3.4 s with velocity error 0.02 mm/sec and zero overshoot. Looking more closely at experimental time 20 s, DMRISMIC velocity error, some chattering still exists i.e. not entirely removed when boundary layer was used.

72.5 Conclusion

The paper has outlined the trapezoidal motion profile performance of DMRISM and PI controllers for improving image quality of Statscan™ X-ray machine. The performances of these controllers were compared by using simulation and experiments. Both controllers are competitive when compared to each other on trapezoidal motion profile tracking at the same 109 ms sampling rate. The DMRISM controller was found to be better in position steady-state error if a low sampling rate (10 ms) is used and requires less control effort than a PI controller. However, when it comes to image quality improvement, DMRISM controllers' performance was expected to improve the image quality since there were no overshoot due to jerk at acceleration and deceleration phases. However, further investigation on image quality assessments methods may be needed in order to clarify this unsubstantiated conclusion. For further more accurate comparison, a model of the Statscan™ X-ray machine is needed instead of a DC motor to match the system linear positioning specifications.

Acknowledgments We would like to thank Lodox Systems and the University of Cape Town (UCT) for their financial support.

References

1. Exadaktylos AK, Benneker LM, Jeger V, Martinolli L, Bonel HM, Eggli S, Potgieter H, Zimmermann H (2008) Total-body digital X-ray in trauma. An experience report on the first operational full body scanner in Europe and its possible role in ATLS. *Injury* 39:525–529
2. Potgieter JH, de Villiers M, Scheelke M, de Jager G (2005) An explanation for the extremely low, but variable, radiation dosages measured in a linear slit scanning radiography system. In: Stepp LM (ed) *Proceedings of SPIE, San Diego, CA, USA, February 13, 2005*, SPIE, Bellingham, WA, pp 1138–1145
3. Li HZ, Gong ZM, Lin W, Lippa T (2007) Motion profile planning for reduced jerk and vibration residuals. Vol 8, pp 32–37
4. Ding -H, Wu J (2007) Point-to-point motion control for a high-acceleration positioning table via cascaded learning schemes. *IEEE Trans Ind Electron* 54:2735–2744
5. Rew KH, Ha CW, Kim KS (2009) A practically efficient method for motion control based on asymmetric velocity profile. *Int J Mach Tools Manuf* 49:678–682
6. Zheng C, Su Y, Müller PC (2009) Simple online smooth trajectory generations for industrial systems. *Mechatronics* 19:571–576
7. Ahmad W, Htut MM (2009) Neural-tuned PID controller for point-to-point (PTP) positioning system: Model reference approach. In: *Proceedings of the international conference on man-machine systems*, Batu Ferringhi, Penang, MA, October 11–13, 2009. pp 140–144
8. Li G, Tsang KM (2007) Concurrent relay-PID control for motor position servo systems. *Int J Control Autom Syst* 5:234–242
9. Ellis GH (2004) *Control System Design Guide: A practical guide*. Elsevier Academic Press, Amsterdam
10. Elker I (2010) Second-order sliding mode control with experimental application. *ISA Trans* 49:394–405

11. Young KD, Utkin VI, Ozguner U (1999) A control engineer's guide to sliding mode control. *IEEE Trans Control Syst Technol* 7:328–342
12. Jamaludin Z, Brussel HV, Swevers J (2007) Classical cascade and sliding mode control tracking performances for a xy feed table of a high-speed machine tool. *Int J Precis Technol* 1:65–74
13. Kaynak O, Yu X (2009) Sliding mode control with soft computing: a survey. *IEEE Trans Ind Electron* 56:3275–3285
14. Wu -J, Ding H (2008) Iterative learning variable structure controller for high-speed and high-precision point-to-point motion. *Robot Comput Integr Manuf* 24:384–391
15. Lai CK, Shyu KK (2005) A novel motor drive design for incremental motion system via sliding-mode control method. *IEEE Trans Ind Electron* 52:499–507
16. Edwards C, Spurgeon S (1998) *Sliding mode control: theory and applications*. Taylor & Francis Press, London
17. Li YF, Wikander J (2004) Model reference discrete-time sliding mode control of linear motor precision servo systems. *Mechatronics* 14:835–851
18. Lee, Edgar TF (2010) Simple graphical method for noisy pulse and step responses. *Chem Eng Sci* 65:2629–2633
19. Rew -KH, Ha CW, Kim KS (2009) A practically efficient method for motion control based on asymmetric velocity profile. *Int J Mach Tools Manuf* 49:678–682
20. Kaya I (2008) Performance improvement of unsymmetrical processes using sliding mode control approach. *Energy Convers Manag* 49:101–106
21. Żabiński T, Trybus L (2010) Tuning P-PI and PI-PI controllers for electrical servos. *Bull Pol Acad Sci Tech Sci* 58:51–58
22. Davies R (2011) Newmat C++ matrix libraries. <http://www.robertnz.net/> Accessed November 2011

Chapter 73

Mechanical Energy Conversion to Electromagnetic Energy for Magnetic Fluids: Theoretical Fundamentals and Applications

Aurel-George Popescu and Adrian Runceanu

Abstract This article treats the static conversion of mechanical energy to electromagnetic energy, known as electromagnetic-acoustic, in magnetic fluids (E.M.A.) showing correlation to the “Theory of Sonics”, part of the domain named “Electro sonicity”. It retains the quality and is named magnetic fluid, a colloidal system formed by dispersing to monodomenic sizes, followed by the stabilization of a solid material with magnetic properties, in a liquid, usually with dielectrically properties. The electromagnetic–acoustic transformation [1] in a medium shows all the phenomenons linked to the kindling and recording of elastic oscillations by using the electromagnetic field or to the electromagnetic waves by using the mechanical oscillations. The original issues contained in this article are: (1) The highlighting of the physical and mathematical connection between the specific electrical formalism and the Theory of Sonics by introducing the sonic power in B8 (B18) formula. (2) The materialization of theoretical results on the conversion of mechanical energy into electromagnetic energy, as a vibration transducer whose originality is certified by the patent no 110872C1/1996 by title “Vibration transducer” [4].

A.-G. Popescu (✉) · A. Runceanu
University Constantin Brâncuși of Târgu-Jiu, Târgu-Jiu, Romania
e-mail: george@utgjiiu.ro

A. Runceanu
e-mail: adrian_r@utgjiiu.ro

73.1 Introduction

The scientist George Constantinescu, author of “Theory of Sonics”, was honored during his lifetime by including his photo in a commemorative picture—“One of seventeen world leaders in science and technology in the March of Progress 1900–1925”, published by the Graphic in 1926. On this commemorative picture the scientist is presented along with Einstein, Kelvin, Edison, Marie Curie, Rutherford etc. After his death (1966), in the Engineers Society Journal of England (volume LVII, January–March 1966, no. 1), in the presented obituary, we can see these words: “His paper Theory of Sonics, always considered by him the most important piece of work, was published by the Admiralty in 1918, but even today almost 50 years later, is unlikely that we truly understood it more than we understand Einstein’s theory” [5–10].

Despite international recognition of lifetime, his work was mentioned along time in obscurity, with little publicized:

1918, British Admiralty, 150 copies

1922, Bucharest, Typography “Cultura”, 2000 copies

1985, Bucharest, Academy Publishing House RSR, 2000 copies

73.2 Theoretical Fundamental of the Electromagneto-Acoustic Transformation in Magnetic Fluids

In the stationary case, the magneto static equations are these:

$$\operatorname{rot} \vec{H}_0 = 0; \quad \operatorname{div} \mu_0 \mu_r \vec{H}_0 = 0 \quad (73.1)$$

In the constant magnetic field \vec{H}_0 the magnetic permeability is a function of the thermo dynamical variables: density ρ and temperature T . The mechanical wave propagation in the probe leads to the adiabatic modification of the magnetic permeability μ' as a cause of the changing in oscillation density and temperature in the sound wave front wave [6–17].

In the case of mechanical wave propagation, Maxwell’s equations can be written as:

$$\operatorname{rot} \vec{E} = -\mu_0 \frac{\partial}{\partial t} \left[(\mu_r + \mu') (\vec{H} + \vec{h}) \right] \quad (73.2)$$

$$\operatorname{rot} (\vec{H} + \vec{h}) = \varepsilon_0 \varepsilon_r \frac{\partial \vec{E}}{\partial t} \quad (73.3)$$

$$\operatorname{div} \vec{E} = 0 \quad (73.4)$$

$$\operatorname{div} \mu_0 (\mu_r + \mu') (\vec{H} + \vec{h}) = 0 \quad (73.5)$$

The (73.2) equation can be written as:

$$\begin{aligned} \operatorname{rot} \vec{E} &= -\mu_0 \frac{\partial}{\partial t} \left[\mu_r \vec{H}_0 + \mu_r \vec{h} + \mu' \vec{H}_0 + \mu' \vec{h} \right] \\ \operatorname{rot} \vec{E} &= \underbrace{-\mu_0 \mu_r \frac{\partial \vec{H}_0}{\partial t}}_{=0} - \mu_0 \mu_r \frac{\partial \vec{h}}{\partial t} - \underbrace{\mu_0 \mu' \frac{\partial \vec{H}_0}{\partial t}}_{=0} \\ &\quad - \mu_0 \mu' \frac{\partial \vec{h}}{\partial t} - \underbrace{\mu_0 \vec{H}_0 \frac{\partial \mu_r}{\partial t}}_{=0} - \underbrace{\mu_0 \vec{h} \frac{\partial \mu_r}{\partial t}}_{=0} \\ &\quad - \mu_0 \vec{H}_0 \frac{\partial \mu'}{\partial t} - \mu_0 \vec{h} \frac{\partial \mu'}{\partial t} \\ \operatorname{rot} \vec{E} &= -\mu_0 \mu_r \frac{\partial \vec{h}}{\partial t} - \mu_0 \vec{H}_0 \frac{\partial \mu'}{\partial t} - \mu_0 \mu' \frac{\partial \vec{h}}{\partial t} - \mu_0 \vec{h} \frac{\partial \mu'}{\partial t} \end{aligned}$$

The (73.3) equation can be written as:

$$\operatorname{rot} (\vec{H}_0 + \vec{h}) = \underbrace{\operatorname{rot} \vec{H}_0}_{=0} + \operatorname{rot} \vec{h} = \operatorname{rot} \vec{h} = \varepsilon_0 \varepsilon_r \frac{\partial \vec{h}'}{\partial t}$$

The (73.5) equation is reduced to:

$$\begin{aligned} \operatorname{div} \mu_0 (\mu_r + \mu') (\vec{H}_0 + \vec{h}) &= \operatorname{div} \left[\mu_0 \mu_r \vec{H}_0 + \mu_0 \mu_r \vec{h} + \mu_0 \mu' \vec{H}_0 + \mu_0 \mu' \vec{h} \right] \\ &= \underbrace{\operatorname{div} (\mu_0 \mu_r \vec{H}_0)}_{=0} + \operatorname{div} (\mu_0 \mu_r \vec{h}) \\ &\quad + \operatorname{div} (\mu_0 \mu' \vec{H}_0) + \operatorname{div} (\mu_0 \mu' \vec{h}) \\ &= \underbrace{\vec{h} \cdot \operatorname{grad} \mu_0 \mu_r}_{=0} + \mu_0 \mu_r \cdot \operatorname{grad} \vec{h} + \vec{H}_0 \cdot \operatorname{grad} (\mu_0 \mu') \\ &\quad + \underbrace{\mu_0 \mu' \cdot \operatorname{grad} \vec{H}_0}_{=0} + \vec{h} \cdot \operatorname{grad} (\mu_0 \mu') \\ &\quad + \mu_0 \mu' \cdot \operatorname{grad} \vec{h} = \mu_0 \mu_r \cdot \operatorname{grad} \vec{h} \\ &\quad + \vec{H}_0 \cdot \operatorname{grad} (\mu_0 \mu') + \vec{h} \cdot \operatorname{grad} (\mu_0 \mu') \\ &\quad + \mu_0 \mu' \cdot \operatorname{grad} \vec{h} \end{aligned}$$

Thus, from (73.2) and (73.5) we obtain:

$$\text{rot } \vec{E} = -\mu_0 \left[\mu_r \frac{\partial \vec{h}}{\partial t} + \vec{H}_0 \frac{\partial \mu'}{\partial t} + \mu \frac{\partial \vec{h}}{\partial t} + \vec{h} \frac{\partial \mu'}{\partial t} \right] \quad (73.6)$$

$$\text{div } \vec{E} = 0 \quad (73.7)$$

$$\text{rot } \vec{h} = \varepsilon_0 \varepsilon_r \frac{\partial \vec{E}}{\partial t} \quad (73.8)$$

By applying the rot operator to (73.6) we obtain:

$$\begin{aligned} \text{rot} \cdot \text{rot } \vec{E} &= \text{rot} \left[-\mu_0 \mu_r \frac{\partial \vec{h}}{\partial t} - \mu_0 \vec{H}_0 \frac{\partial \mu'}{\partial t} - \mu_0 \mu' \frac{\partial \vec{h}}{\partial t} - \mu_0 \vec{h} \frac{\partial \mu'}{\partial t} \right] \\ &= -\text{rot} \left(\mu_0 \mu_r \frac{\partial \vec{h}}{\partial t} \right) - \text{rot} \left(\mu_0 \vec{H}_0 \frac{\partial \mu'}{\partial t} \right) \\ &\quad - \text{rot} \left(\mu_0 \mu' \frac{\partial \vec{h}}{\partial t} \right) - \text{rot} \left(\mu_0 \vec{h} \frac{\partial \mu'}{\partial t} \right) = 0 \\ \text{rot} \cdot \text{rot } \vec{E} &= \frac{\partial \vec{h}}{\partial t} \times \underbrace{\text{grad}(\mu_0 \mu_r)}_{=0} - \mu_0 \mu_r \cdot \text{rot} \frac{\partial \vec{h}}{\partial t} \\ &\quad + \underbrace{\vec{H}_0 \times \text{grad} \left(\mu_0 \frac{\partial \mu'}{\partial t} \right)}_{=0} - \mu_0 \frac{\partial \mu'}{\partial t} \cdot \underbrace{\text{rot } \vec{H}_0}_{=0} \\ &\quad + \underbrace{\frac{\partial \vec{h}}{\partial t} \times \text{grad}(\mu_0 \mu')}_{=0} - \mu_0 \mu' \cdot \text{rot} \frac{\partial \vec{h}}{\partial t} \\ &\quad + \underbrace{\vec{h} \times \text{grad} \left(\mu_0 \frac{\partial \mu'}{\partial t} \right)}_{=0} - \mu_0 \frac{\partial \mu'}{\partial t} \cdot \underbrace{\text{rot } \vec{h}}_{=0} \\ \text{rot} \cdot \text{rot } \vec{E} &= -\mu_0 \mu_r \cdot \text{rot} \frac{\partial \vec{h}}{\partial t} + \vec{H}_0 \times \text{grad} \left(\mu_0 \frac{\partial \mu'}{\partial t} \right) \\ &\quad - \mu_0 \mu' \cdot \text{rot} \frac{\partial \vec{h}}{\partial t} + \vec{h} \times \text{grad} \left(\mu_0 \frac{\partial \mu'}{\partial t} \right) \end{aligned}$$

Analyzing the previous equation terms, using equations, using (73.6–73.8):

$$\begin{aligned}
\mu_0 \mu_r \cdot \text{rot} \frac{\partial \vec{h}}{\partial t} &= \frac{\partial}{\partial t} \mu_0 \mu_r \cdot \text{rot} \vec{h} = \frac{\partial}{\partial t} \mu_0 \mu_r \varepsilon_0 \varepsilon_r \frac{\partial \vec{E}}{\partial t} \\
&= \mu_0 \mu_r \varepsilon_0 \varepsilon_r \frac{\partial^2 \vec{E}}{\partial t^2} = \frac{1}{v^2} \cdot \frac{\partial^2 \vec{E}}{\partial t^2} \\
\vec{H}_0 \times \mu_0 \text{grad} \left(\frac{\partial \mu'}{\partial t} \right) &= \vec{H}_0 \times \left(\text{grad} \frac{\partial \mu'}{\partial t} + \frac{\partial \mu'}{\partial t} \cdot \text{grad} \mu_0 \right) \\
&= \vec{H}_0 \times \text{grad} \mu_0 \frac{\partial \mu'}{\partial t} + \underbrace{\vec{H}_0 \times \frac{\partial \mu'}{\partial t} \cdot \text{grad} \mu_0}_{=0} \\
&= \vec{H}_0 \times \mu_0 \text{grad} \frac{\partial \mu'}{\partial t} = \mu_0 \left(\vec{H}_0 \times \text{grad} \frac{\partial \mu'}{\partial t} \right) \\
\mu_0 \mu' \cdot \text{rot} \frac{\partial \vec{h}}{\partial t} &= \mu_0 \mu' \frac{\partial}{\partial t} \left(\frac{\partial \vec{h}}{\partial t} \right) = \underbrace{\mu_0 \mu' \frac{\partial^2 \vec{h}}{\partial t^2}}_{=0} \\
\vec{h} \times \text{grad} \left(\mu_0 \frac{\partial \mu'}{\partial t} \right) &= \vec{h} \times \left(\frac{\partial}{\partial t} \left(\mu_0 \frac{\partial \mu'}{\partial t} \right) \right) = \underbrace{\vec{h} \times \left(\mu_0 \frac{\partial^2 \mu'}{\partial t^2} \right)}_{=0}
\end{aligned}$$

We obtain:

$$\text{rot} \cdot \text{rot} \vec{E} = -\frac{1}{v^2} \cdot \frac{\partial^2 \vec{E}}{\partial t^2} + \mu_0 \left(\vec{H}_0 \times \text{grad} \frac{\partial \mu'}{\partial t} \right)$$

On the other hand, we know:

$$\text{rot} \cdot \text{rot} \vec{E} = \text{grad} \left(\text{div} \vec{E} \right) - \Delta \vec{E}$$

So:

$$-\frac{1}{v^2} \cdot \frac{\partial^2 \vec{E}}{\partial t^2} + \mu_0 \left(\vec{H}_0 \times \text{grad} \frac{\partial \mu'}{\partial t} \right) = \underbrace{\text{grad} \left(\text{div} \vec{E} \right)}_{=0} - \Delta \vec{E}$$

Thus the (73.6–73.8) equations can be written as a single tridimensional equation:

$$\Delta \vec{E} - \frac{1}{v^2} \cdot \frac{\partial^2 \vec{E}}{\partial t^2} = -\mu_0 \left(\vec{H}_0 \times \text{grad} \frac{\partial \mu'}{\partial t} \right) \quad (73.9)$$

Which shows the dependency between the electrical field \vec{E} generated in the medium as a result of the μ' variation—of the μ_r magnetic permeability—in the

wave front, in the presence of a constant magnetic field of intensity \vec{H} when the magnetic fluid sample is crossed by mechanical waves.

73.3 Vibration Traductor Functioning Fenomenology Theoretical Fundaments

The (73.9) equation confirms the possibility of direct conversion between mechanical energy into electrical energy detectably by a command coil. Applying from the exterior, to the magnetic fluid, a $\vec{B} = \mu \vec{H}_0$ magnetic field of induction along the Ox axis, the (73.9) equation will become:

$$-\vec{B}_0 \frac{\partial}{\partial x} \frac{\partial \mu'}{\partial t} = -B_0 \frac{\partial}{\partial t} \frac{\partial \mu'}{\partial x}$$

Thus, in the wave front, the magnetic permeability μ_r , will simultaneously support a spatial variation and a temporal one, which, phenomenologically, is real.

According the Huygens' Principle, the mechanic waves are transmitted in a medium with elastic properties, from particle to particle, and the medium particles are considered as keeping mechanic energy oscillating to the equilibrium position. Thus, in the positive wave front, appears a local compression of the medium particles, and in the negative wave front, a local relaxation, thus a local variation in the medium density.

Because in a magnetic fluid the particles with magnetic properties are dispersed evenly in the medium, it results that, in the wave front, a local variation in the magnetic phase will appear, and a local variation in space of the relative magnetic permeability μ_r , expressed by $\frac{\partial \mu'}{\partial x}$.

Because the mechanic waves are propagated through the elastic medium, the spatial variation in magnetic permeability μ_r is accompanied by a temporal variation expressed by $\frac{\partial}{\partial t} \cdot \frac{\partial \mu}{\partial x}$.

The left term in the (73.9) equations express the form of the electric field induced to a detection coil by the local spatial-temporal variation in magnetic phase density in the wave front.

The “−” sign shows, according to the electromagnetic induction law, that the electric tension induces is of opposing sign to the spatial-temporal variation in the wave front.

Considering a harmonic variation in the wave front density, the μ' variation in relative magnetic permeability μ_r will be of this form:

$$\mu' = \mu_r \cdot e^{i(\omega t - \vec{k} \cdot \vec{r})} \quad (73.10)$$

To which, by a variation in parameter μ_r at constant entropy:

$$\mu_r = \left(\frac{\partial \mu_r}{\partial \rho} \right)_S \cdot \rho'; \quad S - \text{Entropy} \quad (73.11)$$

Thus:

$$\mu' = \left(\frac{\partial \mu_r}{\partial \rho} \right)_S \cdot \rho' \cdot e^{i(\omega t - \vec{k} \cdot \vec{r})} \quad (73.12)$$

For a cell of section S filled with a magnetic fluid of permeability μ_r , to which is placed a detection coil with N spires introduced into a constant magnetic field of intensity \vec{H}_0 .

The presence of mechanical waves in the magnetic fluid generates a tension to the detection coil terminals, namely:

$$\begin{aligned} e &= -\frac{d\Phi}{dt} = -NS \frac{dB}{dt} = -NS\mu_0 H_0 \frac{d\mu'}{dt} \\ &= -NSB_0 \omega \mu_r \cdot i \cdot e^{i\omega t} \end{aligned}$$

in which $\mu_0 H_0 = B_0$ —is the magnetic induction of the magnetic field applied to the cell.

But $e^{ix} = \cos x + i \cdot \sin x$, thus:

$$e = NSB_0 \omega \mu_r \cdot \sin \omega t - i \cdot NSB_0 \omega \mu_r \cdot \cos \omega t$$

The modulus of the electric tension detected will be:

$$\begin{aligned} |e| &= \left[(NSB_0 \omega \mu_r)^2 + (NSB_0 \omega \mu_r)^2 \right]^{\frac{1}{2}} \\ |e| &= NSB_0 \omega \mu_r \cdot \sqrt{2} \end{aligned} \quad (73.13)$$

Or by using (73.11):

$$|e| = NSB_0 \omega \left(\frac{\partial \mu_r}{\partial \rho} \right)_S \cdot \rho' \cdot \sqrt{2} \quad (73.14)$$

According to the hydrodynamic theory of sound, the ρ' change in density ρ of the medium in the wave front is tied to the instantaneous oscillation speed v of the medium particles, by the relation:

$$\rho' = \rho \frac{v}{u} \quad (73.15)$$

In which u is the sound speed in the respective medium; Thus, (73.14) becomes:

$$|e| = NSB_0 \omega \cdot \left(\frac{\partial \mu_r}{\partial \rho} \right)_S \cdot \rho \frac{v}{u} \cdot \sqrt{2} \quad (73.16)$$

And constitutes with (73.13) the theoretical fundament on which the conception and projection of some installations for static conversion from mechanic to electromagnetic energy is founded.

By analyzing the (73.16) formula, we observe that it can be written as:

$$|e| = \frac{\sqrt{2NB_0\omega\rho}}{\mu} \cdot S \cdot v \cdot \frac{\partial\mu_r}{\partial\rho} = \alpha \cdot i \cdot \frac{\partial\mu_r}{\partial\rho} \quad (73.17)$$

in which:

$\alpha = \frac{\sqrt{2NB_0\omega\rho}}{\mu}$ - Constant which contains static parameters of a vibration transducer.

$i = S * v$ —instantaneous value of the sonic current [2].

$\frac{\partial\mu_r}{\partial\rho}$ - local variation in magnetic permeability μ_r in the wave front, due to the variation in magnetic phase density.

With the expression of sonic current ($i = S * v$) we can rewrite (73.17) as:

$$|e| = \alpha \cdot \frac{\partial\mu_r}{\partial\rho} \cdot I \cdot \sin(at + \Psi) \quad (73.18)$$

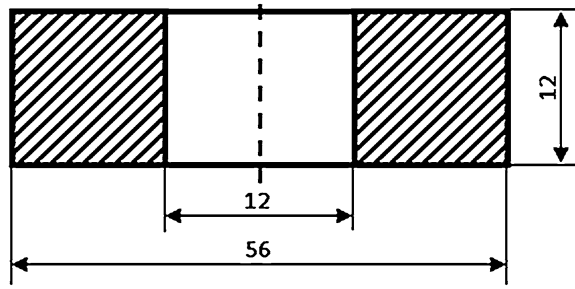
The appearing of phenomenological link which can be express between the E.M.A. Transformation (through which is fundamental the possibility of electric conversion) and the Sonic Theory (which offers the possibility of projecting installations with big efficiency), revealed by formula (73.16), allows projecting a cell in which the sonic parameters are linked with the electric parameters.

The presence of the $S * v$ product in (73.16) or (73.17)—defined in the Sonic Theory as being the instantaneous value of the sonic current is given the rigorous scientific character of being used in making conversion machines with efficiencies near one.

The correlation between the instantaneous sonic current value of the sonic current in formula (73.17) and the static projection parameters and with the local variation in magnetic permeability opens a huge field of research in the lesser-known field of “Electro sonicity”, thus yielding a new scientific dimension of a big interest, namely: “The possibility of projection and construction of static convertors of mechanic waves in electric energy, with increased efficiency”

This interdependencies between the Sonicity Theory and the direct mechanic energy conversion in electromagnetic energy, part of the “Electro sonicity” field, are not found in physics writings.

Fig. 73.1 A permanent magnet (circular crown shaped) with the active section $S=10^{-4} \text{ m}^2$



73.4 The Functioning Principle Behind the Vibration Translator

The permanent magnet (1) is generating a magnetic field whose field lines show their presence in the detection coil (2) and in the magnetic fluid (3), Fig. 73.3.

The magnetic wave presence in the magnetic field shows a spatial-temporal modification in the relative magnetic permeability value μ_r in the wave front, after which the magnetic field lines intensity becomes variable in space and time.

There are local modifications in magnetic reluctance, and the magnetic field lines tend to close through the minimum magnetic reluctance zones becoming variable by position and intensity, in time and space.

Thus, the detection coil will be swept by a variable magnetic field, which will induce through it electromotive tension.

The experimental results obtained [3]

The permanent magnet used is a type of circular crown with the dimensions in the figure:

In the experiments that we've done, we used a permanent magnet (circular crown shaped) with the active section $S = 10^{-4} \text{ m}^2$, having a magnetic induction of B_0 between $0.01 \text{ T} \div 0.08 \text{ T}$. (Fig. 73.1)

Visualizing the magnetic field lines, with iron powder, we noticed that they limit the active surface in three domains, namely:

Domain (1)

The magnetic field lines close in the exterior corner of the circular crown. The magnetic induction the magnet's active border is $B_1 = 0,08 \text{ T}$ (Fig. 73.2).

Domain (2)

This is the transition zone. The magnetic field lines are dispersed to the exterior and interior of the active surface of the magnet, and the magnetic inductance value is in this case $B_2 = 0,01 \text{ T}$.

Domain (3)

The magnetic field lines close to the interior of the active surface. The field line polarity is reverse to the zone (1). The maximum magnetic inductance value is $B_3 = 0,08 \text{ T}$.

The permanent magnet was introduced into iron housing.

Fig. 73.2 The magnetic induction the magnet's active border

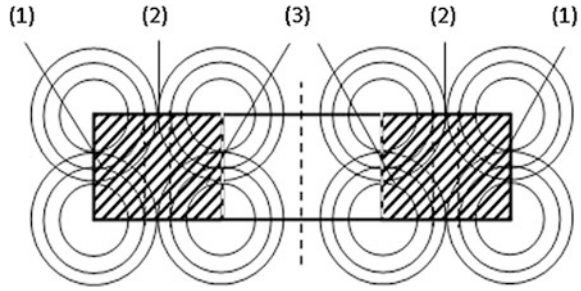
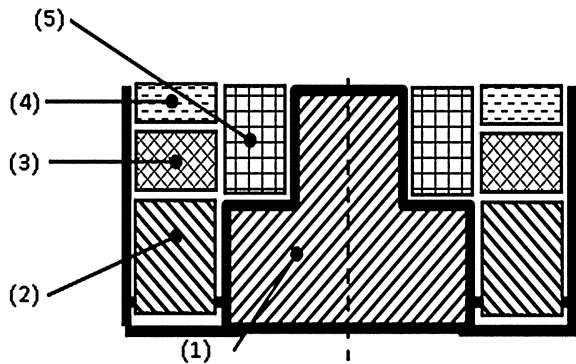


Fig. 73.3 The housing structure



The housing structure are introduced a flat coil with $N = 1000$ spires and a magnetic fluid inside, both placed on a ferromagnetic ring and applied on top of the permanent magnet. The whole system is presented in Fig. 73.3.

- (1) ferromagnetic housing
- (2) permanent magnet
- (3) flat coil
- (4) magnetic fluid inside
- (5) ferromagnetic ring

For the experimentation of the vibration translator we used a variable frequency between 50 Hz – 200 kHz.

73.5 Conclusions

After the experiments we drew the following conclusions:

- The current intensity through the detection coil is of tens of microamperes and as such, the vibration that we have made can be used safely in explosive environments.

- The electric signal amplitude goes from a few millivolts to a few volts.
- The link between the formula results (73.16) and the experimental value was proven.

References

1. Burzo E (1983) Physics of magnetic phenomena, vol I, II, III. RSR Academy Press, Bucharest
2. Constantinescu G (1985) Theory of sonics. 2nd edn. RSR Academy Press, London
3. Constantinescu G (1918) Theory of sonics: a treatise on transmission of power by vibrations, copy nr. 110, British Navy, London
4. Popescu AG (1996) Patent no. 110872C1/1996: Vibration transducer
5. Berkowetz AE, Lahut JA, Van Buren CG (1980) Properties of magnetic fluid particles. IEEE Trans Mag MAG-18(2):185
6. Dubbledday PS (1980) Application of ferrofluids as an acoustic transducer material. IEEE Trans Mag MAG-18(2):372
7. Jackson JD (1991) Classical electrodynamics. Tehnica Press, Bucharest
8. Sedov L (1975) Mecanique des Milieus Continus, tome 1st, 2nd edn. MIR, Moscow
9. Tarapov I, Patsegon NF, Phedonenco AI (1983) Some physical and mechanical phenomena in magnetizable fluids. JMMM 3(9):51
10. Vausovski SV (1983) Magnetism. RSR Academy Press, Bucharest
11. Grofu F, Popescu L, Cercel C (2007) Acquisition system for monitoring vibrations, international joint conferences on computer, information, and systems sciences, and engineering (CISSE 07)—international conference on industrial electronics, technology & automation (IETA 07), University of Bridgeport, USA, 03–12 December 2007, Springer Press, pp 89–92, ISI proceedings, ISBN 978-1-4020-8736-3, e- ISBN 978-1-4020-8737-0
12. Grofu F, Popescu M, Popescu L (2006) Data acquisition system for vibration signal, international journal of computers, communications & control, CCC Publications, 1–3 June 2006, Băile Felix-Oradea, Romania, pp 251–255 ISSN 1841-9386
13. Berkovsky B, Bashtovoy V (1996) Magnetic fluids and applications handboock, Begell Home, Inc., New York
14. Shliomis MI (1997) Energy conversion in ferrofluids: magnetic nanoparticles as motors and generators. Phys Rev E 56(1):614–618
15. Ruiz-Suarez C (1999) Sound waves in a magnetic fluid. Phys Rev Focus 3:5
16. Kodama S, Takeno M (1999) Sound-responsive magnetic fluid display, Department of Human Communications, The University of Electro-Communications 1-5-1, Chofugaoka, Chofu-shi, Tokyo, Japan
17. Park GS, Park SH (1999) Design of magnetic fluid linear pump. IEEE Trans Magn 35(5)
18. Kamiyama S (1995) Hydrodynamics of magnetic fluids. Brazilian J Phys 25(2)
19. Seiichi SUDO & Co. (2005) Magnetic fluid devices for driving micro machines. JSME Int J Ser B 48(3)

Chapter 74

Initial Steps Towards Distributed Implementation of M-Urgency

Shivsubramani Krishnamoorthy, Arun Balasubramanian
and Ashok K. Agrawala

Abstract M-Urgency is a public safety system that redefines the way an emergency call is made to the public safety services such as 911. It enables mobile users to stream live video and audio feed from their Smartphone to the local Public Safety Answering Point (PSAP) along with real time location information and other contextual information. In this paper, we present our effort to design and implement a distributed M-Urgency system so as to make the same more efficient and scalable. Our main focus is on (1) handling load balancing of the incoming calls within the servers of the distributed architecture, (2) and handling the case of failures of servers within the same. Simulation of the distributed system and our proposed algorithm for load balancing showed promising results and our approach was also able to handle failures and minimize its impact.

74.1 Introduction

Savannah¹, just 5 years old then, became famous by saving her father's life, when he had a heart attack. Because of her presence of mind she called 911. She became a sensation because this may not always be the case. It is usually a difficult task for the 911 responder to gather information about the situation. A 911 dispatcher has

S. Krishnamoorthy (✉) · A. Balasubramanian · A. K. Agrawala
Departmentt of Computer Science, University of Maryland, College park, MD, USA
e-mail: shiv@cs.umd.edu

A. Balasubramanian
e-mail: arunb@cs.umd.edu

A. K. Agrawala
e-mail: agrawala@cs.umd.edu

to put forth a lot of questions to get the precise idea about the situation, which causes delays and exchange of inaccurate information in a stressful situation and when the time is critical. M-Urgency is the next generation emergency response system that can provide a more efficient and timely service. It is a significant advancement in how emergency calls are made to systems such as 911. M-Urgency enables a person to establish a full-fledged audio and video stream connection with a PSAP (Public Safety Answering Point), to give the dispatcher the most precise idea about the situation. More interestingly, it also enables the dispatcher to forward the stream to a responder such as a squad car, nearest to the location of emergency or the most appropriate one, to ensure a timely service.

The current system encounters various issues considering the centralized approach adopted in designing it. We, thus, propose a distributed architecture, inspired from the p2p approach of object location techniques, to make the M-Urgency system more efficient and scalable. As part of this paper, we intend to take our initial steps towards this effort by verifying our approach and our algorithm through simulations, which we present also.

74.1.1 Objectives

The main objectives of our work described in this paper include:

- Simulation, mainly, of the interface tier of Rover server (described in [Sect. 74.2.1](#)) and partially of the other required components.
- Evaluating our algorithm on the distributed simulation and seeing how the Rover servers balance the load of incoming calls within themselves.
- To evaluate how the servers handle the situation of failure within the distributed architecture.

74.1.2 Overview

The paper is organized in the following manner. [Sect. 74.2](#) describes the current M-Urgency system and also specifically describes the Rover server and the issues in the current system. [Sect. 74.3](#) presents the proposed distributed architecture and our approach for load balancing and handling failures. We discuss the initial results in [Sect. 74.4](#). [Section 74.5](#) discusses the experiments and the performance results. Related work is discussed in [Sect. 74.6](#).

74.2 M-Urgency System

M-Urgency is a public safety system for tomorrow. It enables a user to establish an audio/video stream connection with an emergency dispatcher, to give him the most precise idea about the situation. M-Urgency also enables the dispatcher to forward the stream to a responder(s) nearest to the location of call by a drag and drop action. Three types of applications communicate with Rover 2.0:

- The caller application, running on a mobile phone, facilitates the location and the streaming service.
- The Dispatcher application, running on a desktop, receives emergency calls with the audio/video feeds. It enables grouping of calls, assigning responder(s) to the incident etc. through drag and drop operations.
- The responder application, running on laptops or cell phones receives the forwarded audio/video streams of the assigned caller(s) along with their location information.

The system has been designed comprising of the following components:

- The Rover server that handles the overall administration of the system including initiating and handling user/application connections, user context information management, a/v streaming management etc.
- The streaming server that enables each application to stream audio and video feeds.
- The three kinds of user applications discussed above.

74.2.1 Rover Server

Figure 74.1 provides a high level layout of the Rover architecture. It derives the ideas and is an enhancement to earlier versions [1, 2]. The developed architecture is centered on the information flow between the various sources and the end application or the user. The architecture is laid out in three tiers:

74.2.1.1 Service Tier

The key role of the service engines is to fetch or to translate information to meaningful and useful form for the user. The service we are interested here is the streaming server which enables an application to establish an audio/video streaming service.

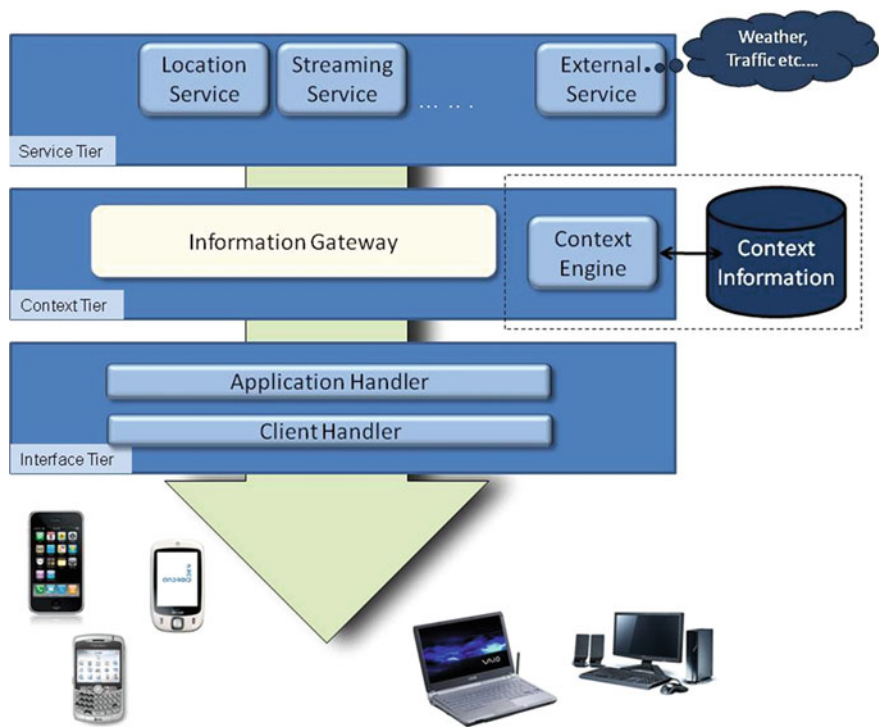


Fig. 74.1 High level Rover architecture

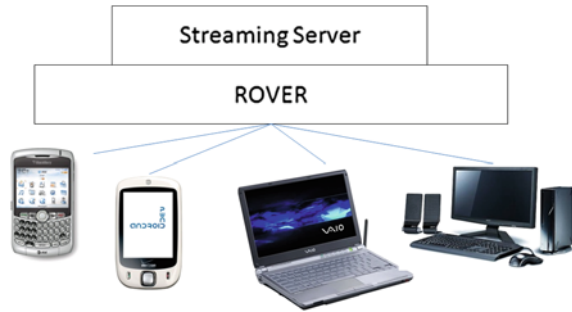
74.2.1.2 Interface Tier

This tier provides a well-defined interface for the applications to associate with the system. It also plays the important role of communicating among many Rovers in the distributed architecture. The communication is done through TCP sockets.

74.2.1.3 Context Tier

We would not really focus on this layer as part of this project. This is the part of the system which makes use of the context of each user to mediate the flow of information from the source to the user. Based on the contexts defined for the user, along with the general context, the information is filtered, adjoined or rearranged.

Fig. 74.2 The existing centralized approach



74.2.2 Issues

Taking into account the critical nature of the system, dealing with emergency situations, we discuss below some of the challenges that we face now or expect to face in the near future as the system is rolled out to the University community:

- *Bulky Data:* For every call made, we deal with real time audio/video stream which consumes high bandwidth. The police department requires good quality video which would facilitate eventual investigations.
- *Occasional, but critical heavy call rate:* Though, practically, not many emergency calls could be made simultaneously, we have learnt from the police department that on particular days (usually Thursday to Saturday) they experience a spike in the call rate. Also, an unfortunate event during a public gathering could produce heavy call rate.
- *Centralized:* The whole system is dependent on a Rover and a Streaming server. Failure of either of them would bring the whole system down; an uncalled for development in an emergency situation.

One of the most significant issues that we have noticed in existing system is that when the number of calls on an average, exceeds 20, we experience issues such as significant delays in the stream, frequent freezing of audio/video and the overall performance degradation in the Dispatcher and Responder applications .

74.3 Distributed Architecture

We propose a distributed architecture for M-Urgency wherein a single Rover server, as depicted in Fig. 74.2, is replicated and organized as a peer to peer network. The Rover server is replicated as depicted in Fig. 74.3.

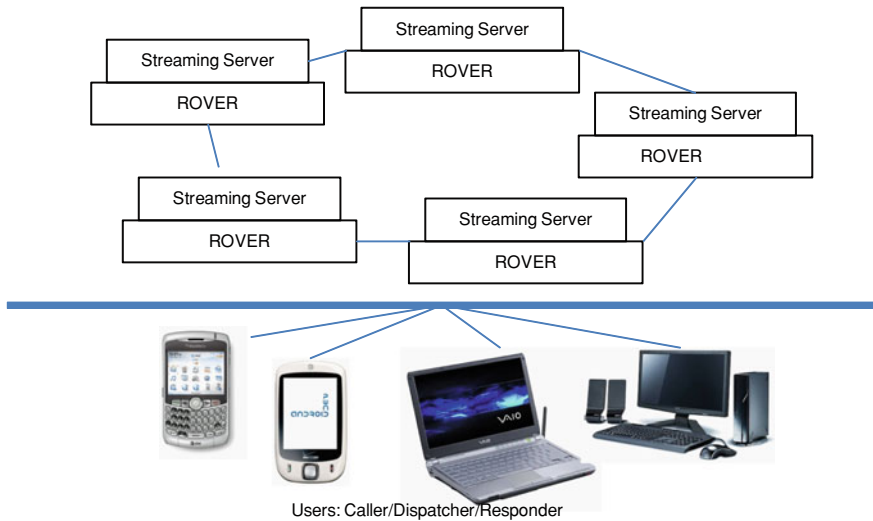


Fig. 74.3 The distributed approach in M-urgency

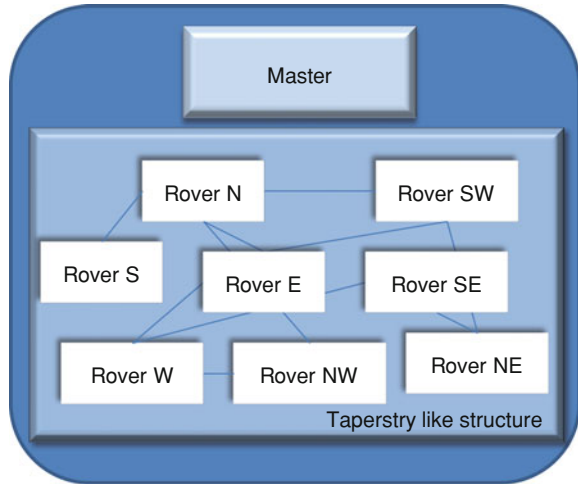
74.3.1 System Design

Our approach derives inspiration from Tapestry [3, 4] to design the distributed system. Each Rover is assigned a unique nodeid that is generated using Secure Hash Algorithm—SHA-1 [5]. Though the Rover servers are organized in a decentralized manner, as in Tapestry, our approach is partially centralized as shown in Fig. 74.4. A *Master node* maintains certain global information, which the Rover servers share within themselves. Following a partially centralized approach helped us reduce the routing cost within the network as it substantially minimizes the number of messages sent between the Rover servers. We discuss the role of the Master node in Sect. 74.3.2

Each Rover server (at the Interface tier) maintains the following:

- *Routing table*—identical to that in Tapestry, the table has the entries of the neighboring nodes. As SHA-1 generates 40 digit hexa-decimal nodeid for each Rover server, the routing table has rows ranging from 0 to 39 and columns ranging from 0 to F. An entry in the m th row and n th column signifies that the nodeid of that particular node shares the same first n digits with the node owning the routing table and the $n + 1$ th digit is m . For example, a node with nodeid as 102ABD... will have 102CD2... entered in the routing table with row = 3 and column = C, because they share the same first three digits (102) and the 4th digit is C.
- *Threshold value*—a limit on how many calls can the particular Rover server handles at a point of time. All the Rover servers maintain the same threshold at any given time and the master intimates each of them to increase the same as and when required.

Fig. 74.4 A partially de-centralized approach



- *Reject list*—once the number of incoming calls goes beyond the threshold, the Rover Server forwards the excessive call loads to other Rover servers. If the recipient Rover server rejects a forwarded call, the same is maintained in the Reject list, until the threshold value is updated globally.

Addition of a Rover server node into the network is the same as in Tapestry as explained in [3]. We have also handled the case of voluntary deletion of the node from the network in a similar way as [3].

74.3.2 Master Node

Though the Rover servers are organized in a decentralized, distributed fashion, we found it important to have a partially centralized approach to maintaining certain global information. This substantially brought down the number of communication messages between the Rover servers, thus reducing the routing cost. The *Master node* is assigned the following responsibilities:

- Direct incoming calls to appropriate Rover server based on the location of the caller
- Re-route calls to other Rover servers when either the appropriate rover server fails or has reached the threshold value.
- Receive regular pings from the Rover servers and keep track of number of calls handled by them
- Maintain the global threshold value wherein when every Rover server reaches the threshold, they are all instructed to increase their respective thresholds

Fig. 74.5 Pseudo code for the master node

```

initialThreshold=n
currentThreshold=initialThreshold
While(true)
{
  Wait for request
  If(request from caller)
  {
    if(appropriate Rover is alive)
      return appropriate Rover Info
    else
      return info about first Rover from the list
      with least clientCount
  }
  If(ping from a Rover)
  {
    update the client count value
    reset corresponding timer
    acknowledge ping
    if(client count of all Rover = currentThreshold)
    {
      currentThreshold+=initialThreshold
      update all Rovers about new threshold
    }
  }
  if(forward request from a rover)
  {
    select first Rover from the list with least clientCount
  }
  if(any of the rover timer has exceeded 1 second)
  {
    Inform all Rovers about the failure
    Remove the entry of the rover information
  }
}

```

Figure 74.5 shows the pseudo code for the Master node. The Master can receive a call request from a caller wherein it calculates the nodeid of the Rover server from the location of the caller and return the information like the IP address and the port number. The caller device then initiates a direct connection with the respective Rover server. In case the concerned Rover has failed, the details of the least busy Rover server are returned.

The Master could also receive a call forward request from a Rover server (details of this situation are discussed in [Sect. 74.4.1](#)). The least busy Rover server is assigned the particular call.

The Master receives the heartbeat ping messages from every Rover server. The Rover servers intimate the Master about the number of calls they are handling at that point of time. If all the Rover servers have reached their threshold, the Master informs them to increase the threshold by the “initial threshold” value. This ensures that the incoming calls are well distributed before each Rover server can begin taking more calls. In case any particular Rover server does not ping the Master for more than 1 s, the Master treats it as a failure and informs all the other Rover Servers about the failure.

74.4 Algorithms

In this section we explain our approach to maintaining the load balance of incoming calls in the distributed system and also handling cases of failures of either of the Rover server nodes or the Master node.

74.4.1 Load Balancing

Figure 74.6a presents the algorithm for each of the Rover server, on how do they handle the incoming calls directed to them and share the overload with the other Rover servers.

When a Rover server receives an incoming video call from a client, it simply accepts it if the number of calls it is currently handling is below the threshold. Otherwise, it forwards the call to the node having the longest matching prefix nodeID in the routing table. If this neighboring node is incapable of accepting calls, it sends a *reject* message after which the current node adds the call to the *Reject List* so that no forwarding attempt is made to the particular node (until the threshold value is increased globally as discussed in [Sect. 74.3.2](#)).

If the host node is unable to find a free neighboring node to forward the call to, it forwards the call to the Master. The Master then allocates a Rover server to handle the call as described in [Sect. 74.3.2](#).

Besides handling calls, the Rover servers also periodically ping the Master by sending heartbeat messages every 100 ms, intimating the Master about the number of calls it is currently handling. Fig. 74.6b provides the pseudo code for this function. After every third ping, the Rover server waits for a response from the Master. If no response is received, the Master is considered failed and all future communication is directed to the Master2 (backup Master). The Rover server also increases its threshold as and when the Master sends the *threshold update* message.

Fig. 74.6 **a** Routing algorithm for each Rover server. **b** Heartbeat ping message to the Master

```

while(true)
{
  If (connection request arrives & current connections <
  threshold)
  {
    Accept call request.
  }
  else if (connection arrives from client)
  {
    flag=false;
    Repeat until request succeeds
    {
      Select longest matching prefix not
      in reject list.
      Request call forwarding.
      If rover is ready,
        forward call.
        flag=true;
      Else if declined,
        Add nodeid to reject list.
    }
    if(flag==false)
      forward to Master
  }
  else if( connection request form rover)
  {
    reply "reject"
  }
}

```

```

masterPingCount=0
while(true)
{
  ping Master updating with client count
  masterPingCounter++

  if(masterPingCounter==3)
  {
    set timer
    Wait for response from Master
    stop timer
    if(changed threshold)
      update threshold.
    masterPingCounter=0
  }
  TimerTimeOut()
  {
    change ip to secondary Master.
  }
}

```

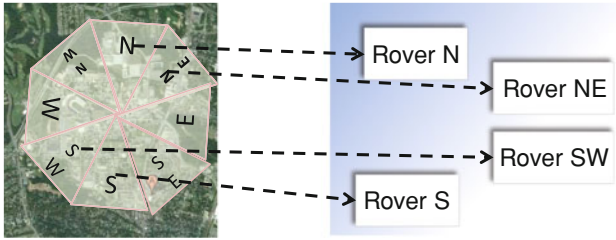


Fig. 74.7 Incoming calls directed to particular Rover servers based on the callers' location

74.4.1.1 Initial Load Balancing

As an attempt for better load balancing, the system has been designed in such a way that the incoming call is directed to a particular Rover server based on the location from where it originates. The Rover servers obtain their *nodeid* by *secured hashing* a string describing the geographical region. For our experiments we considered 8 regions described as S (south), N (North), NW (Northwest) etc. Thus the load balancing of the incoming calls begins right at the incoming stage itself reducing the effort put forth by the individual Rover servers to do it (Fig. 74.7).

74.4.2 Fault Tolerance

We can expect failures at two levels (a) any one of the Rover Servers fails or (b) the Master node fails. We have devised mechanisms for handling both the cases which we discuss in this section.

74.4.2.1 Rover Server Failure

Master Side Recovery—Whenever a Rover server goes down, the master detects it if it doesn't receive a ping from it for 1 s. It updates the local information regarding the particular Rover server and informs all other Rover servers to update their routing tables.

Rover Side Recovery—The other Rover servers remove the entry of the failed node from their routing tables when they are informed by the master.

Client Side Recovery—When initiating a call, the Master automatically redirects the client to another appropriate Rover server. In case of a call that was already being handled by the failed Rover, the client reconnects to the Master. The Master then provides the information of the new Rover that is responsible for handling the client and the connection is established.

74.4.3 Master Failure

To ensure reliability in the event of Master's failure, the distributed system comprises of a secondary Master. The primary Master periodically sends its bookkeeping statistics (like threshold and client-counts) to the secondary Master. This ensures that at almost all times the secondary Master has the necessary information to take over from the primary.

Rover Side Recovery—When a master node exits or goes down abruptly, the Rovers detect Master's failure by a timeout in not receiving an acknowledgement for the pings and update their Master IP to the secondary Master.

Master Side Recovery—When the secondary Master receives a ping from any of the Rovers, it realizes its role as the master and becomes the new Master for the system.

Client Side Recovery—When a client detects that the primary Master has gone down, it automatically tries to connect to the secondary Master for any further communications.

74.5 Experiments and Results

As mentioned before, our objective at this point was to simulate the *Interface Tier* of the Rover server and see how they organized themselves in the distributed setup. Our main focus, in this work, is on how the incoming calls traffic is handled and Rover servers balance the load; and how quickly does the system recovers to a consistent state in case of a failure.

74.5.1 Load Balancing

We considered two scenarios—(1) when the calls come in a random order from the different regions and (2) a less probable but more critical case, where all the calls originate from the same region (for example, an unfortunate incident occurs in a public place and many callers call into report the same).

Calls from same region—We simulated 119 calls to originate from the same region, so that they are all directed to Rover S (south). We monitored the load balance at intervals after 40 calls, 97 calls and finally 119 calls. We performed the same experiment setting the initial threshold value as 5 and 3.

Tables 74.1 and 74.2 show the distribution of the calls within the eight Rover servers. It is clear from the tables that our algorithm produced a good distribution of the incoming calls. Instead of Rover S handling all the 119 calls, the load was well balanced with each Rover server handling a maximum of only 15 calls. The difference in the threshold did not show a significant difference in terms of

Table 74.1 Distribution of incoming calls within the Rover servers (initial threshold = 5)

| | | | | | |
|-----------------|----|----|----|----|------------------|
| After 40 Calls | S | N | NW | E | Routing messages |
| | 5 | 5 | 5 | 5 | 65 |
| | W | SE | NE | SW | |
| | 5 | 5 | 5 | 5 | |
| After 97 Calls | S | N | NW | E | Routing messages |
| | 15 | 15 | 15 | 12 | 132 |
| | W | SE | NE | SW | |
| | 10 | 10 | 10 | 10 | |
| After 119 Calls | S | N | NW | E | Routing messages |
| | 15 | 15 | 15 | 15 | 172 |
| | W | SE | NE | SW | |
| | 15 | 15 | 15 | 14 | |

distribution of the calls, but the number of routing messages between the Rover servers and the Master was observed to be less in case with threshold = 3.

Calls from different regions—We, then, simulated 43 calls to originate from the eight different regions as shown in Fig. 74.8. The calls were made to originate one by one, in a round robin fashion, beginning from the North region.

Tables 74.3 and 74. 4 show the distribution of the calls within the 8 rover servers.

One interesting thing to notice here is the number of routing messages. The number of routing messages is significantly less than when the calls originated from the same region. This shows that our *initial load balancing* approach was effective. Since the calls are directed to their respective Rover servers, based on the location of the caller, the number of routing messages sent between the Rover servers is significantly low as the Rovers automatically receive the incoming calls in a distributed manner. We can compare the number of messages after 40 calls in Table 74.1 with the number of messages in table 74.3 (after 43 calls). With initial load balancing, the number of messages is about 62 % less in this case.

74.5.2 Fault Tolerance

We conducted experiments with three scenarios. In scenario 1, we brought down a Rover and observed the behavior of the system. We noticed that the system converged to a consistent state within 2 s (i.e. the other Rovers updated the routing information accordingly).

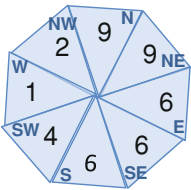
In scenario 2, we brought down a Master and observed that the Rovers detected the failure and contacted the secondary Master within 2.5 s.

In scenario 3, we brought down a master and a rover together and observed that the system converged to a consistent state within 4.5 s. We noticed that calls

Table 74.2 Distribution of incoming calls originating from the same region (initial threshold = 3)

| | | | | | |
|-----------------|----|----|----|----|-------------------------|
| After 40 Calls | S | N | NW | E | Routing messages 50 |
| | 6 | 6 | 6 | 6 | |
| | W | SE | NE | SW | |
| | 4 | 4 | 4 | 4 | |
| After 97 Calls | S | N | NW | E | Routing messages 132 |
| | 13 | 12 | 12 | 12 | |
| | W | SE | NE | SW | |
| | 12 | 12 | 12 | 12 | |
| After 119 Calls | S | N | NW | E | Routing messages 163 |
| | 15 | 15 | 15 | 15 | |
| | W | SE | NE | SW | |
| | 15 | 15 | 15 | 14 | |

Fig. 74.8 Number of calls originating for each region



originating from client, within the above mentioned period of recovery time, were handled successfully by the Master by forwarding it to another active Rover.

74.5.3 Inferences

We list a few inferences that we have made from our experimental results:

- Our approach does produce a good level of distribution of incoming calls within the Rover servers, thus enabling good load balancing.
- The initial load balancing helped us significantly bring down the number of routing messages, thus reducing cost.
- We also observed in our experiments that the performance of the system depends on:
 - Number of Rover server nodes in the network
 - The way the network is formed (with links between the Rovers) based on the nodeid assigned to them
 - Number of neighbors a node, that receives significantly high number of calls, has.
 - The initial threshold value

Table 74.3 Distribution of incoming calls originating from different regions (initial threshold = 5)

| S | N | NW | E | Routing messages |
|---|----|----|----|------------------|
| 5 | 6 | 5 | 5 | 25 |
| W | SE | NE | SW | |
| 5 | 5 | 7 | 5 | |

Table 74.4 Distribution of incoming calls originating from different regions (initial threshold = 3)

| S | N | NW | E | Routing messages |
|---|----|----|----|------------------|
| 6 | 6 | 3 | 6 | 17 |
| W | SE | NE | SW | |
| 6 | 6 | 6 | 4 | |

74.6 Related Work

A lot of emphasis has already been made, in the literature, regarding the significance of a distributed architecture over a centralized approach. Many have exhibited the advantages of a distributed approach like in [6, 7].

The first generation of peer-to-peer applications, like Napster and Gnutella had restricting limitations such as a central directory for Napster and scoped broadcast queries for Gnutella limiting scalability [8]. To address these problems a second generation of P2P mechanisms were developed like Tapestry [3, 4], Chord [9], Pastry [10], and CAN [11]. Our approach derives its approach from that of Tapestry. Tapestry is an extensible infrastructure that provides decentralized object location and routing focusing on efficiency and minimizing message latency. This is achieved since Tapestry constructs locally optimal routing tables from initialization and maintains them in order to reduce routing stretch. Chord is a protocol and algorithm for a peer-to-peer distributed hash table. A distributed hash table stores key-value pairs by assigning keys to different computers (known as “nodes”); a node will store the values for all the keys for which it is responsible. Chord specifies how keys are assigned to nodes, and how a node can discover the value for a given key by first locating the node responsible for that key. Pastry is an overlay and routing network for the implementation of a distributed hash table similar to Chord. The protocol is bootstrapped by supplying it with the IP address of a peer already in the network and from then on via the routing table which is dynamically built and repaired. Because of its redundant and decentralized nature there is no single point of failure and any single node can leave the network at any time without warning and with little or no chance of data loss.

A number of approaches for distributed video streaming have been discussed in the literature [12– 14], addressing issues of source and channel coding, implementation of transport protocols, or modifying system architectures in order to deal with delay, loss, and time-varying nature of Internet. Our focus, in this paper,

is mainly on how to bring in a p2p approach in handling and routing of video streams.

Cool streaming [8] is a P2P data driven Overlay distributed users. Notable features of the protocol network for live media streaming that achieves good streaming quality for globally include its intelligent scheduling algorithm that copes well with the bandwidth differences of uploading clients and thus minimises skipping during playback, and its swarm-style architecture that uses a directed graph based on gossip algorithms to broadcast content availability.

Though the critical parts of our system are totally decentralized, we drew inspiration of having a partially centralized approach from [15, 16]. Though they adopted this approach in a different context, it proved beneficial for us as explained in [Sect. 74.3](#).

74.7 Future Work

The results we have obtained are very promising and encouraging for us to extend this work in many dimensions.

- We would like to assess the performance of the algorithm for different network formations and with more threshold values.
- We would also want to assess our algorithm with different patterns of incoming calls
- Though we are confident that our system can handle multiple Rover failures, we are yet to verify the same experimentally.
- Our final goal is to create the distributed architecture and apply our algorithm on the actual M-Urgency system and verify the performance
- It would be interesting to see how the distributed approach actually addresses the problem of video latency and freezing.

74.8 Conclusion

We attempted to take our initial steps towards creating a distributed architecture for M-Urgency, a next generation public safety and emergency application. Our main focus was to simulate and analyze how our distributed system handles the incoming call traffic efficiently and the events of failures. The performance of our architecture and our algorithms are promising and encouraging. We plan to implement the same in the actual M-Urgency system and evaluate the performance.

References

1. Almazan CB (2010) Rover: architectural support for exposing and using context. Ph.D. Thesis. University of Maryland, College Park. College Park, Maryland, United States
2. Almazan CB, Youssef M, Agrawala AK (2007) Rover: an integration and fusion platform to enhance situational awareness. In: 26th IEEE international performance computing and communications conference, pp 582–587
3. Zhao BY, Huang L, Stribling J, Rhea SC, Joseph AD, Kubiawicz JD (2004) Tapestry: a resilient global-scale overlay for service deployment. *IEEE J Sel Areas Commun* 22(1):41–53
4. Zhao BY, John Kubiawicz, and Anthony D. Joseph Tapestry: an infrastructure for fault-tolerant wide-area location and routing. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.24.7439&rep=rep1&type=pdf>
5. Robshaw MJB (1995) MD2, MD4, MD5, SHA and other hash functions. RSA labs, vol 4.0. Technical report, TR-101
6. Rosenblatt JK (1995) DAMN: a distributed architecture for mobile navigation. : AAAI technical report SS-95-02
7. Diot C, Gautier L (1999) A distributed architecture for multiplayer interactive applications on the Internet. *IEEE Netw* 13(4):6–15
8. Saroiu S, Gummadi KP, Gribble SD (2003) Measuring and analyzing the characteristics of Napster and Gnutella hosts. *Springer J Multimed Sys* 9(2):170–184
9. Stoica I, Morris R, Karger D, Kaashoek MF, Balakrishnan H (2001) Chord: a scalable peer-to-peer lookup service for internet applications. In: proceedings of the ACM SIGCOMM '01, Aug 2001
10. Rowstron A, Druschel P (2001) Pastry: scalable, distributed object location and routing for large-scale peer-to-peer systems. In: proceedings of IFIP/ACM international conference on distributed systems platforms (middleware), Nov 2001
11. Ratnasamy S, Francis P, Handley M, Karp R, Shenker S (2001) A scalable content-addressable network. In: proceedings of ACM SIGCOMM, Aug 2001
12. Nguyen T, Zakhor A (2004) Multiple sender distributed video streaming. *IEEE Trans Multimed* 6(2):315–326
13. Mungee S, Surendran N, Schmidt DC (1999) The design and performance of a CORBA audio/video streaming service. System Science 1999. In: Proceedings of the 32nd annual Hawaii international conference on HICSS-32, vol Track 8, p 14
14. S.-H. Gary Chan. 2001. Distributed servers architecture for networked video services. *IEEE/ACM Trans. Netw.* 9, 2 (April 2001), 125-136.
15. Le Pape C (1990) A combination of centralized and distributed methods for multi-agent planning and scheduling. In: Proceedings of the IEEE international conference on robotics and automation, vol 1. pp 488–493, 13–18 May 1990
16. Ghemawat S, Gobioff H, Leung S-T (2003) The Google file system. *SIGOPS Oper Sys Rev* 37(5):29–43

Chapter 75

Optimal Selection of Components in Fault Detection Based on Principal Component Analysis

Patricia Helen Khwambala

Abstract Selection of the optimal number of principal components (PCs) in fault detection using principal components analysis (PCA) is considered in this paper. The focus is on the relationship between the sensitivity to a particular fault and the number of PCs retained. The selection method that is based on signal to noise ratio of the fault detection (known as fault SNR) is compared to the cumulative percent variance (CPV) and the Scree methods of determining the optimal number of PCs to be retained for fault detection based on PCA. The SNR fault detection method shows different dependencies on the number of PCs for different kinds of faults. The number of PCs that gives the maximum sensitivity is easily determined for sensor faults by examining the fault SNR. If apriori data is available as operational data that has been measured during faulty conditions, then optimization of the number of PCs for the process fault is possible. The methods are applied to a thermal system.

75.1 Introduction

Numerous methods exist for selecting the number of PCs when the PCA technique is used for fault detection. Some of the popular methods are: Scree plot, eigenvalue limit, cumulative percent variance, cross validation, variance of construction error, and there are many others [1] and [2]. The Scree plot and eigenvalue limit are based on the concept that components with small eigenvalues are not important for modeling the data that is under investigation. In the cumulative percent variance

P. H. Khwambala (✉)

Department of Electrical Engineering, University of Cape Town, Rondebosch, South Africa
e-mail: patriciakhwambala@yahoo.co.uk

(CPV) method, the minimum model dimension that can express a substantial part of the total variance of the data is selected. The cross-validation method [3] and [4] uses part of the training samples for model construction; the remaining samples are compared with the prediction by the model and when the prediction residual sum of squares (PRESS) is less than the residual sum of squares of the previous model, the new component is added to the model.

Several comparative studies have been conducted on methods for determining the number of PCs for fault detection. In [5] four methods were compared using the Tennessee Eastman data. In [6] eleven methods were applied and it was concluded that the VRE criterion is preferable, because amongst other reasons it is simple to apply. Apart from these comparative studies, [1] and [2] also mentioned that fault detection ability depends on the number of PCs retained in the PCA model.

Most of the previous studies studied were largely based on the concept that the best solution is given by the number of components that represent the ‘true’ dimension of the system, which is usually unknown. In this paper the main focus is on the sensitivity of the prediction to a fault, which is the most important and practical issue in fault detection. T^2 statistic, which is used for fault detection, show different behaviors with different numbers of PCs retained that gives the maximum sensitivity of fault detection depending on the type of fault. The signal to noise ratio of fault detection is considered as an indicator of the fault detection ability of the PCA model. In the case of process faults, optimization of the number of PCs is possible if a priori information of the fault is available, as data measured that is during faulty conditions. In cases where data in faulty conditions is not available, monitoring many models with various numbers of PCs in parallel is commonly used but this approach is not the focus in this paper.

75.2 Principal Component Analysis

PCA is a statistical technique that transforms a set of correlated original data to an uncorrelated data set that represents most of the information of the original data. Let $\mathbf{X} = \mathbf{R}^{n \times m}$ denote the original data matrix with n samples and m variables. In the PCA method the original data is first scaled to produce a matrix \mathbf{X} with zero mean and unit variance [5]. Then based on a singular value decomposition (SVD) algorithm, the matrix \mathbf{X} of the original data can be decomposed as follows;

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \quad (75.1)$$

where $\mathbf{T} \in \mathbf{R}^{n \times l}$ and $\mathbf{P} \in \mathbf{R}^{m \times l}$ are the score matrix and the loading matrix respectively, while \mathbf{E} is the error matrix. The PCA transforms the original set of m variables to a reduced set of l principal components. PCA can be regarded as a classical linear dimension reduction technique and the number of PCs is commonly determined by using CPV and the Scree plot method.

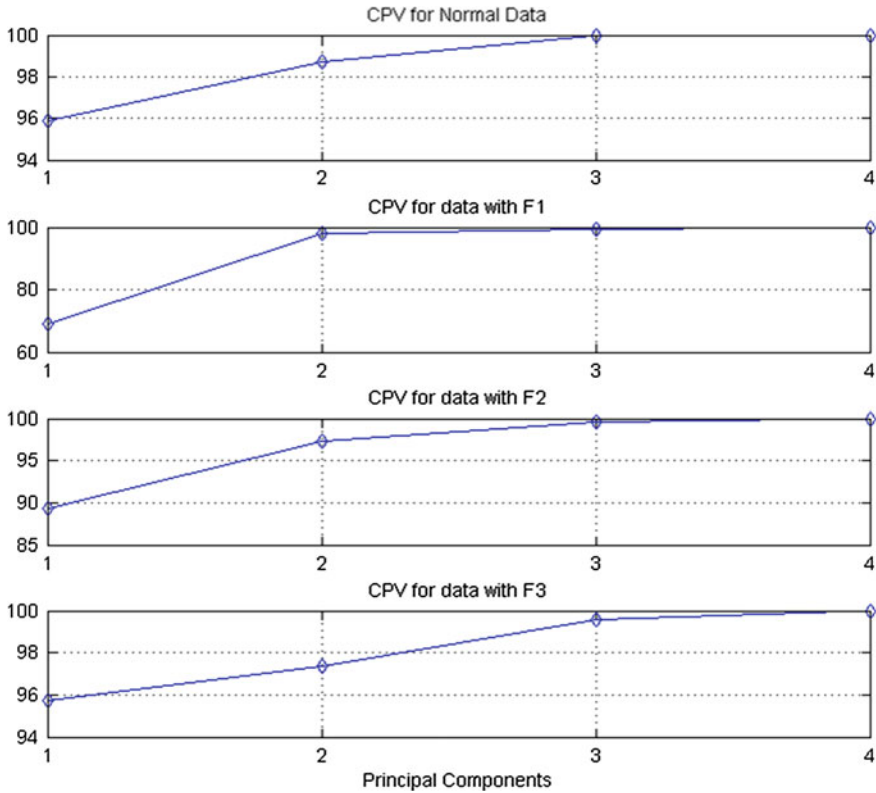


Fig. 75.1 CPV plot for normal data and the data with faults

The CPV method is given as follows:

$$\frac{\sum_{i=1}^a \lambda_i}{\sum_{i=1}^m \lambda_i} \times 100 \% \geq 85 \% \quad (75.2)$$

where λ_i is the variance of the score vector and a is the number of PCs that are retained. When CPV is larger than 85 %, the corresponding number a of PCs is determined.

The Scree method involves plotting the eigenvalues against the number of principal components. These eigenvalues of matrix X are arranged in ascending order of magnitude and linked with a line. The number of PCs to be retained is decided by the value at the knee point of this plot (as shown in Fig. 75.1). Specifically the point where the slope joining the plotted points is large to the left of the particular PC and not to the right of the same PC point identifies the ideal number of PCs to retain for the given data set.

In PCA based fault detection, statistics and their control limits need to be established to determine whether a process is within the control limit or not. Common statistics include the Q -Statistic, which indicates the degree of deviation of each sample from the model, and the T^2 statistic, which reflects on variations with PCA model [7]. The two statistics and corresponding control limits are given as follows;

75.3 Q-Statistics

The residual space which corresponds to the smaller singular values can be monitored efficiently by the use of Q -statistic [6, 8 and 9].

$$Q = \mathbf{r}^T \mathbf{r} \quad (75.3)$$

$$\text{Where } \mathbf{r} = (I - \mathbf{P}\mathbf{P}^T)\mathbf{x} \quad (75.4)$$

Residual vector $\mathbf{r} \in \mathbf{R}^{n \times 1}$ is the projection of the observation X into the residual space and \mathbf{x} is the observation row vector. It has been found that the Q -statistic measures the total sum of variations in the residual space. The Q -statistic is not over sensitive to the inaccuracies in the smaller singular values [3] and [9]. The Q -statistic is sometimes known as the Squared Prediction Error (SPE), which is the squared 2—norm measuring the deviations of the observations to the low dimensional PCA presentation.

The distribution of the Q -statistic can be approximated as;

$$Q_\alpha = \theta_1 \left[\frac{h_0 C_\alpha \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{\frac{1}{h_0}} \quad (75.5)$$

$$\text{where } \theta_i = \sum_{j=a+1}^n \delta^{2i}, \text{ where } \delta^{2i} = \lambda_i \quad (75.6)$$

$$\text{and } h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2} \quad (75.7)$$

Constant C_α is the normal deviate corresponding to the $(1 - \alpha)$ percentile. The Q -statistics measures the random variations of the process while the threshold Q_α is applied to define the normal variations of the random noise, and any violation of the threshold can indicate that the random noise has significantly changed, hence this is used to detect faults.

75.4 T^2 Statistics

This is another technique for fault detection when PCA is used simultaneously for both dimensional reduction and detecting faults. The calculation of the T^2 statistic enhances fault detection when PCA is used. For observation vector \mathbf{x} and the singular value decomposition given as $\mathbf{\Lambda} = \sum^T \sum$ (singular values of the matrix \mathbf{X}) which is invertible, T^2 statistic can be calculated as;

$$T^2 = \mathbf{x}^T \mathbf{v} \left(\sum^T \sum \right)^{-1} \mathbf{v}^T \mathbf{x} \quad (75.8)$$

where \mathbf{v} is taken from the covariance matrix \mathbf{S} for the data set \mathbf{X} and is a single value of \mathbf{V} in the process of dimensional reduction, as in $\mathbf{S} = \mathbf{V}^T \mathbf{V}$ and $\delta_i^2 = \sum^T \sum$ which are equal to the diagonal elements of the matrix $\mathbf{\Lambda}$ [7].

When the number of observation variables m is large and the amount of data available is relatively small, this is like the number of rows in matrix \mathbf{X} is large and the number of columns is relatively small, the T^2 statistic tends to be an inaccurate presentation of the in-control process behavior [7] and [10]. This is more in the loading vector directions corresponding to the smaller singular values. These inaccuracies may have a big effect on the calculation of the T^2 statistic. This is due to the square matrix $\mathbf{\Lambda}$ been inverted as indicated in Eq. (75.8) above.

As a way of avoiding this problem, the loading vectors (that form the loading matrix \mathbf{P}) associated with the largest singular value should be used for the calculation of the T^2 statistic.

Including the \mathbf{P} matrix, the first a loading vectors with number of largest singular values ($a = m - l$); the T^2 statistic for the data becomes [8];

$$T^2 = \mathbf{x}^T \mathbf{P} \sum_a^{-2} \mathbf{P}^T \mathbf{x} \quad (75.9)$$

where \sum_a matrix has the first a rows and columns of \sum . As it is T^2 only measures the variations in the score space. If the actual mean and covariance are known as;

$$T_\alpha^2 = X_\alpha^2(a) \quad (75.10)$$

with the known actual covariance matrix from the sample covariance matrix, the T^2 statistic threshold becomes [6];

$$T_\alpha^2 = \frac{(a(n-1)(n+1))}{(n(n-a))} F_\alpha(a, n-a) \quad (75.11)$$

However to come up with the outliers in the training set, the threshold becomes [6]:

$$T_{\alpha}^2 = \frac{(n-1)(n-1)(a/(n-a-1))F_{\alpha}(a, n-a-1)}{n(a/(n-a-1))Fc(a, n-a-1)} \quad (75.12)$$

The T^2 in Eq. (75.9) is not affected by the inaccuracies in the smaller values of the singular values; hence it better represents the normal behavior of the process (i.e. the part that has the signal information). Now this can be used for fault detection when compared to T^2 for Eq. (75.8). This is because the T^2 statistic is the measurement of the systematic variations of the process, and any violation of the threshold would indicate that the fault occurred because the systematic variations are above the control limit.

75.5 Application for Laboratory Thermal System

75.5.1 Process Description

The thermal consists of two hot-air blowers and two temperature sensors that form an interactive multivariable dynamic system. By adjusting heat transfer based on the input voltage, it is possible to alter the output temperature.

The hot-air blowers produce a column of air transfers that the heat in air to the temperature sensors. The air is cooled down or heated up depending on the input voltage applied to the corresponding power electronic circuit. An inverse relationship existed between heat and voltage in that an increase in voltage produced a decrease in heat. The temperature is sensed by two electronic elements that give a voltage proportional to the temperature around the sensors.

The two sensors are positioned in such a way that the heat from heater number 1 affects sensor 1 directly and sensor 2 partly. Similarly heater number 2 affects sensor 2 directly and sensor 1 partly. This leads to interaction between the two heater systems and the plant forms a multi-input-multi-output (MIMO) dynamic system. The input signal to the hot-air sub-system was constrained to a range of 4.5 and 6.5 V to ensure safe operation of the heater coils. Once controlled the set point for the temperatures was stepped by 1 V.

75.5.2 Selecting the Number of Optimal PCs Based on CPV

The cumulative variance percent (CPV) method, as stated above, is one of the many methods used for determining the number of PCs to be retained when using PCA [6].

This method selects the percentage of the total variation which is desired for the selected PCs [9]. The user is at liberty to choose the percentage that seems reasonable, as long as it's above 85 %; this could be 85, 95 or 99 % of the total variation.

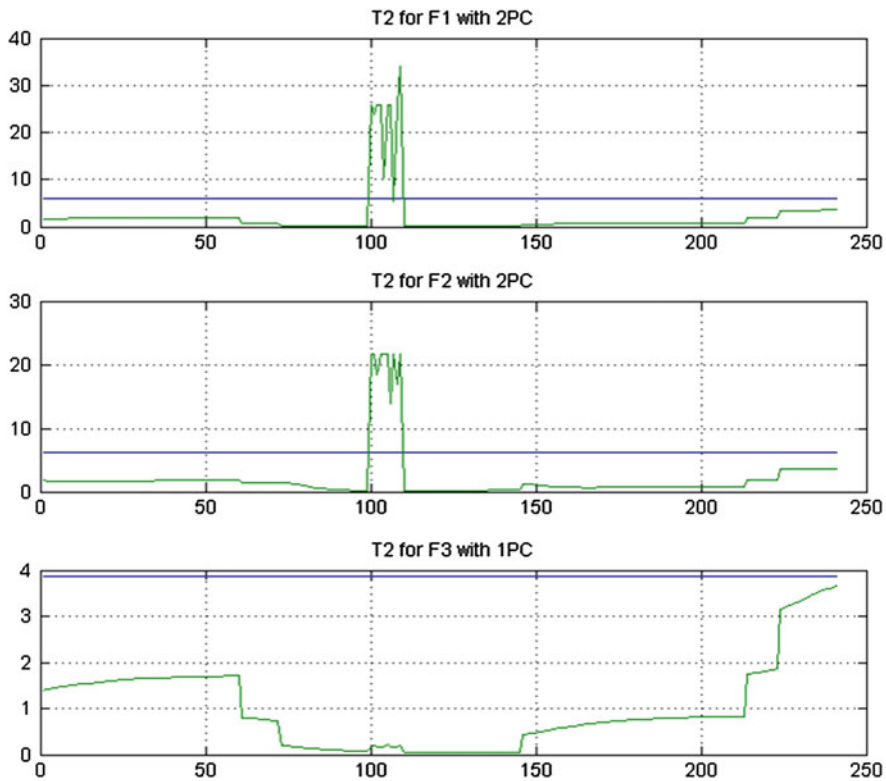


Fig. 75.2 T^2 Statistics for all samples for F1, F2 and F3

Three types of faults were generated in such a way that Fault F1 has voltages over 100 % above the normal output voltage, Fault F2 has voltages 100 % below the normal output voltage and Fault F3 has voltages less than 20 % above the normal output voltage. All three faults were introduced to 10 samples from 100 sample to 110th sample. This was to cater for all fault types that can be detected by these PCA fault detection techniques.

$$CPV = 100 \left[\frac{\sum_i^a \lambda_i}{\sum_i^m \lambda_i} \right] \% \quad (75.13)$$

Figure 75.1 has the cumulative variance percentage for the normal data, the data with fault F1, data with fault F2 and data with fault F3. The author considered 95 % to be the cumulative variance for determining the number of PCs to be retained for dimension reduction for the normal data. The choice 95 % was made in order to get most of the information of the signal into the score space. The same CPV was used for the determination of PCs for both dimension reduction and fault detection for the data with faults.

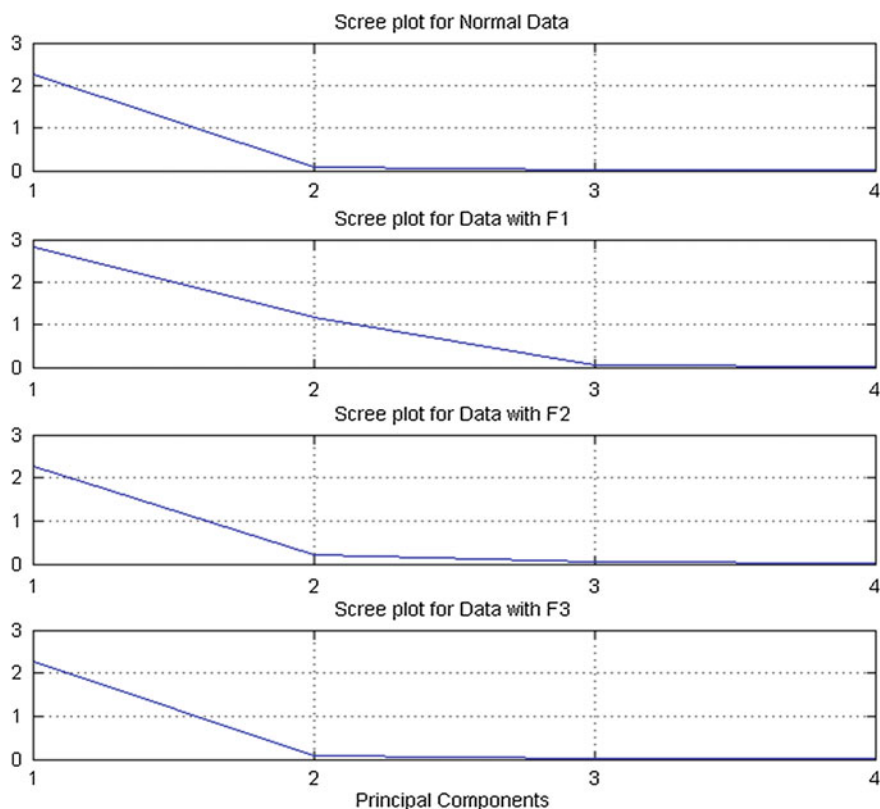


Fig. 75.3 Scree test plots for normal data, F1, F2 and F3 data

It is observed from the plots in Fig. 75.1 that with the 95 % threshold CPV determined that one PC was to be retained for normal data and fault F3, and two PCs for both fault F1 and fault F2.

These numbers of retained PCs were then used in fault detection with the calculation of T^2 statistics values. The results are plotted in Fig. 75.2.

From these three plots it was observed that fault F3 could not be detected with the retention of one PC as determined by the CPV method. It worked for the dimension reduction but was not feasible for fault detection. The T^2 statistics plots in Fig. 75.2 shows the detection of fault F1 and fault F2 but not fault F3. Thus it can be concluded that fault F3 was not projected in the score space which T^2 statistics used, this is because the fault was in the residual space projected by the least significant eigenvalues which are not used in this technique. This fault was in the part of the signal that had noise information, hence could not be detected.

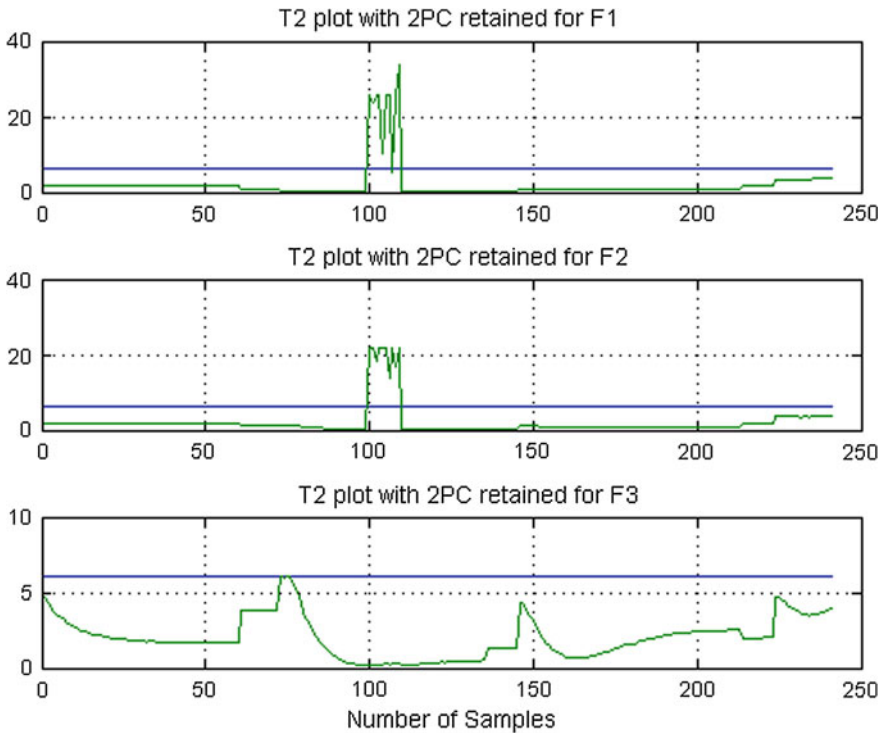


Fig. 75.4 T^2 statistics values for F1, F2 and F3 with 2PCs retained

75.5.3 Scree Test Method Used for PCs Selection

This is one of the most commonly used methods [1] which has been explained above. This method was applied and the figures below show the results.

Figure 75.3 gives the number of PCs to be retained for the normal data, fault F1, fault F2 and fault F3.

It can be observed that for normal data, F2 and F3, two PCs are determined by this method to be retained, this works both for reduction of the dimension of the system data and for fault detection. In the case of F1 it is three PCs that are determined to be retained for fault detection and dimension reduction.

The T^2 statistics plots follow in Fig. 75.4. This technique of fault detection is applied with the number of PCs determined by the Scree method and the result shows that faults F1 and F2 could be detected and fault F3 could not.

Fault F3 could not be detected with the two PCs that were retained after being determined by the Scree method. The result shows that this method is not right all the times when it comes to determining the optimal number of PCs to be retained for fault detection when using PCA. The same reason noted for the CPV method and fault F3 can be applied for the failure of the Scree method to detect fault F3, as

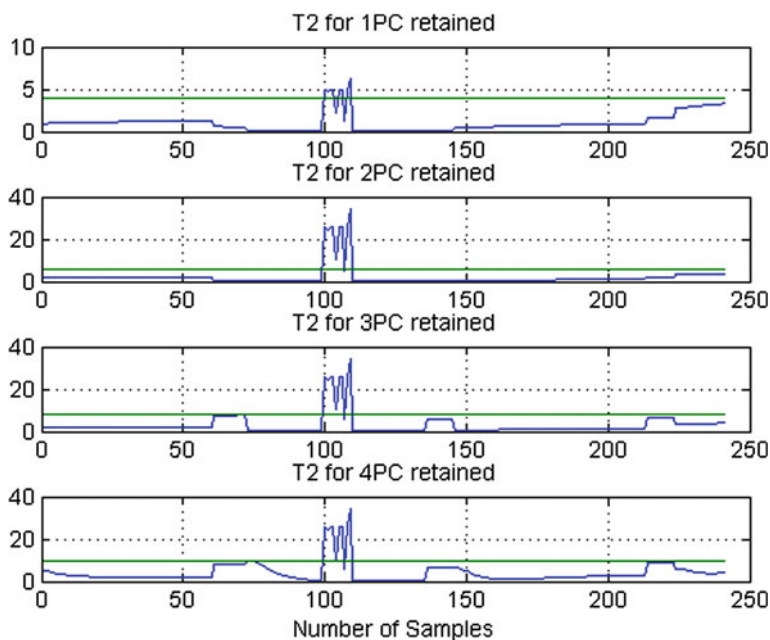


Fig. 75.5 T^2 statistics plot for F1 of the thermal data for various PCs retained

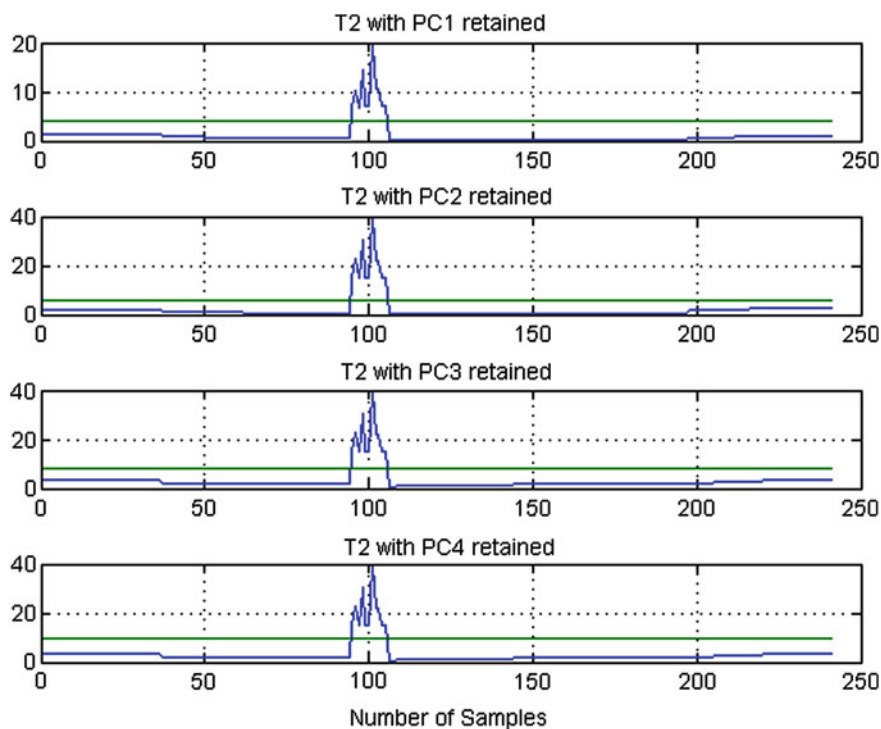


Fig. 75.6 T^2 statistics plot for F1 of the thermal noisy data for various PCs retained

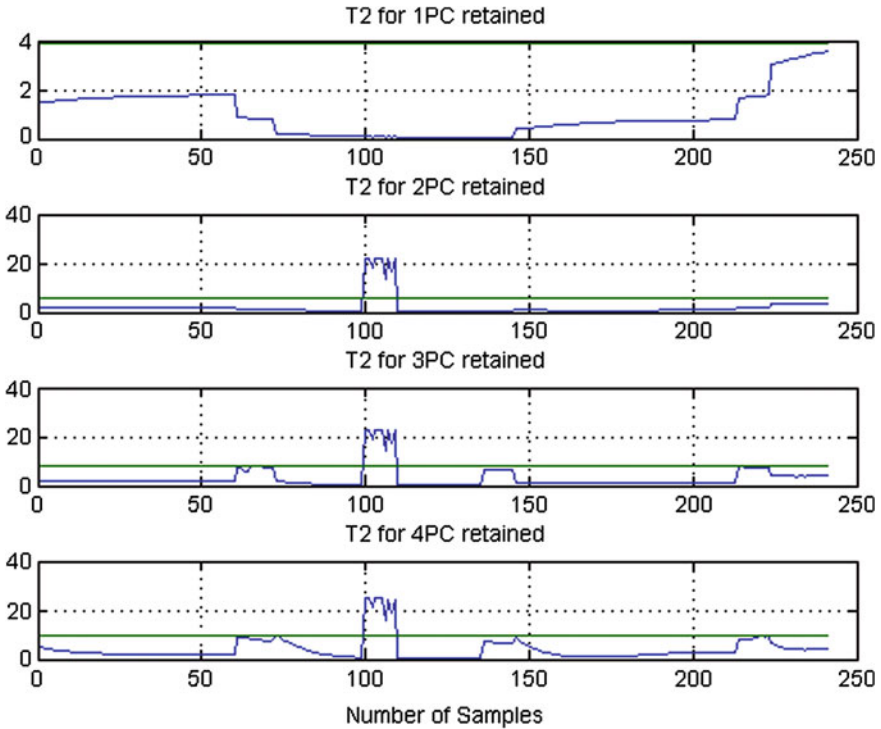


Fig. 75.7 T^2 statistics plot for F2 of the thermal data for various PCs retained

the fault was projected to the noise part of the signal no considered by the T^2 statistics technique.

75.5.4 Selecting the Number of Optimal PCs Based on SNR

This is the most recent method of those considered for determining the number of PCs to be retained when focusing mainly on fault detection [1]. This method can only be applied to data for which prior information exists and the results give the number of PCs which is more sensitive to which particular fault [6].

The method is based on the proposed fault signal to noise ratio (fault SNR). As stated above, it indicates the relationship between the sensitivity of fault detection and the number of PCs retained [6].

Figure 75.5 has the T^2 statistics values plotted with various retained number of PCs determined for fault F1. Figure 75.6 has the plot for fault F1 noisy data of (0.5 dB). The significance of the added noise was used to set the threshold at 0.5 dB rather than a lower value.

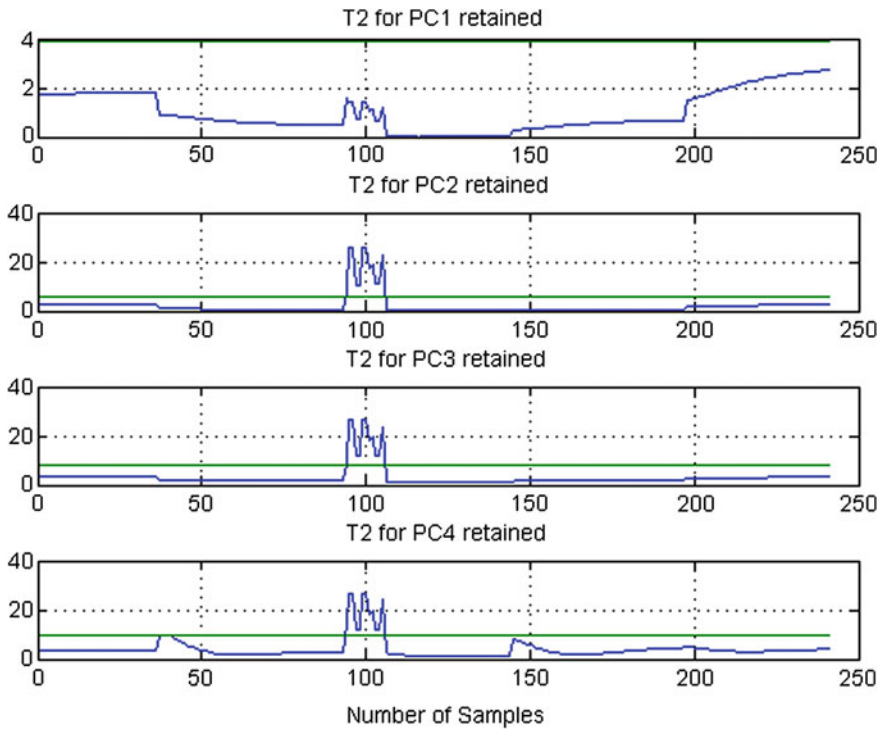


Fig. 75.8 T^2 statistics plot for F2 of the thermal noisy data for various PCs retained

It can be seen that in both data plots, fault F1 could be detected with the retention of all the various number of PCs.

One has to consider the sensitivity of the fault with a particular number of PCs retained and at the same time consider the dimension reduction of the data. In this case the most optimal number of PCs to be retained should be two for fault F1, to ensure a high sensitivity for fault F1 detection.

Fault F2 could not be detected by retaining one PC, but with the rest of the PCs as it shows in Figs. 75.7 and 75.8. The sensitivity of the fault F2 detection is almost the same for the three combinations of the PCs that could detect the fault, making it more plausible to consider two PCs as the optimal number of PCs to be retained.

The same procedure was done for fault F3 and the results given in Figs. 75.9 and 75.10.

The two plots show that fault F3 could be detected only with the retention of all (four) of the PCs as the rest of the combinations failed to detect fault F3. The case was different with the noisy data, the retention of three PCs and four PCs could detect fault F3 and not two PCs or one PC. The introduction of more noise in fault F3 data pushed the fault to the score space hence the detection by the retention of the 3PCs. For fault sensitivity factor, one could retain four PCs to detect fault F3.

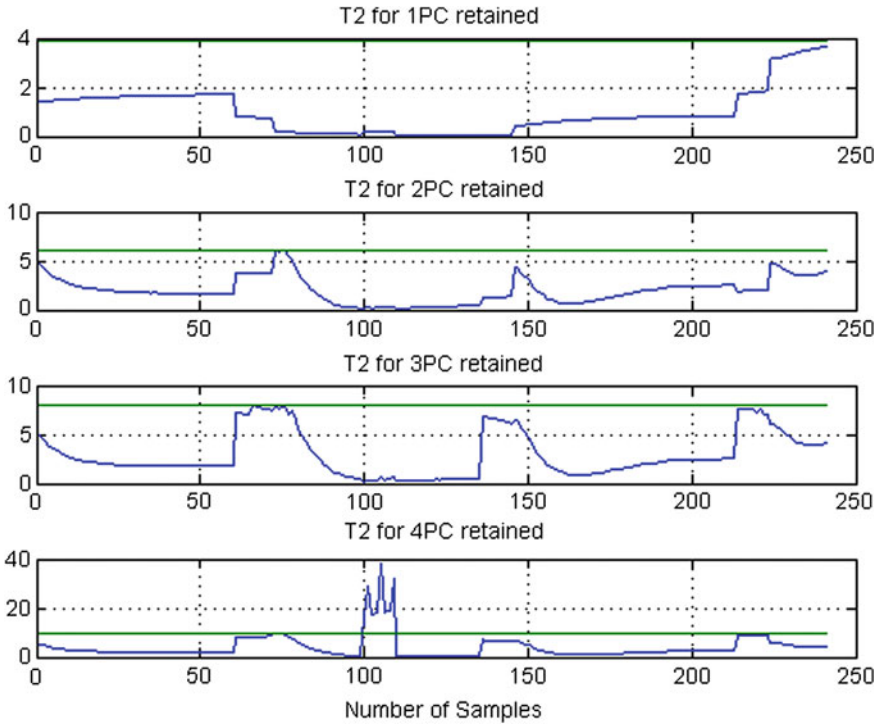


Fig. 75.9 T^2 statistics plot for F3 of the thermal data for various PCs retained

Three PCs could be retained for the same fault F3 in the noisy data to serve the dimension reduction factor which is the most important factor in PCA.

75.6 Conclusions

In the fault detection method based on PCA, the number of PCs retained greatly affects the ability of the method to detect faults. In this paper it has been shown that the number of PCs which maximizes fault detection depends on the kind of fault that is experienced. Examining the fault SNR the number of PCs can be determined for detecting faults on the thermal system process. If a reference data set of a faulty plant operation is available, optimization of the number of PCs with fault SNR for the process faults can be done. This is good for plant operators if a certain kind of faults occurs repeatedly in the plant, because this method enables sensitive detection of the fault from its second appearance. Plots of the threshold serves as a control limit to demonstrate that the number of PCs determined by fault SNR is the most optimal. This helps in the indication of the sensitivity to the particular fault. Furthermore this fault SNR method is compared to CPV and Scree

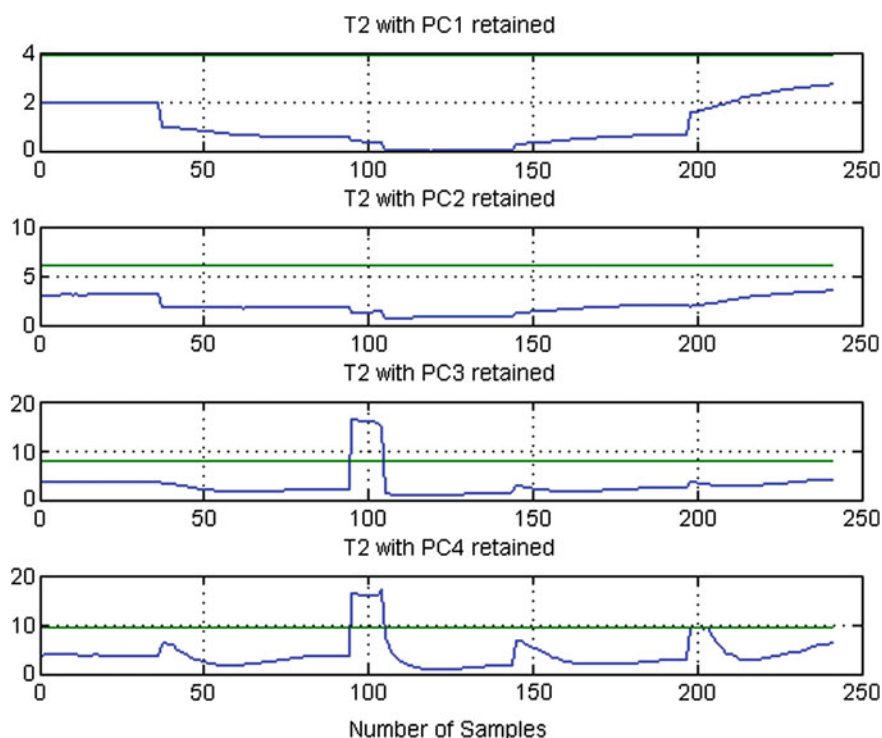


Fig. 75.10 T^2 statistics plot for F3 of the thermal noisy data for various PCs retained

test method and the results from the plots show that the selection of the number of PCs based on fault SNR provides superior performance of fault detection for different kinds of thermal process faults.

References

1. Mnassri B, Eladel EM, Ouladsine M, Ananou B (2010) Selection of the number of principal components based on the fault reconstruction approach applied to a new combined index, 49th IEEE conference, 2010
2. Valle S, Li W, Joe Qin S (1999) Selection of the number of principal components. *Ind Eng Chem res* 38:4339–4401
3. Abdi H, Williams J (2010) *Principal component analysis*. Wiley, Hoboken
4. S. Wold.1978 Cross—validatory estimation of components in factor and principal components models. *Technometrics* 20:397 – 405, 1978.
5. Himes DM, Storer RH, Georgakis C (1994) Determination of the number of principal components for disturbance detection and isolation. In: *Proceedings of American control conference*, IEEE Press, Piscataway, New Jersey, pp 1279–1283
6. Jackson JE (1991) *A user's guide to principal components*. Wiley, New York

7. Jackson JE (1959) Quality control methods for several related variables. *Technometrics* 1:359–377
8. Jackson JE, Mudholkar GS (1979) Control Procedures for residuals associated with principal component analysis. *Technometrics* 21:341–349
9. Xiao F, Wang S, Xu X (2009) An isolation enhanced PCA with expert based multivariate decoupling for sensor FDD in air conditioning systems, *Applied Thermal Engineering*, 2009
10. Dunia R, Qin SJ (1998) Joint diagnosis of process and sensor faults using principal component analysis. *Control Eng Pract* 6:245–255

Chapter 76

E-Learning Environment Identification System: Error Injection and Patterns Dynamics

Deniss Kumlander

Abstract The paper describes a security subsystem of the e-learning environment designed to identify users using so called soft methods. The approach combines the soft biometric characteristics of the individual and environments parameters, i.e. characteristic measuring, which we do not need any specific additional hardware or software installed, in order to capture patterns. Those patterns are applied later on the reactive base identifying persons. In order to improve the identification process an error injections technique is proposed. The subsystem is educated during the e-learning process of students and provides an output during the examination process in order to avoid both submitting answers by other persons and using an external help. The security subsystem is not a strict one and is designed to produce alerts to teachers, so the “hard” approach can be requested, i.e. examination using a web cam or the peer-to-peer method. The paper discusses how the security subsystem can be designed and implemented in order to increase the general reliability of the identification process.

76.1 Introduction

Importance of an electronic learning as an alternate solution to the traditional education is practically not debatable any longer since it sufficiently improves the availability of the education and increases the personal freedom—the freedom to be anywhere, move or stay at home, obtain knowledge from classes available

D. Kumlander (✉)

Department of Informatics, Tallinn University of Technology,

Raja St.15, 12617 Tallinn, Estonia

e-mail: kumlander@gmail.com

within the country instead of been restricted to schools close to you. Accordingly to several studies the availability of e-educational system has sufficiently increased with development of electronic channels and world globalization [1, 2]. Thousands students are already attending e-courses in developed countries and millions are looking for that in so called “third world”, which quickly bridging the gap in living and educational standards. The recent financial crisis and evolution of technologies have shown to us that re-education is also crucial in order to avoid the structural unemployment. The traditional education cannot provide the required flexibility in many cases and e-educational approach becomes the only feasible the facto. Finally to be mentioned that a lot of studies among young people also have shown their high devoutness to media and electronic channels including socializing, educating, obtaining information etc. and decreasing interest to standard or traditional educational approaches of mentoring face to face in classes.

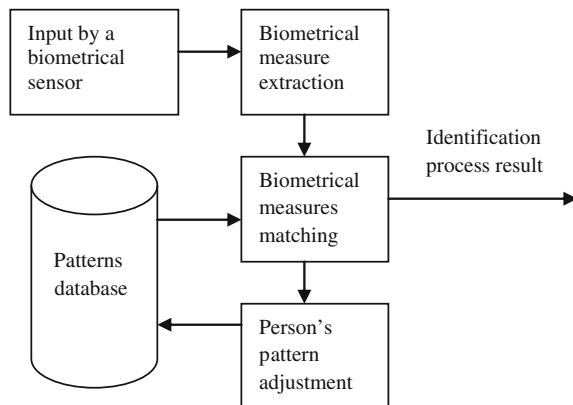
The main obstacle in the e-learning process arises when the educational process is designed to end up with recognition of achievements by the educational organisation or government releasing certificates. Here the examination process and the student identification become crucial [1]. There are several methods that can be applied to solve this problem including certified centers or “hard” biometrical methods as an entrance requirement. Unfortunately those approaches are not always possible or sufficiently restrict people’s ability to attend the examination or wiliness to enter into the course as there could be a common fear or inability to install and make to work additional hardware especially among middle age students. The wide usage of ID cards and similar methods do not help us also since unlike banks we deal with an examination environment where the side help for some students are not afraid, but instead asked and appreciated and teachers vice versa try to avoid that. Summarizing all previously said the current students’ identification methods sometimes prevent people from using e-courses [3] or not enough to identify students, which is required to issue reliable certificates.

Therefore, from our point of view, it is important to develop a security identification subsystem that will be transparent for the end-users of e-courses, but will be good enough to warn teachers in case of suspicious activity during the examination process, which refers to potential manipulation within the identification process.

76.2 Biometrical Markers

Biometric is a relatively novel approach to the computer security that uses either physical or behavioral person characteristics to identify that person. The major idea is to learn computers to identify persons as people do it in the everyday life. The most well-known system of identification is finger prints. In fact, it was used much longer than electronic computers do exist at all. Nowadays, there are much more types of different security approaches for computers relying on biometrical methods. Consider for example face recognition, eyes’ patterns, signature or voice

Fig. 76.1 Adjustable biometrical identification systems architecture [3]



recognition [4, 5]. Unfortunately most biometric approaches require special hardware systems and therefore are not good enough to be used in e-learning as rarely students have such devices or are ready to buy those.

The biometrical security system is relatively easy to implement and use. It is possible to vary the factors to be measured and users will be verified by. The system can start from some basic measures and could evolve by implementing more and more features. The general approach to storing, capturing and using data is very similar to other computer based recognition systems. Typically there are three basic elements as it is demonstrated in Fig. 76.1. The first part is training block responsible for registering users in the system by producing patterns for each user during users' communication with the system. The sensor produces raw data and an analyzer generates patterns, which are:

1. Bounded to a particular measure, like time intervals;
2. Pre-processed and stored in a compact way ready for later use.

The registered patterns are saved into a patterns database, which is the second part of the system. After that, when user logs into the system and uses it, the third element of the system activates. During the matching step, a user pattern is once again captured via the same sensors and compared to the saved one. If those are different then the system issues an alert to the environment administrator on this user. The similarity level of the compared patterns is normally tuned by setting a critical level, lower which patterns are called different. The system should adjust patterns constantly in case individual's behavior is changing as s/he become more familiar with computers etc.

One of the simplest and so the best method to be used in such biometric security system is keystroke patterns, where the system monitors first of all the keystroke typing dynamic. It is well-known among telegraph operators that writing/typing dynamics are assumed to be unique to a large degree among different people [6]. Actually the keystroke typing dynamic can be measured by the following sub/characteristics [2, 6]:

1. Duration of the keystroke, hold time;
2. Latency between consecutive keystrokes;
3. Overall typing speed;
4. Habits by using additional keys on the keyboard like typing numbers via number pad part;
5. Frequency errors and backspace versus delete key using; and some others.

The keystroke recognition systems research is started in 1980s [7] and includes a set of specific recognition methods like [2]:

1. Statistical methods [7–9];
2. Artificial intelligence methods/machine learning and data mining, including but not restricted by neural networks, graphs and Euclidean distance metrics, decision trees etc. [10–12];
3. Generic algorithms [12] and fuzzy classification methods [13].

76.3 Error Injection

The error injection idea is to unbalanced the system in the controlled manner, so we can compare the expected variations to the actual one during the feedback cycle in both learning and examination phases of the system work and so collect additional information that can be used to strengthen the biometrical patterns described above and so improve the overall quality of the identification process.

The error injection can have several forms as described below and be used depending on parameters specific for each individual e-educational system.

1. Error injection into questions—define a faulty statement or provide no correct answer except the free style text box where the student can formulate additions and other clarifying statement. The idea of this error injection is to motivate to act, write longer explanations, write complains to admin and so provide more possibilities to measure biometric and form the individual patterns;
2. Error injection into list of questions—re-ask questions that do not belong to the current topic, but belongs to the previous one and was correctly answered before. Here the system evaluates things and facts the student knew before forming the knowledge dynamic as will be described later and tries to question already formulated patterns. Notice that building up the questioner we can divide questions among several pages leaving us a possibility, in case of alerts, to add extra checks on fly to prove the alert and so suspicious activity during the current check or exam;
3. Error injection into the context. Here we can make it impossible to use certain controls like checkboxes to train the system simulating other kind on environment, for example been on desktop simulate a tablet.

76.4 Patterns Dynamics

The original approach to pattern based recognition described above is rather static and pattern adjustments are designed to detect the pattern more precisely the more student collaborate with the system (biometrical sensors). This approach does not consider both different types of units the e-learning system is consumed from [14] and the knowledge evolution of students. Both of them sufficiently changes the biometric pattern and so requires sufficient modifications before we can apply the system in the real world environment targeting set of students we have no control over.

The simplest solution for different kind of units would be creating several patterns per users for each individual unit (computer, tablet, phone etc.) she can be using, choose the current session pattern basing on a unit used at the moment and consume only that to identify the person. In fact those patterns are not fully independent and evolution of user knowledge within the learning subject as well as dynamic of biometric measures changes for the individual should be included into all of them. Considering previously said we can easily conclude that the different units patterns have a lot of common and varying part is changing just because one or another part of the system is either more or less important. For example the screen manipulation with touch is mostly available on tablets and nearly not possible on desktops. This makes it possible to model the pattern as a set of layers where the main biometric characteristic for the individual stays the same, but their weight is changed depending on the kind of units he is using at the present moment. In other words the system can be represented as a classical neural network, where w_i is the weight of each individual characteristic for the current case and AI is the engine of the security pattern based system extracting patterns and weights for the current session person/type of consumed device. Notice that the adjustment feedback loop remains the same and is not shown on the picture (Fig. 76.2).

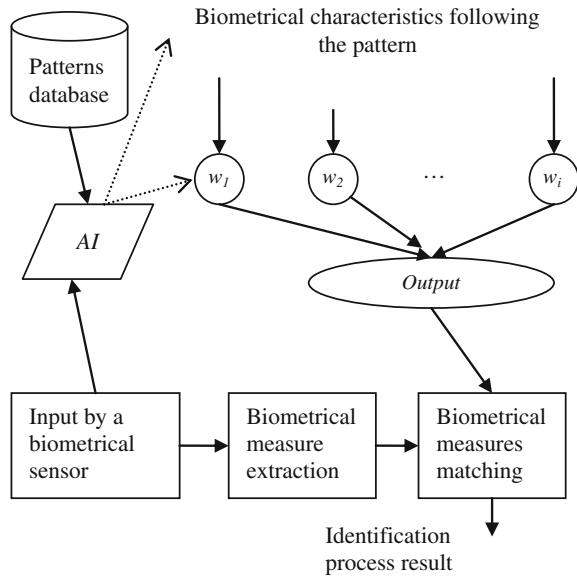
Into addition to the direct biometrical characteristics there are some important measures usage of it can sufficiently improve the recognition process. Those are:

1. Lexical pattern
2. Knowledge pattern.

The lexical pattern recording process is similar to process widely used to identify an author of unknown or disputable books. It is well known that each person uses rather unique set of words and idioms that has been formed by his education, background, processed literature and so forth. Using a reasonably restricted set of sentences (several pages during the system learning phase and 10–15 sentences during the examination) is sufficient to derive the lexical patterns and use them for the identification.

The knowledge pattern is used to record what the student has learned during the course and its' dynamic. This pattern uses the fact that the amount of known for the person rarely decreases during the course—for majority of us it is constantly

Fig. 76.2 Pattern base on weights depending on type of device used to access the e-learning system



increasing since that is what we measure during the examination—obtained knowledge during the course. This can be relaxed to a very simple rule we are going to employ: if a student knew an answer on some kind question during the first check then it is impossible that he will forget it by the next check or exam (otherwise the initial knowledge he has shown was artificial or faulty and so this should produce a knowledge gap alert). This kind of pattern is the most evolving one and is one of the best goals for the error injection approach described above.

The knowledge dynamic can be used to weaker the acceptance level for other perceptions as too large gap between past and current could refer to external help and so other sensors information should be evaluated strictly i.e. with smaller variance allowed.

76.5 Concluding Remarks

It will be incorrect to rely only on the proposed approach building a security system for identification of students during an examination phase of e-learning. The soft biometric measures and earlier mentioned approaches to increase their reliability do not guarantee the full identification of students by themselves, but do generate an alert so the additional methods can be executed to verify the identification knowledge match like face-to-face examination—5 min long over web cam repeating the question the student was able to answer during the suspicious examination.

The recorded patters are considered in many countries to be personal data that should be in now circumstances made public, so the required level of security and access control should be built around it.

76.6 Conclusion

A student identification process is a key element planning, implementing and executing an e-learning environment. The malfunctioning identification subsystem can lead to creditability decrease in case of inability to identify the person claiming the certificate during after the successfully passed examination. At the same time setting too high restrictions on the online examination requirements like demanding web cameras or even requiring attending the examination in the offline mode in dedicated centers could be a sufficient demotivating factor to attend the course. Therefore in this study we explored possibilities to identify students without such strict restrictions.

The paper proposes an identification sub-system that is designed to be executed during the educational process in the learning mode and as an alerts generator during the examination. The system is based on individual patterns like biometrical and lexical, knowledge dynamic and error injection.

The proposed system is able to handle correctly also different types of devices used to access the e-learning environment keeping the individual patterns independent on them using weights based neural network like system allowing adopting the recorded pattern to the device type increasing or decreasing weights of the certain elements of the pattern.

References

1. González-Agulla E, Argones-Rúa E, García-Mateo C, Flórez ÓWM (2004) Development and implementation of a biometric verification system for e-learning platforms, EDUTECH, computer-aided design meets computer-aided learning. In: IFIP 18th world computer congress, pp 155–164
2. Guven O, Akyokus S, Uysal M, Guven A (2007) Enhanced password authentication through keystroke typing characteristics. In: Proceedings of the 25th IASTED international multi-conference: artificial intelligence and applications, pp 317–322
3. Kumlander D (2008) Soft biometrical students identification method for e-Learning. In: Sobh T (ed) Advances in computer and information sciences and engineering. Springer, Netherlands, pp 114–118
4. Jain AK, Ross A, Prabhakar S (2004) An introduction to biometric recognition. *IEEE Trans Circuits Syst Video Technol* 14(1):4–19
5. Jain AK, Hong L, Pankanti S (2000) Biometric identification. *Commun ACM* 43(2):91–98
6. Illonen J (2003) Keystroke dynamics. In: Advanced topics in information processing lectures
7. Gaines R, Lisowski W, Press S, Shapiro N (1980) Authentication by keystroke timing: some preliminary results, Rand Report R-256-NSF, Rand Corporation
8. Joyce R, Gupta GK (1990) Identity authentication based on keystroke latencies. *Commun ACM* 33(2):168–176
9. Sogukpinar I, Yalçın L (2004) User identification via keystroke dynamics, *Ist. Univ. J Electr Electron Eng* 4(1):995–1005
10. Bleha S, Slivinsky C, Hussien B (1990) Computer-access security systems using keystroke dynamics. *IEEE Trans Pattern Anal Mach Intell* 12:1217–1222

11. Sheng Y, Phoha VV, Rovnyak SM (2005) A parallel decision tree-based method for user authentication based on keystroke patterns. *IEEE Trans Syst Man Cybern* 35(4):826–833
12. Yu E , Cho S (2003) GA-SVM wrapper approach for feature subset selection in keystroke dynamics identity verification In: *Proceedings of the international joint conference on neural networks*, vol 3, pp. 20-2253–2257
13. Hussien B, McLaren R, Bleha S (1989) An application of fuzzy algorithms in a computer access security system. *Pattern Recognit Lett* 9(1):39–43
14. Kumlander D (2010) Context driven personalized e-learning environment provided as a state service. Technological developments in networking, education and automation. In: *CISSE 2009*, pp 43–48

Chapter 77

Energy Consumption by Deploying a Reactive Multi-Agent System Inside Wireless Sensor Networks

Alcides Montoya and Demetrio Ovalle

Abstract Intelligent software agents can be a valuable tool to model and implement wireless sensor networks (WSN). Such networks have a set of inherent limitations, such as energy, limited resources, limited computing, and unreliable wireless links. These limitations make the design and development of intelligent software agents and multi-agent systems in such networks hard and complex. In near future, WSNs will be more robust and highly supported by intelligent agents that will allow WSNs to behave like intelligent systems. This paper presents the results of an experimental WSN system executing reactive agents in nodes. We measure the energy consumption and propose a possible integration model of multi-agent architectures, with WSNs using plug computers as a strong base station.

77.1 Introduction

The need for long lifetimes and small form factors, common to most WSN, does not match up well with the power density of the available battery technology. This could limit the use of WSNs because of the need for large batteries. Better batteries for small devices are not expected to become available in the near future. Therefore, energy harvesting could be a solution to make WSNs autonomous,

A. Montoya (✉)

Scientific and Industrial Instrumentation Group, National University of Colombia—Medellin Campus, Autopista Norte, Carrera 65 A.A, 3840 Medellín, Colombia
e-mail: amontoya@unal.edu.co

D. Ovalle

Artificial Intelligence Research and Development Group, National University of Colombia—Medellin Campus, Autopista Norte, Carrera 65 A.A, 3840 Medellín, Colombia

which could enable the widespread use of these systems in many applications. Such WSNs would be able to perform their sensing functions and wireless communication without any supervision, configuration, or maintenance. Tapping into freely available power from sources such as motion, heat, the sun, and radio waves could enable a whole new class of portable devices with no batteries to charge.

Multi-agents inside embedded systems are a new area of research. The WSNs are based on the client/server computing model, where each sensor node sends its sensory data to a back-end processing center or a sink node. However, despite many research works that emphasize the advantages of agent technology in the context of distributed systems, few reports have been made regarding experiences on the use of agents in real wireless environments and measures of the energy consumption of these agents. Section 77.2 of this paper presents an energy consumption model proposed by [2], and this model is used as the base for our experimental study in Sect. 77.3. We show the results in Sect. 77.4, and propose possible model in Sect. 77.5.

77.1.1 Processor Energy Model

Processor operation state: The processor module is the node control and data processing center, responsible for sensor control, protocol communication, and data processing. The microprocessor normally supports three operation states (sleep, idle, and run) and has five state transitions [1]. Using the equations and proposed model as in [2], we have the Processor energy function:

$$E_{\text{cpu}} = E_{\text{cpu-state}} + E_{\text{cpu-change}}$$

$$E_{\text{cpu}} = \sum_{i=1}^m P_{\text{cpu-state}}(i) T_{\text{cpu-state}}(i) + \sum_{j=1}^n N_{\text{cpu-change}}(j) e_{\text{cpu-change}}(j) \quad (77.1)$$

where $P_{\text{cpu-state}}(i)$ is the power of state i that can be found from the reference manual, and $T_{\text{cpu-state}}(i)$ is the time interval in state i that is a statistical variable calculated in the model. $N_{\text{cpu-change}}(j)$ is the frequency of state transition j , and $e_{\text{cpu-change}}(j)$ is the energy consumption of one-time state transition j , which can be expressed as

$$e_{\text{cpu-change}}(j) = T_{\text{init-end}}(j) \left(\frac{P_{\text{init}}(j) + P_{\text{end}}(j)}{2} \right) \quad (77.2)$$

where $P_{\text{init}}(j)$ and $P_{\text{end}}(j)$ are the power of state init and end in the state transition j , respectively, and $T_{\text{init-end}}(j)$ is the time interval for the state transition j from the init state to the end state. The power state is considered as the average power of state init and end.

77.1.2 Transceiver Energy Model

The communication module includes baseband and radio frequency. The transceiver normally has six states (T_x , R_x , Off, Idle, Sleep, CCA/ED) and nine state transitions [3].

The transceiver energy function can be expressed as in [4]:

$$\begin{aligned}
 E_{\text{trans-state}} &= E_{TX} + E_{RX} + E_{\text{Idle}} + E_{\text{sleep}} + E_{CCA} \\
 E_{\text{trans-state}} &= \sum_{i=1}^{N_{TX}} P_{TX} \frac{L_i}{R} + \sum_{i=1}^{N_{RX}} P_{RX} \frac{L_i}{R} + P_{\text{Idle}} T_{\text{Idle}} + P_{\text{sleep}} T_{\text{sleep}} + P_{CCA} T_{CCA} \\
 E_{\text{trans-state}} &= \sum_{i=1}^{N_{TX}} V_{tr} I_{TX} \frac{L_i}{R} + \sum_{i=1}^{N_{RX}} V_{tr} I_{RX} \frac{L_i}{R} + V_{tr} (I_{\text{Idle}} T_{\text{Idle}} + I_{\text{sleep}} T_{\text{sleep}} + I_{CCA} T_{CCA})
 \end{aligned} \tag{77.3}$$

where E_X , P_X , I_X , and T_X are the energy consumption, power, electric current, and time interval of transceiver in state x , V_{tr} is the working voltage, L_i is the size length of the i^{th} packet received or sent, R is the data transferring rate, and N_{TX} and N_{RX} are the local numbers of sending and receiving packets. $E_{\text{trans-change}}$ can be expressed as in the next equations, where $j = 1, 2, \dots, n$ is the type of state transition and n is the number of the state- transition ($n = 9$), $N_{\text{trans-change}}(j)$ is the frequency of state transition j , and $e_{\text{trans-change}}(j)$ is the energy consumption of one-time state transition j , which can be expressed as

$$E_{\text{trans-transition}} = \sum_{j=1}^n N_{\text{trans-change}} e_{\text{trans-change}}(j) \tag{77.4}$$

$$\begin{aligned}
 e_{\text{trans-change}}(j) &= T_{\text{init-end}}(j) \left(\frac{P_{\text{init}}(j) + P_{\text{end}}(j)}{2} \right) \\
 e_{\text{trans-change}}(j) &= V_{tr} T_{\text{init-end}}(j) \left(\frac{I_{\text{init}}(j) + I_{\text{end}}(j)}{2} \right)
 \end{aligned} \tag{77.5}$$

77.1.3 Sensor Energy Model

The sensing module consists of sensors and digital-analog converters. The energy consumption of the sensing module comes from multiple operations, including signal sampling, signal conversion, and signal modulation. The sensing module can operate either in burst or periodic mode. In general, it operates in periodic mode. Assuming that the energy consumption levels of the open and close operations are constant, the sensor energy consumption can be expressed as

$$\begin{aligned}
 E_{\text{sensor}} &= E_{\text{on-off}} + E_{\text{sensor-run}} \\
 E_{\text{sensor}} &= N(e_{\text{on-off}} + e_{\text{off-on}} + V_s I_s T_{ss})
 \end{aligned} \tag{77.6}$$

where $e_{\text{on-off}}$ is the one-time energy consumption of the closing sensor operation, $e_{\text{off-on}}$ is the one-time energy consumption of the opening sensor operation, $E_{\text{sensor-run}}$ is the energy consumption of the sensing operation, V_s and I_s are the working voltage and current of the sensor, T_s is the time interval of sensing operation, and N is the number of sensor opening and closing operations.

77.1.4 Complete Node Energy Model

In real systems, the processor, transceiver, and sensor components of WSN nodes must work cooperatively to perform a task, and thus have mutual relationships, especially where the energy issue is concerned. The event-driven mechanism of different node modules is as follows:

- Event trigger in the sensor module: The sensor energy model (SEM) enters the 'on' state periodically triggered by the external clock event. After sensing and statistically calculating sensor energy consumption, the sensor module enters the 'off' state automatically.
- Event trigger in the processor module: The processor energy model (PEM) enters the 'run' state triggered by the following three events: the periodic data collection event generated by the SEM, the sending packet requests generated by external applications or protocols, and the packet arriving action generated by the transceiver energy model (TEM).
- Event trigger in the transceiver module: The TEM enters the T_X state triggered by the sending packet event generated by PEM, enters the R_X state triggered by the external packet arriving action, and enters the CCA/ED state triggered by channel detection commands.

Figure 77.1 lists the state current and state transition time for the Chipcon CC2420 transceiver. H.Y. Zhou and colleagues consider a typical node WSN that consists of an Intel Strong ARM SA-1100 Microprocessor [5], a Chipcon CC2420 Transceiver [6], and a Dallas digital temperature DS18B20 [7], and in their paper shows the results of the simulation using this model for energy analysis without multi-agents.

77.2 Experimental Study

77.2.1 Node Architectures and Energy Models with Agents Running Inside the Node

In a WSN, the battery is generally not replaceable due to the randomness of the sensing device's position and sometimes due to the danger posed by the sensing

Fig. 77.1 Current and state transition for the Chipcon CC2420

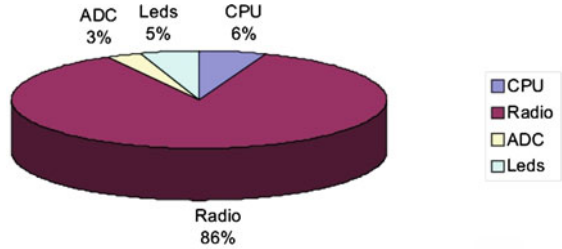
| State | Current |
|--------------------------|-----------------------|
| I _{off} | 0.02 μ A |
| I _{tx} | 17.4 mA |
| I _{rx} | 19.7 mA |
| I _{idle} | 426 μ A |
| I _{cca/ed} | 17.4 mA |
| I _{sleep} | 20 μ A |
| State transition Class | State transition time |
| T _{off-idle} | 1 ms |
| T _{cca/ed-idle} | 2 μ s |
| T _{idle-cca/ed} | 192 μ s |
| T _{sleep-idle} | 0.6 ms |
| T _{idle-sleep} | 192 μ s |
| T _{tx-idle} | 2 μ s |
| T _{idle-tx} | 192 μ s |
| T _{rx-idle} | 2 μ s |
| T _{idle-rx} | 192 μ s |

field. Therefore, battery lifetime is synonymous with sensor lifetime, and must be extended as much as possible. Furthermore, the progress made in battery capacity, lifetime, is at best limited when compared with the progress made in processing power, and storage capacities. Finally, with the ever-increasing performance constraints, power management always needs improvement.

In the experiments, we start by using the simplest type of multi-agent system, namely, reactive agents. These agents do not have any internal symbolic models of their environment; they act with a stimulus/response type of behavior that responds to the present state of the environment in which they are embedded.

A reactive decision agent is a 9-tuple one, denoted by $\langle I_d, A, D, S, E, E', O, O', \text{act, dec, sig} \rangle$ [8, 9]. I_d is the agent identity. A is the set of actions executed on the agent, each action representing a possible operation to be carried out on this object to achieve a specific goal. D is the set of decisions generated by the agent, each decision being a solution concerning process behavior in the future, and is characterized by its action horizon. S is the set of signals received by the agent; each signal received by an object reflects at any given time the state of the controlled tools used to achieve a specific goal. E is the set of external states delivered by the agent, each state representing the object state emitted to the environment. E' is the set of the agent's internal states, each state indicating the current state of the agent. O is the set of agent's internal objectives; each decision is elaborated to achieve an internal objective according to the current external objective and the actual internal state. O' is the set of external objectives that can be achieved, these objectives representing the agent's interpreter of each action. Finally, act, dec, and sig are the three decision functions that define the behavior of a reactive decision agent.

Fig. 77.2 Components of the energy consumed in a WSN node



77.2.2 Experimental Description of the WSN and the Reactive Decision Agent

For this experiment, we use Mica2 crossbow nodes. The consumption of one node has different components. Figure 77.2 shows the component of the energy consumed and gives an idea of the parameters that must be optimized to obtain better performance in time.

The experiment is a WSN composed of four nodes and the management node (sink). Each node uses: two SPI devices (Radio/Flash), two I2C sensors (Humidity/Temp). The two I2C sensors are on the same chip (same I2C address), but require separate sensing command sequences. Every 2 min, the application samples four sensors and logs the reading in flash. Every 10 min the application retrieves new readings from the flash and sends them to the gateway. This is the key in this WSN: store the data in a flash memory and transmit only important data over a long time. Table 77.1 provides the basic measures for calculating the energy costs of different I/O operations and sleep states. Sampling and sending are completely decoupled: the two have a producer/consumer relationship. The reactive agent takes the data (Humidity and Temp) every 2 min, loads the data in the flash memory, and sends the data to the gateway every 10 min. There are no mobile or belief-desire-intention (BDI) agents, only this simple agent that stores and saves the data in the flash memory.

Figure 77.3 shows the experimental data.

77.3 Results

The four nodes execute a simple reactive agent. This agent takes the data every 2 min and sends the data every 10 min. A typical operation in 120 min uses 389 mA of current; using AA batteries, this node will work for approximately 11.3 h.

$$x = \frac{(2200 \text{ mAh})(120 \text{ min})}{389 \text{ mAh}} = 678.7 \text{ min}$$

Table 77.1 Energy decrement for reactive agents running in four nodes MICA2

| Time (min) | N1 (I, mA) | N2 (I, mA) | N3 (I, mA) | N4 (I, mA) |
|------------|------------|------------|------------|------------|
| 0 | 2,200 | 2,200 | 2,200 | 2,200 |
| 10 | 2,172 | 2,183 | 2,176 | 2,186 |
| 20 | 2,142 | 2,165 | 2,154 | 2,166 |
| 30 | 2,110 | 2,145 | 2,134 | 2,147 |
| 40 | 2,075 | 2,127 | 2,116 | 2,124 |
| 50 | 2,047 | 2,112 | 2,095 | 2,106 |
| 60 | 2,013 | 2,092 | 2,071 | 2,077 |
| 70 | 1,979 | 2,076 | 2,050 | 2,059 |
| 80 | 1,943 | 2,062 | 2,033 | 2,045 |
| 90 | 1,915 | 2,040 | 2,009 | 2,023 |
| 100 | 1,875 | 2,022 | 1,992 | 1,997 |
| 110 | 1,846 | 1,999 | 1,971 | 1,979 |
| 120 | 1,811 | 1,982 | 1,947 | 1,955 |

Fig. 77.3 Four Mica2 nodes running a simple reactive agent

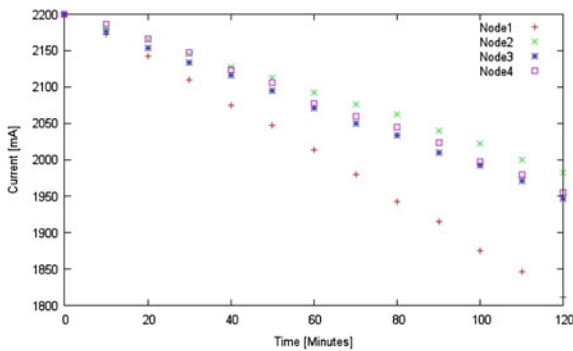
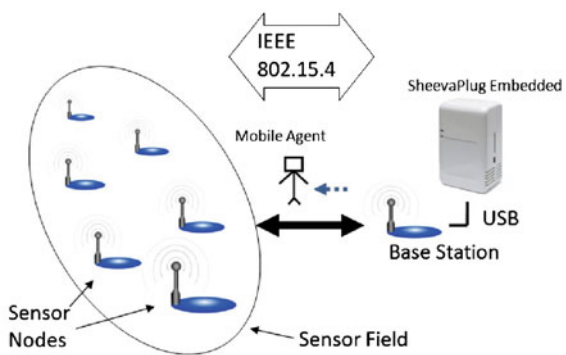


Fig. 77.4 The proposal architecture using a plug computer using reactive agents in the nodes and deliberative agents in the base station



Suppose that the reactive agent only reports the data when there is a critical condition. For example, if a WSN is used to monitor and control the leak of ammonia, the agent can take the data and will send the data only when the ammonia level reaches a critical value, e.g., 30 ppm. Suppose this happens once a day. The equation may then be modified:

$$x = \frac{(2,200 \text{ mAh})(1,440 \text{ min})}{28 \text{ mAh}} = 113,142 \text{ min}$$

which translates to 1885 h or 78.5 days.

Scalability of the system: Using four nodes without routing and having only one reactive agent for each node, we have energy for 78.5 days for each node. For ten nodes, the system may be similar; we do not need low-energy adaptive clustering hierarchy routing or head clusters for the system. It is possible for the network to work using only one base station; in this case, the system would be similar to that presented in the above-mentioned discussion.

77.4 Proposed Approach

An intelligent sensor modifies its internal behavior to optimize its ability to collect data from the physical world and communicate these data in a responsive manner to a base station or a host system. The functionality of an intelligent sensor includes self-calibration, self-validation, and compensation. Self-calibration means that the sensor can monitor the measuring condition to decide whether a new calibration is needed or not. Self-validation applies mathematical modeling error propagation and error isolation or knowledge-based techniques. Self-compensation entails making use of the compensation methods to achieve high accuracy.

The proposal is a combination of mobile agents running in a strong base station and reactive agents in nodes, and the system is highly dynamic compared with the implicit nature of multi-agent systems and the artificial intelligence inside the WSN. The architecture and solution need a combination between an embedded system that can be run by our agents, the middleware, and the classical nodes used in the WSN. Normally, the base station in a WSN does not have a lot of resources. Recent advances in embedded systems allow having a new base station; in this case, it is a plug computer system, which has a complete Linux system. Specifically, our system has Linux debian kernel 2.6.33, inside which we have Java and openjdk or J9 and Equinox installed. Equinox is used to have an implementation, which can run bundles as components and services. In our development, we have the middleware SIXTH, a complete application that is useful for fast applications for WSNs. Our middleware use Agent Factory Micro-Edition [10–12] as part of its implementation (Fig. 77.4).

The system has a strong base station and similar nodes for the classical nodes in WSN, but in the future it may use 32-bit microprocessors running real-time

operating systems and using energy harvesting. In this proposal, the BDI agents work in the plug computer, and only small mobile agents can be used between the reduced functional devices (nodes) and the full functional device or base station. The intelligent multi-agent runs inside the plug computer and makes decisions about WSN management and performance; the reactive agents in the nodes send only the critical data, and the lifetime of the network will be increased, working as an autonomous system. The RFD runs only reactive agents, similar to the analysis for the reactive agents inside the four Mica2 nodes.

77.5 Discussion

Using a clear energy model, similar to that proposed by H.Y. Zhou et al., we propose a simple reactive agent and combine it with an experimental multi-agent system developed to test the possibility of using artificial intelligence at a primitive level in the WSN. As energy consumption and constraints are a barrier to the development of a new proposal for autonomous systems, this simple system integrates the energy model and the measures with a simple multi-agent. The experimental data point to very good results: it is possible to use simple agents inside the nodes without compromising the battery life or the energy consumption, and it leads to fewer transmissions of the data samples and to an increase in the periodic data samples sent at the base station.

For more complex systems, we propose a new architecture based on BDI systems executing hard tasks inside the new proposed base station of the plug computer system. Using the strong base station and AFME system, we have an autonomous system that can executed hard tasks, such as monitoring ammonia or methane in factories.

77.6 Conclusions

The results from this study show that the energy consumed by a simple reactive multi-agent is low compared with other processes, and carrying out tasks, such as taking sensor samples, storing the values in the flash memory, and sending the data every 10 min, increases battery life and guarantees a longer lifetime for energy in the nodes.

The energy consumed by deploying a reactive multi-agent is under normal conditions and can increase the operation modes with the addition of some intelligence in the WSN nodes.

The tasks of the reactive multi-agent inside the node, such as taking the data, analyzing the data, storing the sensor data in the flash, and sending the values each time in one packet, guarantee the first autonomous type of WSN.

Other architectures can be proposed for the use of multi-agent systems inside WSNs. For example, using plug computers as strong base stations for indoor or factory WSNs can increase the autonomous level of the WSN. Strong BDI agents can execute hard tasks inside the plug computer as in a human brain, and simple tasks can be executed by the nodes using simple reactive agents.

Architectures and new proposed models for the use of artificial intelligence inside WSNs can be explored through techniques that not only use mobile agents between nodes in a classical way, but also make better use of simple reactive agents working in simple tasks and mobile agents between a plug computer base station and the nodes.

Acknowledgments The authors thank The National University of Colombia-Campus Medellin and the Colombian Institute for the Development of Science and Technology, COLCIENCIAS program to train leaders in innovation, and the research project titled “Development of a model of intelligent hybrid system for monitoring and remote control of physical variables using distributed wireless sensor networks” Contract RC. No. 472, Code 20201007027.

References

1. Lu F, Xu GZ, Ying RD (2005) Power consumption simulation model based on the working status of intel PXA250 processor. *Control Autom* 21(1):131–132
2. Zhou HY, Luo DY, Gao Y, Zuo D (2011) Modeling of node energy consumption for wireless sensor networks. *Wirel Sens Netw* 3(1)
3. Wang XF, Xiang J, Hu BJ (2009) Evaluation and improvement of an energy model for wireless sensor networks. *Chin J Sens Actuators* 22(9):1319–1321
4. Han B, Zhang DZ, Yang T (2008) Energy consumption analysis and energy management strategy for sensor node. International conference on information and automation. In: *Proceedings of the 2008 IEEE*, vol 6, pp 211–214
5. Intel Corp (1999) Intel strongARM SA-1100 microprocessor, Developer’s Manual
6. Tsiatsis V, Zimbeck S, Srivastava M (2001) Architectural strategies for energy efficient packet forwarding in wireless sensor networks. In: *Proceedings of the 2001 international symposium on low power electronics and design*, vol 3, pp 25–31
7. Jiang TH, Jiang ZW (2003) Characteristics and application of the digital temperature sensor DS18B20. *Electron Technol* 12:46–49
8. Romadi R, Bounabat B, Lafont et S, Labhalla JJC (1998) Users’s behavioral requirements specification for reactive agent. In: *IEEE publication-CESA’98*, Nabeul-Hammamet, Tunis, April 1998
9. Romadi R, Bounabat B (1999) Designing multi-agent reactive systems: a specification method based on decisional reactive agent. In: *PRIMA’99*, Japan, 2–3 December 1999
10. Muldoon C (2008) An agent framework for ubiquitous services. Ph D thesis, School of Computer Science and Informatics, University College Dublin, Ireland
11. Muldoon C, O’Hare GMP, Collier RW, O Grady MJ (2006) Agent factory micro edition: a framework for ambient applications. In: *Intelligent agents in computing systems. Lecture Notes in Computer Science*, vol 3993. Springer, Reading, pp 727–734, 28–31 May 2006
12. Muldoon C, O’Hare GMP, O’Grady M (2007) Managing resources in constrained environments with autonomous agents. In: *ESAW’06: Proceedings of the 7th international conference on engineering societies in the agents world VII*. Springer-Verlag, Berlin

Chapter 78

Network Intrusion Detection System Based on SOA (NIDS-SOA): Enhancing Interoperability Between IDS

Wagner Elvio de Loiola Costa, Denivaldo Lopes, Zair Abdelouahab
and Bruno Froz

Abstract Anti-virus and firewall protection systems are designed to prevent the execution of evil deeds in the network, thus constituting a barrier to invaders (e.g. viruses, worms and hackers). However, there is no guarantee to full protection of the computer network, because invasions may occur. In this case, Intrusion Detection System (IDS) provides intrusion detection and subsequent notification to the network administrator, or in conjunction with the firewall it blocks the port used in the invasion or the IP address of the attacker. An important factor for intrusion detection is the quality of database signatures. However, IDS systems are isolated; they do not share the signatures, and do not cooperate and the database signatures are not easily reused. Generally, they communicate using different protocols and are designed with different programming paradigms. In this paper, we present Network Intrusion Detection System based on SOA (NIDS-SOA) in order to allow interoperability between two or more IDSs for exchanging subscription information and notifications of occurrences of invasions and provide support for isolation of an invasion.

W. E. de Loiola Costa (✉) · D. Lopes · Z. Abdelouahab · B. Froz
Federal University of Maranhão – UFMA, Campus do Bacanga, 65080-040
São Luís - MA, Brazil
e-mail: welc@dee.ufma.br

D. Lopes
e-mail: dlopes@dee.ufma.br

Z. Abdelouahab
e-mail: zair@dee.ufma.br

B. Froz
e-mail: brunofroz@dee.ufma.br

78.1 Introduction

Computer networks offer extensive benefits to users; these benefits range from simply sending an e-mail, acquisition of products with electronic commerce to Internet banking. However, these benefits may only occur in a secure environment. A secure environment or a secure computer network can be defined (a) as an environment or system that has computer programs and operating system updated and configured correctly and a protection system (antivirus, firewall and IDS). Antivirus and firewall protection systems are designed to prevent evil deeds in the network, thus constituting a barrier to invaders (e.g. viruses, worms and hackers). An IDS is a system designed to detect the intrusion which passes through barriers such as antivirus and firewall. An IDS is a system that is behind the protective barriers and aims to detect an intrusion as soon as possible to minimize damages. Assuming that there is a completely safe environment or a network computer completely secure; minimizing the detection time of an invasion and the time to take countermeasure is of fundamental importance to users. Several IDS products are available; they use different mechanisms of detection, countermeasures, and different paradigms in their design, implementation and communication protocols. Consequently, existing IDS systems are isolated and reusing features of other IDSs are difficult to implement. The reuse of functionalities between intrusion detection systems (IDS), i.e., the interoperability between different IDS, is a challenge and can significantly improve the detection time countermeasures [1]. In this paper we present how to make information available to other IDSs through a Service Oriented Architecture (SOA) approach and composition as a form of reuse.

This paper is organized as follows: In [Sect. 78.2](#) we describe IDS, SOA and Web services. In [Sect. 78.3](#), we describe the functionality of NIDS-SOA framework. In [Sect. 78.4](#), we describe the adaptation of NIDIA to NIDS-SOA. [Section 78.5](#) describes the tests performed with the NIDS-SOA framework and in [Sect. 78.6](#) we present our conclusions about this work.

78.2 Overview

IDS systems are isolated systems and interoperability between different vendors' IDS is complex and difficult to implement. Some of IDSs use non-standard formats and protocols for interaction between their modules [2]. IDS systems in the literature are isolated and are not easily reused [3]. Generally, they communicate using different protocols and are designed with different programming paradigms. The creation of a SOA-based IDS will allow information sharing and reuse of features of other IDSs.

78.2.1 Intrusion Detection Systems: IDSs

An IDS has the ability to detect various attacks coming from both external and internal computer network. It helps in protecting corporate environment and its installation points as of fundamental importance [1]. Intrusion Detection System is originally suggested by James Anderson in 1980 in an article entitled “Computer Security Threat Monitoring and Surveillance”. This paper describes the concepts that an audit system can provide important information on the system misuse by users. Later, these concepts are circulated in 1987 by Dorothy Denning in an article entitled “A Model for Intrusion Detection” [3, 4]. A year later Dorothy Denning has developed a model for intrusion detection. Todd Heberlein introduced the concept of network intrusion detection, and has developed the Network Security Monitor (NSM). The use of NSM in large networks has provided with a large amount of information that is generated by the network. The concept of network intrusion detection started by Todd Heberlein attracted a great interest in this area of intrusion. The principle of an IDS [4] is to capture traffic from TCP/UDP flows in the network of computers and shortly thereafter starts analysis. Thus, an IDS analyzes the content of captured traffic in order to find some resemblance to a package that contains characteristics of vulnerabilities or malicious code previously known in its database that we commonly call rules. In the event finding a package with features that resemble the characteristics that are in its database, the intrusion protection system takes action to protect the system, ranging from sending an alert to the administrator and stronger actions such as blocking the sender through the firewall.

78.2.2 Web Services: Description and Features

Web services have emerged as a natural consequence of the Internet usage growth [5]. Software systems are built to withstand interaction between machines on a network or distributed systems, having a standard communication interface [6, 7]. Web services are components or applications that are accessible via standard Internet protocols. They are self-contained applications, with XML-based interfaces that describe a collection of operations accessible in the network, regardless of the technology of the service implementation. A web service can be available in the Internet by publishing it, having its access settings described in Web Services Description Language (WSDL). A web services can be accessed by any client using SOAP protocol.

Figure 78.1 presents the description of web services and its standard protocols [8].

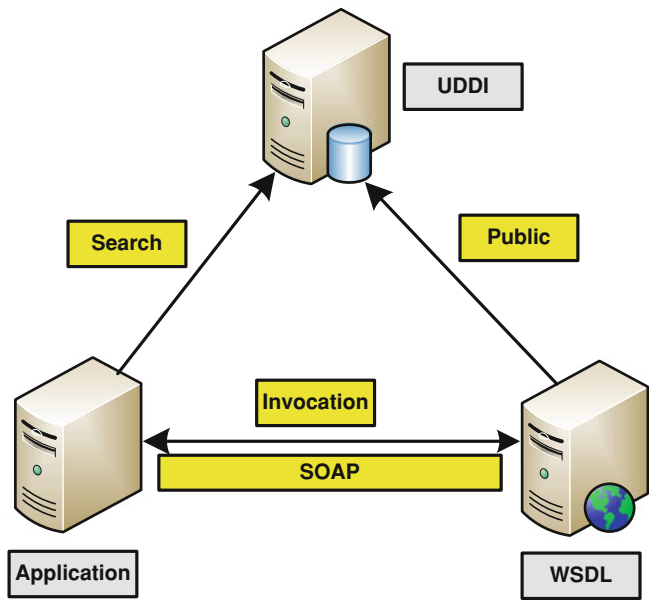
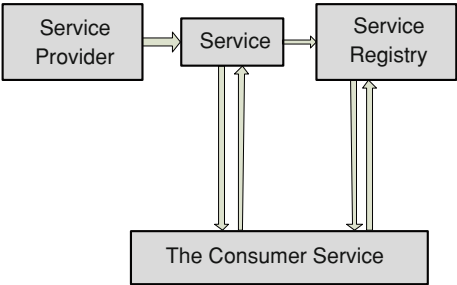


Fig. 78.1 Web services architecture: standards and features

Fig. 78.2 Service oriented architecture



78.2.3 SOA Description and Features

SOA describes the keys concepts of software architecture and their relations, where a service and its use are the key concepts that are involved, following a model of publishing services and applications and their universal access [9, 10]. SOA has an interface that describes a collection of operations accessible over the network via a standardized format (e.g. XML) [8]. These requirements are activated anywhere in a dynamic computing environment and/or pervasive computing where service providers offer a range of services. Figure 78.2 illustrates a service-oriented system with a service consumer, a service provider and a service repository.

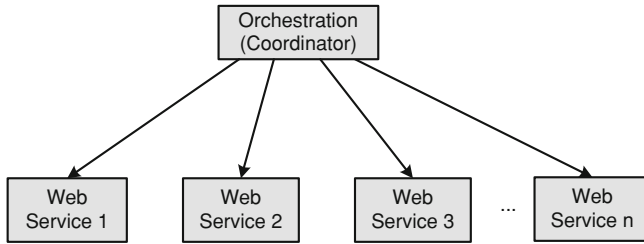


Fig. 78.3 Orchestration in service composition

SOA creates an environment in which distributed applications and components may create independently of language and platform and focuses on the use of a relatively widespread pattern of communication between operations, enabling thus a model for homogeneous distribution and composition of components.

SOA is an approach that aims to integrate existing applications (legacy systems) with new applications. This integration can be done using web services, DCOM, Simple Object Access Protocol (SOAP), Representation State Transfer (REST), Common Object Request Broker Architecture (CORBA) and Remote Procedure Call (RPC). SOA has three technical concepts that are: service, interoperability through a bus service and low coupling [11]. A service is a software component that represents a process, a business task or activity and has an interface and it is invoked via messages. The bus service (bus) is an infrastructure that enables interoperability between distributed systems by making communication between different services or business processes of different languages and technology. The loose coupling refers to the reduction of dependency on one service to another one.

With the evolution of SOA using web services, a language to manage web services for their use in service-oriented architecture was demanded. Several languages were created for this purpose, for example: electronic business XML (ebXML); RosettaNet; Universal Business Language (UBL); and Business Process Execution Language (BPEL) [11].

Orchestration and choreography are approaches employed to compose business processes using web services. In a web service, orchestration is for central control that manages the composition of other web services involved in this operation. In choreography, there is not a web service that performs this service coordination, because each web service “knows” exactly with which other web service to interact, it is embedded in its functionality. These two techniques are shown in Figs. 78.3 and 78.4.

Fig. 78.4 Choreography in service composition

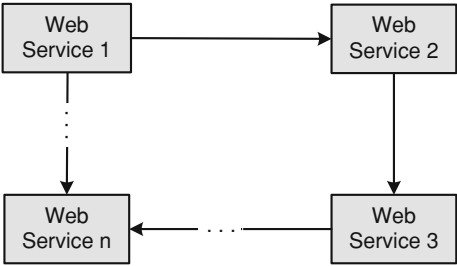
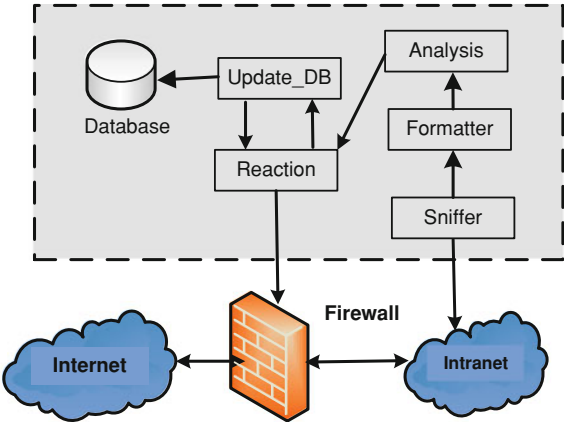


Fig. 78.5 Functions of NIDS-SOA



78.3 Framework of Network Intrusion Detection System Based on SOA

In this paper, we propose the Network Intrusion Detection System Based on SOA (NIDS-SOA) to support interoperability between IDS systems following a SOA approach.

78.3.1 Introducing the NIDS-SOA Framework

Figure 78.5 presents an architecture for NIDS-SOA framework, which aims to detect attacks behaviors in a computer network and performs action to remedy to attacks.

This architecture has the following components: a sniffer captures packets in the network and/or in a log file servers; these packets are passed to the next component *packet formatter* that formats the packets in a XML data structure and with some important information such as source IP, destination IP, TCP and UDP ports access time etc. Packages already formatted are analyzed by the next component based on some rules already defined (in a database). If a rule is find

Fig. 78.6 Services offered by NIDS-SOA

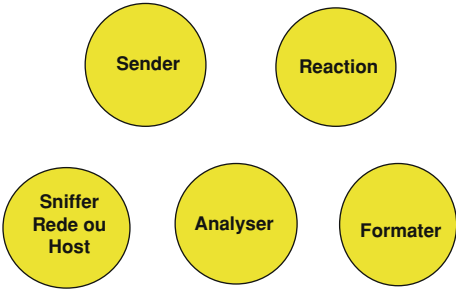
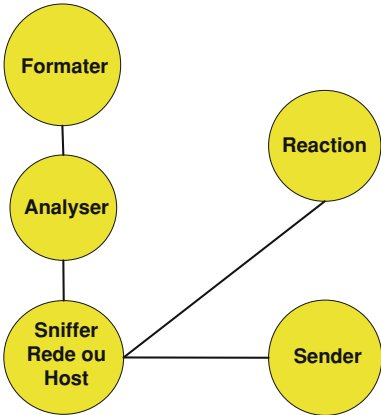


Fig. 78.7 Service composition in NIDS-SOA



compatible, an action (such as blocking the IP or port) is performed in order to deny access to that IP. The DB-Update component is responsible for updating a database (for example inserting a new rule) if needed.

In Fig. 78.6, the NIDS-SOA framework is presented in a service-oriented architecture SOA. The components shown in Fig. 78.5 are transformed into services. The layers (capture, format, analyze, update and reaction) are now services (Host or Network Sniffer, Analyzer, formats, and Reaction). Each service presented in this framework is considered a basic service that can be composed with other services and thus providing a service composition, as shown in Fig. 78.7.

Figure 78.8 shows the class diagram of the NIDS-SOA framework. There are seven classes: WsSniffer, WsFormater, WsAnalyzer, WsSender, WsReaction, RequestMessage, and ResponseMessage.

78.3.2 NIDS-SOA and Web Services

The NIDS-SOA framework uses web services to perform service composition according to SOA approach. Each component or module of Fig. 78.5 has a web service associated with its function. Thus the NIDS-SOA, shown in Fig. 78.9

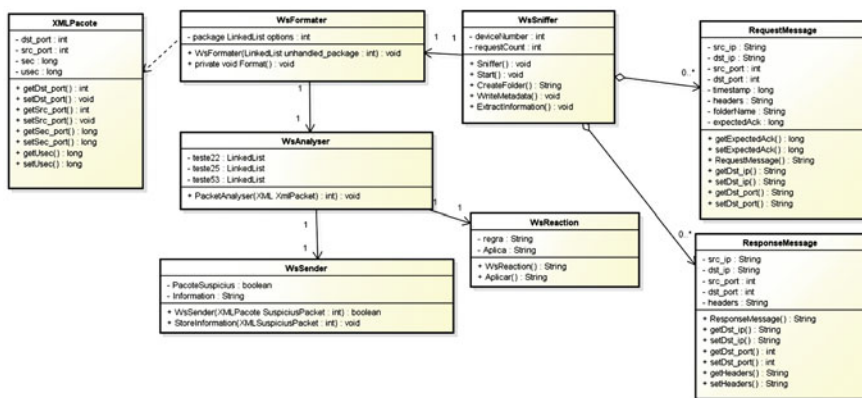


Fig. 78.8 Class diagram of the NIDS-SOA framework

consists of five web services, where each has a web service WSDL acting as a service contract for other layers. In this diagram there is a layer of administration which is performed by the system administrator when it is necessary.

- *WsSniffer*. This web service is listening and is capturing every packet in the network and/or in the host. To start capturing packets, this service must be instantiated, passing the parameters “Number of the device” (integer) and “options” (list of integers) and then call its start() method. With the start() method, the parameter “Number of the device” is passed to JpCap class, which is responsible for receiving and handling the packets.
- *WsFormatter*. This web service formats the packets in XML files, in order to facilitate updating the database as well as a comparison with other packets in the analysis service. This service receives the packets as a parameter which are formatted in XML, with the attributes “DestinationPort”, “Destination IP”, “source port”, “source IP”, “Seconds” and “Milliseconds”. These last two parameters represent the exact time that the packet is captured.
- *WsAnalyser*. This web service looks at each packet. This service receives the packet already formatted in XML and analyzes it according to the rules of danger if they are met. The packet goes through these rules, and if it fails it is passed to the WsSender, but ignored if it passes the rules.
- *WsSender*. This web service updates the database information. This web service receives the packet when it fails in the analysis from WsAnalyzer and saves it in the packets repository to be used in future comparisons with others.
- *WsReaction*. This web services takes actions according to the analysis and based on the available signatures in the database. It may act in the firewall protection (lock access port or lock the source IP of the attack).

Figure 78.10 shows the sequence diagram of the NIDS-SOA framework. The description is as follows:

- The web service (WsSniffer) listens for packets in the network or in a log file.

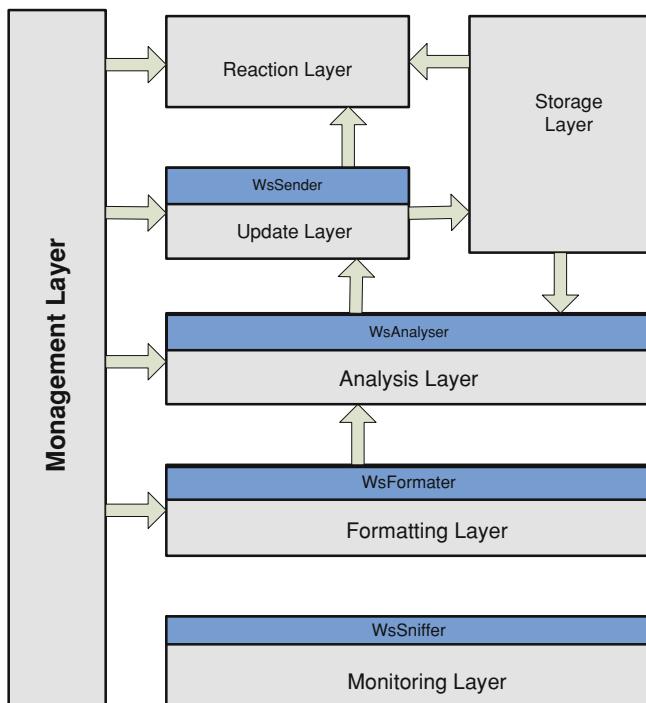


Fig. 78.9 Details of layers of NIDS-SOA and its services

- All packets are formatted by the web service (WsFormatter) in XML file format to facilitate comparison and update later by another web service.
- The web service (WsAnalyser) performs a query check in the database for all formatted packets and if this query is positive, a message is sent to the web service (WsReaction) so that necessary actions are taken to prevent access from that IP (e.g. blockingIP and/or port in the firewall).
- If the query web service (WsAnalyser) does not find any details about the attacker in the database, then a message is sent to the WsSender to insert some information about the attacker in the repository.

78.4 Adapting the IDS-NIDIA to be Conform to NIDS-SOA: Prototyping

The IDS NIDIA is an IDS based on a society of agents. The system is based on the Common Intrusion Detection Framework (CIDF) model that develops protocols and application interfaces with the goal of sharing information [1, 12]

The IDS NIDIA has a set of agents that perform actions according to their functionality. The NIDIA project [13] lists agents with a function of generating

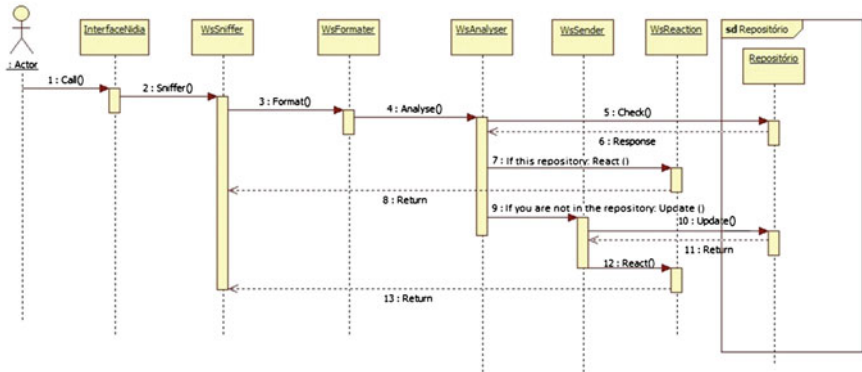


Fig. 78.10 Sequence diagram of the NIDS-SOA

events (sensor agents), agents with a function of monitoring and evaluating safety, agents with a function of storage of historical mechanisms and agents with a function of controlling actions. In Fig. 78.11 we present the architecture of the IDS-NIDIA [14].

78.4.1 Historic of IDS-NIDIA

The IDS-NIDIA has already the following features: a model of automatic update; a model for monitoring and automated responses to attacks [15], a model of automatic update using web services [7]; a model management and integration of databases [16], a reliable communication model specifications using Extensible Markup Language (XML) [17], a model of fault tolerance [13], a model of remote IDS based on web service and MDA [18], a model of IDS for mobile devices [19] and a model of safety and reliability among IDS agents [14]. Our contribution consists of adapting IDS-NIDIA to be conform to NIDS-SOA framework as shown in Fig. 78.9.

78.4.2 Implementing the NIDS-SOA

The NIDS-SOA framework (see Figs. 78.7, 78.8 and 78.9) is applied to IDS-NIDIA and implemented in the NetBeans IDE [20]. The services are implemented in a distributed fashion.

Fig. 78.11 Architecture of IDS-NIDIA

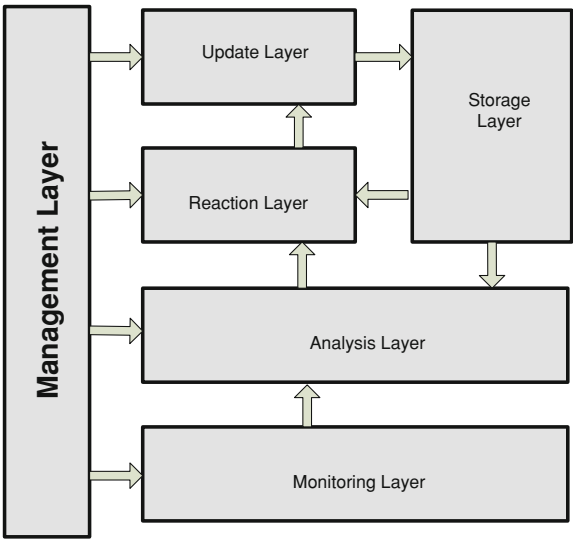


Fig. 78.12 Package in XML format

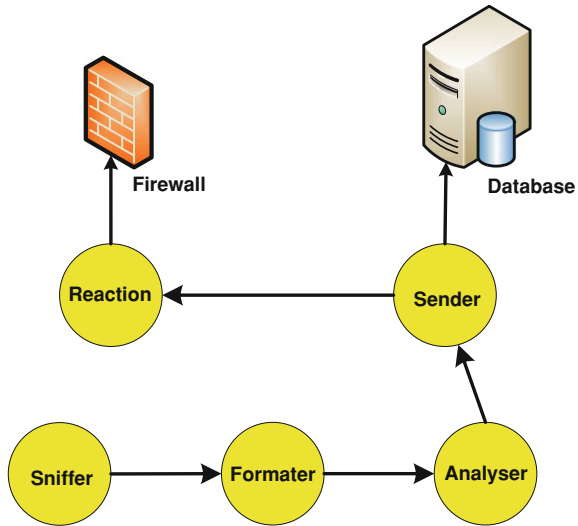
- 1. <pacote>
- 2. <dst_port>49195</dst_port>
- 3. <dst_ip>/192.168.177.103</dst_ip>
- 4. <src_port>80</src_port>
- 5. <src_ip>/129.128.5.191</src_ip>
- 6. <sec>1319653063</sec>
- 7. <usec>523886</usec>
- 8. </pacote>

78.5 Tests Carried out with NIDS-SOA Framework

Tests are conducted with the framework NIDS-SOA using standalone basic services and with service composition. The tests with standalone basic services consist of testing the services separately (e.g. performing only the sniffing). The tests with service composition consist of composing complex services using basic services, example performing sniffing-formatting, sniffing-formatting-analysis, sniffing-formatting-analysis-reaction. Figure 78.12 presents the results of a composition Sniffer and format.

A complete composition of web service can be accomplished by using all the features of all the services of NIDS-SOA framework, where the captured packets from the Sniffer are formatted by the formatter services (Fig. 78.12) then passes to the service Analysis (based on a set of rules) to determine the repeated occurrence of some parameters presented in the XML format packet. If the occurrence of one or more of the parameters listed in the packet formatted in XML is observed (e.g, continuous access to a particular port on the system from a specific IP), the

Fig. 78.13 Service composition in NIDS-SOA



analyzer service informs the Sender service to update the database and the reactions services to take some actions such as blocking the attacker’s IP or port in the firewall. Figure 78.13 shows the composition of all these services.

78.6 Conclusion

The use of the framework NIDS-SOA presents great advantage because it can combine the data system intrusion detection, a priori, work in isolation. The union of these information will help in protecting computer systems against evil actions. The results provided by NIDS-SOA are satisfactory, because the functionalities of IDS-NIDIA can be reused by other IDS through the implementation of services described in the WSDL of NIDS-SOA.

In future research works, we aim employ NDIS-SOA to extend other IDS in order to create a cooperative network constituted by several IDS, thanks to the interoperability and the support provided by the SOA approach.

Acknowledgments This research work was sponsored by Government of Maranhão State through FAPEMA and Federal Government through CNPq and CAPES.

References

1. Nakamura E, Geus P (2007) *Segurança de Redes em Ambientes Corporativos*. São Paulo Ed, Novatec
2. Brandão J, Fraga J, Mafra P (2008) “Composição de IDSs Usando *Web Services*.” V Simpósio Brasileiro em Segurança da Informação e de Sistema Computacionais, pp 339–342
3. Lima C (2002) *Agentes Inteligentes para Detecção de Intrusos em Redes de Computadores*. 2002. 176f. Dissertação, Universidade Federal do Maranhão. São Luis, MA
4. Lauffer R (2011) *Introdução a Sistemas de Detecção de Intrusão*. Disponível: <http://www.gta.ufrrj.br/grad/03_1/sdi/sdi-1.htm>. Acesso em 20 set. 2011
5. Abinader J, Lins R (2006) *Web Services em Java*. Rio de Janeiro. Ed. Brasport
6. What is a Web service? Available at <<http://www.w3.org/TR/ws-arch/#whatis>>. Accessed 20 Sept 2011
7. Júnior F (2005) *Proposta de atualização automática dos sistemas de detecção de intrusão Por meio de web services*. 2005. 166f. Dissertação, Universidade Federal do Maranhão. São Luis, MA
8. Michels P (2007) *Coreografia de serviços web*. Monografia de Conclusão de Curso. 2007, 76f. Universidade Federal de Santa Catarina. Florianópolis, SC
9. Michael P (2011) Papazoglou, Paolo Travesso, Schahram Dustdar, Frank Leymann. *Service-Oriented Research Roadmap*. Available <<http://infolab.uvt.nl/pub/papazogloup-2006-96.pdf>>. Accessed 31 Oct 2011
10. de Sene Fonseca J, Abdelouahab Z, Lopes D, Labidi S (2009) A security framework for SOA applications in mobile environment. *Int J Netw Security Appl (IJNSA)*1:90–107
11. José R (2008) *Orquestração e Composição de Serviços Web Usando BPEL*. 2008. 87f. Dissertação, Universidade de Aveiro
12. The Common Intrusion Detection Framework (CIDF). Available at <<http://gost.isi.edu/cidf>>. Accessed 19 Oct 2010
13. Siqueira L, Abdelouahab Z (2006) A fault tolerance mechanism for network intrusion detection system based on intelligent agents (NIDIA). The fourth IEEE workshop on software technologies for future embedded and ubiquitous systems, and the second international workshop on collaborative computing, integration, and assurance. SEUS/WCCIA 2006, pp 49–54. IEEE Computer Society
14. Moraes F (2009) *Security and reliability-based IDS agents*. 108f. Dissertation (Master in Electrical Engineering) Federal University of Maranhão. Sao Luis, MA
15. Oliveira AAP, Nascimento E, Abdelouahab Z (2005) Using honeypots and intelligent agents in security incident responses and investigation of suspicious actions in interconnected computer systems. In: *Proceedings of the E-crime and computer evidence conference 2005*. Technip, Monaco, pp 13–25
16. Abdelouahab Z, Costa Claudino Silva E (2006) Management and integration of information in intrusion detection system: data integration system for IDS based multi-agent systems. In: *Proceedings of the web intelligence and international agent technology workshops (2006) WI-IAT 2006 Workshops*. IEEE/WIC/ACM international conference on 2006, Hong Kong pp 49–52
17. Oliveira E, Abdelouahab Z, Lopes D (2006) Security on MASs with XML security specifications, IEEE 9th international workshop on network-based information systems (NBIS 2006)

18. Silva Lopes M, Lopes D, Abdelouahab Z (2006) A remote IDS based on multi-agent systems, web services and MDA. In: IEEE international conference on software engineering—ICSEA 2006. IEEE Computer Society
19. Lopes da Silva A, Abdelouahab Z, Lopes D (2009) Intelligent IDS for mobile devices: modeling and prototyping. *Int J Commun Netw Inf Security* 1:12–23
20. Netbeans (2011) Available at <<http://netbeans.org/>> Accessed 19 Oct 2011

Chapter 79

Lyrebird: A Learning Object Repository Based on a Domain Taxonomy Model

Ingrid Durley Torres, Jaime Alberto Guzman Luna
and Jovani Alberto Jimenez Builes

Abstract Taking advantage to a maximum of the records that contain data regarding learning objects has been a widely studied task in various fields of research. Web semantics is one of the most promissory; this paradigm is used with the purpose of dealing with the significant heterogeneity expressed by different users of objects themselves (authors, teachers and students). *Lyrebird* was born under this perspective; it is a repository of learning objects which provides sufficient semantic information to facilitate the reuse of learning objects ranging from processes such as classification, search and retrieval of information implementing three ontologies which are: one to specify a Domain of knowledge, another to define a learning object and lastly a to model metadata.

I. D. Torres (✉) · J. A. J. Builes
Artificial Intelligence for Education Research Group, Universidad Nacional de Colombia,
Medellín Campus, Medellín, Colombia
e-mail: idtorresp@unal.edu.co

J. A. J. Builes
e-mail: jajimen1@unal.edu.co

I. D. Torres · J. A. G. Luna
SINTELWEB Research Group, Universidad Nacional de Colombia,
Medellín Campus, Medellín, Colombia
e-mail: jaguzman@unal.edu.co

J. A. G. Luna · J. A. J. Builes
School of Systems, Universidad Nacional de Colombia,
Medellín Campus, Medellín, Colombia

79.1 Introduction

A natural activity in the field of Learning Objects (LO) is represented by compilation processes in containers that organize them and make them available for various uses. These containers are known as Learning Object Repositories (LORs) [1]. LORs, in addition to facilitating storage, also promote reuse and interoperability of their own LOs which are readily available within them. Nevertheless, it is not as easy a task as people think due to their multiplicity and the autonomy of suppliers which do not have a significant conceptual model widely accepted by suppliers themselves to use and specify LOs, nor is it common acceptance by the end users of those LOs.

Thus, more and more efforts have been focused on the development of technologies that allow standardization mainly covered by a set of metadata which may be used to describe an LO's main features [2] or to establish packaging norms within a structure having educational content [3]. Despite said standardization, the reuse of LOs is still a complex issue, and even more, when you add user diversity connected to the process (authors, teachers and students). Arriving at a comprehensive solution in which all the components involved (LOs and students) perform their tasks, interact and obtain intended results is a complex job that requires important efforts to achieve partial or total communication and interaction between software applications (LO) and people [1].

An emerging solution being applied successfully consists in using ontological models that make it possible to define a unique specification that provides an unequivocal comprehensible interpretation for the various parties (software and people).

Current repositories as MERLOT [4] and CAREO [4], already provide a first approach in the work to implement semantics related to an LOR scope. Nevertheless, even if they allow software agents which are no longer exclusively human to search and look up data related to them, their only concern is centered only on this process. SLOR [5], itself, involves another important aspect such as the definition of a shared conceptualization concerning the proper specification of what an LO is; thus, supporting the attainment of a more flexible LOR. Nonetheless, it omits domain intentionality. Hence, one can state that the three projects converge in a way that they contribute to what may be considered a prototype for autonomous data processing. That prototype inspired *Lyrebird*, which extends the aforesaid features adopting three ontological models that are able to represent: (i) the description of the definition of an LO, (ii) a semantic specification of an IMS-MD standard (named SIMS—Semantic IMS), and lastly (iii) an ontology which will be used to represent all the concepts that symbolize the intentionality of learning in a domain of particular knowledge.

To provide a more detailed vision in this paper, the article is organized as follows: [Sect. 79.2](#) defines the main concepts related to this proposal's framework. [Section 79.3](#) describes each of the ontologies developed. [Section 79.4](#) presents all the details of repository architecture along with a brief description of the

functionality of each one of the component modules. [Section 79.5](#) is an experimental case with its respective test results. Finally, [Sect. 79.6](#) compiles conclusions and future work arising from the development proposed herein.

79.2 Frame of Reference

79.2.1 Description of the LOs

Formally, there is not just one definition of the concept of an LO [1, 2, 6]. Nonetheless, it is convenient to consider it as an attempt to unify; the following definition: an LO will be understood as all the material structured in a significant way and it must be related to a learning objective which must correspond to a digital resource that can be distributed and consulted online. An LO must also have a registration form or metadata that includes a list of attributes which not only describes the possible attributes of an LO, but also allows to catalogue and exchange it. In this setting, standardization is a notably recurrent topic since when one handles various types of resources for different applications and with different technologies, it becomes a key topic to continue operating current applications and even to make them grow.

Among these initiatives it is worth highlighting: (i) an LOM [2]; this standard specifies the syntax of a minimum set of metadata required to complete and to adequately identify, administrate, locate and evaluate an LO. Its purpose is to facilitate the task of searching, sharing and exchanging LOs for authors, students and automatic systems. A second standard is the (ii) IMS (by the *Global Learning Consortium*) [6], it is a consortium whose mission is to develop and promote open specifications to facilitate online learning. Its objective was the design of a form to put into practice IEEE and AICC recommendations. To do so, an XML-type file was defined to describe course contents. This is done in such a way that any LMS may upload the course reading its set up file, IMSMANIFEST.XML. The following is only a description of this committee's main initiative: (i) *IMS Learning Resources Meta-Data*. This specification provides a structure for the elements (metadata) that describe or catalogue learning resources. This specification is based on an LOM. (ii) *IMS Learner Information Package (LIP)*. This corresponds to the interoperability of systems and student information to support an internet learning environment. (iii) *IMS Learning Design*. This design provides a flexible generic language to express various pedagogical models. IMS-LD, is based on a series of elements that include: roles that people play (who), the activities they perform (do) and the settings where they perform (services) and what they use to perform them (LO), all within a series of simultaneously executed acts.

79.2.2 Web Semantics and LO Search and Retrieval

Formal models that support web semantics [7], provide more knowledge for contents (images, videos, links), enabling the automation of many tasks currently performed by humans. In particular, semantics seeks to produce a world where ontologies [8] allow greater task automation by structuring resources available on the web, so that software agents may analyze and execute processes such as searching, retrieving, invocation, interoperability, and automatic execution [9]. To fulfill these tasks, said ontologies must be sufficiently expressive and must be able to describe the properties of related Domains.

Historically, Web semantics was introduced with an Resource Description Framework (RDF) [10] which allows the representation of classes, properties, sub-classes and more class hierarchies. That RDF has evolved into a more expressive language called OWL [11]. OWL (1.0) enabled the following properties: (i) class definition by means of restrictions on property, values or cardinality; (ii) class definition by means of Boolean operations on other classes as intersection, union and complement; (iii) relations among classes (for ex. inclusion, disjunction, and equivalence); (iii) properties of relations (as inverse, symmetric, transitive); (iv) cardinality (i.e. “just one”); (v) equality and inequality of classes; (vi) enumerated classes; (vii) cardinality restrictions; (viii) asymmetrical properties, reflexive and disjunctive, among others.

79.3 Ontological Formalization

79.3.1 Semantic Characterization of a Learning Object

As mentioned in the previous section, LOs adjust to the definition described in Sect. 79.2.1. To mold this sort of LO, it is necessary to approach each element mentioned in this definition indicating how it will be characterized semantically. The common one and apparently the simplest refers to a web resource, which according to W3C corresponds to that which has identity (whether it be called video, audio, and text among others), identified by a URI, that resides on the internet and which is accessible by means of any implemented version of an http protocol or its equivalent. The second element defines metadata and in this aspect, there is one already referenced, an IMS-MD standard, which focuses on the description of an LO. The third relates to the formative intention, which in this case refers to the knowledge that will be acquired with that LO.

The three elements have been represented semantically by the ontology described in Fig. 79.1. Nevertheless, to move these specifications semantically implies constructing in only one language a description that is not only easy to read and understand, but also easy for the user to write and which at the same time allows the exchanging and processing of data via internet without human intervention.

ontologies is that they may be extended at any given moment without any computational cost.

It is important to highlight that IMS *Learning Resource Metadata* (IMS-MD) has adopted an LOM standard. This is a reason why the SIMS semantic specification coincides with three of the nine main categories of an LOM, that group the rest of fields of the standard which are: (i) *General* (identifier, Title, Language Description), (ii) *Technical* (Format, Size, Requirement, Duration), (iii) *Educational* (Learning Resource Type, Difficulty); nonetheless, the fields grouped in these three categories have not been totally used either. A semantic representation of an IMS-MD allows a user to construct metadata instances that the user will use to describe each one of the LOs, just as it was done syntactically. Nevertheless, the great advantage lies in that the fields that formerly were optional or freely-expressed using natural language are now represented as “facts” which a machine can understand and process, despite being found in “human knowledge”; thus empowering a base of knowledge upon which complex reasoning and inferences can be made.

The process of semanticizing the metadata referenced in the paragraph above requires considering their specification in an OWL ontological language. This development generally implies three steps: (i) defining classes of ontology, (ii) hierarchically organizing classes (class-subclass), (iii) defining the properties of the classes and values allowed for them, (iv) creating instances assigning values to properties.

The first and second steps are represented by the structure defined by the LOM model. Thus, the root node corresponds to a *Metadata* concept while their direct off springs will be represented by the three categories selected and referenced as a “parent element” (General, Educational and Technical). The sub-categories of these elements correspond to the next level of hierarchy and in it we found the elements that have been enumerated as “off springs of this parent element”. Nevertheless, there is a specification that has been extended from the original standard which involves the definition of the properties of the classes and of the characteristics of these properties. Among those characteristics, it is possible to mention rank, Domain, cardinality and permitted values for the enumerated types among others. To detail this behavior, this paper indicates that a class has one or more properties (commonly called *has <nom_property>*), for each one of the characteristics it groups. Those properties may be at the same time *ObjectProperty*, and through both of them, it is possible to relate some classes and objects with others. Properties may also be *DataProperty* type when the property is a simple data type. This type of simple datum corresponds to a basic element of an XML schema type (*String*, *boolean*). Finally, for each of the properties of each class there will be a definite domain (*Domain*); in other words, to specify the object upon which it is applied and the range (*Range*), which corresponds to the type of values that it may take from the type of object (*ObjectProperty*) or the type of data (*DataProperty*). Figure 79.1 represents the abstract model of the *SIMS: Metadata* ontology. For example, one of the properties of language is to correspond to one and just one of the elements on a list of languages. This category then

refers to an *SIMS: Language* class because it is to whom it is applied; this being the domain as long as its range is of the *xsd: String* type of data which allows taking one of these values: sp (*spanish*), en (*english*), and ru (*russian*), among others.

79.3.3 Semantic Characterization of Formative Intentionality

Since one of the current priorities lies in the reusing of the teaching content; an LO will always be used as device to acquire the same knowledge. This is perhaps the main reason why when an LO is born, its creators provide it with the greatest freedom of use and the association of the formative intentionality becomes the responsibility of the person who implements it in a learning setting.

Under these guidelines, the semantic formalization of this element is directly associated with an instance of the *concept* class, of the *Learning Ontology Domain* (LOD) ontology. With the development of LOD, it is necessary to demonstrate how it is possible to manage LOs with a unique information model and this way face the semantic heterogeneity associated to the diversity of meanings present among various participants (LO creators, teachers and students) involved in a learning process. This way it is possible to define a non-ambiguous shared vocabulary that involves various mental representations that a participant has for each concept, superimposed on contrasting syntactic specifications. In this sense, this ontology allows classification, storage, search and retrieval of LOs that have been instantiated under such concepts, and at the same time, it is an important part of the description of course contents and the formulation of the requirement of the learning objectives a user wishes to fulfill [12]. For this specific case, an LOD has been inspired on the SKOS-Core Model [13] (see Fig. 79.2).

SKOS-Core defines an RDF vocabulary to describe the structure and contents of a variety of a concept schema, as LOs instanced with the concept of “offspring” generating the possibility of duplicating or not precising correctly the objectives to be taught nor the objectives which are to be reached. This is the main reason why it was decided to retake only the definitions of concept (*skos: Concept*), which is defined as a unit of thought that can be described. Likewise, the idea that each concept may only have a preferred label has been retaken, which is what documentalists call descriptor or preferent term and a limited number of alternative labels called non-descriptor or non-preferent. The coding of the labels corresponds to preferent and non-preferent terms belonging to a concept which is carried out by means of *skos: prefLabel* and *skos: altLabel* properties respectively. This second label represents the relation of synonymy or equivalence between two concepts. Moreover, you contemplate the *skos: Hidden*, which allows the generation of the same *skos: prefLabel*, but this time considering some orthographical errors or even typos. Inclusively, specification of *skos: languageconcept* was contemplated to specify the same concept in other languages. On the other hand, relations defined by *kos: Semantic Relation* have been completely ignored.

Fig. 79.2 Definition LOD, in the domain of robotics

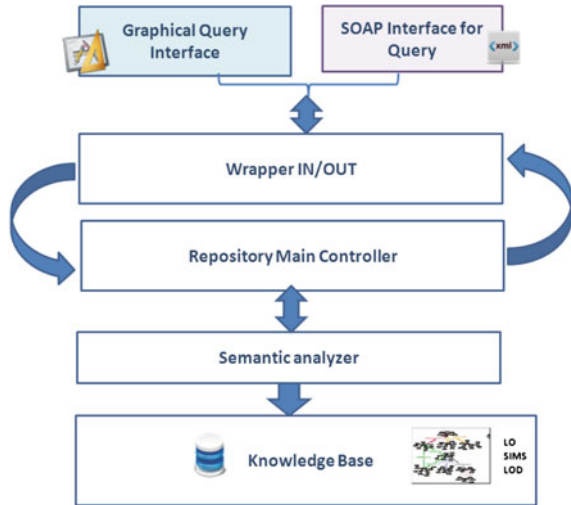
| robotics | |
|---------------------------|-------------|
| skos:semanticRelation = | pioneers |
| | history |
| | legislation |
| | ... |
| skos:altLabel = | robotics |
| | robotica |
| skos:narrowerTransitive = | pioneers |
| | history |
| | legislation |
| | ... |
| skos:narrower = | pioneers |
| | history |
| | legislation |
| | ... |
| skos:prefLabel = | Robotics |

79.4 Architecture of the Semantic Repository

The architecture of the semantic repository described in Fig. 79.3 is specified in five basic levels: the first corresponds to what has been called graphic search interface; it acts as a bridge between a user and a Lyrebird repository.

The SOAP interface appears at the same level, which allows the unification of search criteria issued by the wrapper standardizing it to an ontological concept defined in an XML file. This case only occurs when the wrapper explicitly requires its service. The following level is the wrapper; it is a software program that receives String or XML files and it is in charge of sending petitions or receiving results from the semantic analyzer. It acts with the SOAP interface only when it must send the string it received from the interface to have it unify it to a concept. The third level is represented by the main controller of the repository; in this case, it is a typical Model Vista Control (MVC) since it does not execute a thing; it is not an administrator that coordinates other components in compliance with the order from the GUI standardized by the *wrapper*. The fourth level corresponds to a semantic analyzer which is a software in charge of interpreting the capture of each search the controller has structured; to do so, the controller constantly uses the last level, which is where each one of the ontologies described in the previous number. If the order is to store in function of a specific LOD concept, it generates an OWL instance for each one of the ontologies and stores them persistently in a database. If it is otherwise and the petition is to search or visualize, it works along with the standardized order issued by the wrapper to search amongst all the stored OWL models; those that are specifically related to the concept of the order.

Fig. 79.3 Architecture of the *Lyrebird* semantic repository



79.5 Experimental Case

The functioning of the repository depends on the role the user identifies with, which corresponds to the traditional model of an ROA (administrator, teacher and student). The user having most functionality is the administrator user therefore, it is explained first. The administrator counts on various cases of use: (i) One which corresponds to administration as a “super user” of all the users; (ii) Receive the request to store an LO, in this case, *Lyrebird* enables the metadata fields associated to SIMS, so that this user may store them the way the user sees it fit. In this process, besides demanding the association of one or more concepts that can be associated to the formative intention of this LO, these concepts can only be selected directly from the same LOD being used at that given moment. To help users, a system of meanings has been enclosed which will briefly describe each concept; this option is activated when a user points with the mouse to each concept; (iii) Search, retrieve and show summarized information of the metadata that identifies the LOs related to the concept of ordering a search, or (iv) finally visualizing the LO properly in a pop-up window. It is important to mention that (ii), (iii) and (iv) are also implemented for a teacher-user; while the student-user will only have options (iii) and (iv) activated. Because of lack of space, only case (iii), shown in Fig. 79.4 will be mentioned.

As you can see, at the upper left hand side of the graphic interface are the administration options and the respective interface of each case being used, which as it was previously mentioned, is alike for the three users. In this case, it is required to search a specific concept of the LOD ontology, which as an example formalizes the knowledge surrounding the domain of the principal concepts of robotics (however, it may be replaced by another in another domain, as long as it

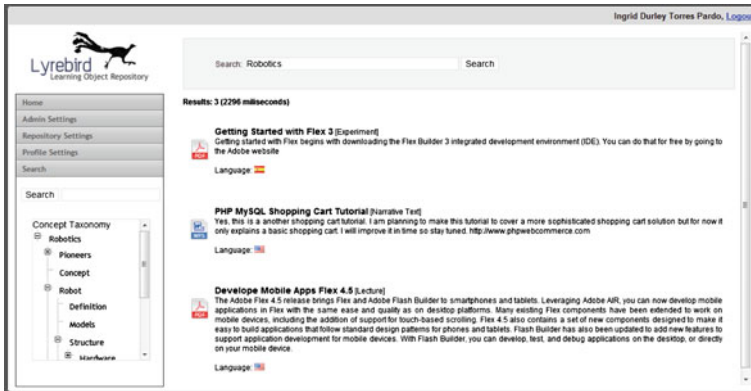


Fig. 79.4 Graphic interface search in Lyrebird

always complies with the OWL specification, SKOS-Core described in Sect. 79.3.3 regarding these searches). Continuing with the example, the search showed that there is a wish to search and retrieve all the LOs which have formative intentionality related to the concept of “Robotics”, which in this case has been keyed in the text field, was not directly selected from the ontology; even though, an equal option is not activated. Moreover, this field allows the keying in of the concept considering some word typos and will return the results as long as that that error is error is stored in the LOD specification in the *skos: Hidden* field. Likewise, a search can be conducted looking for some synonym concepts (Robot) obtaining the result.

Finally, as one can see, the system returns all the LOs marked semantically with that concept when they were stored along with them it shows the SIMS metadata recorded for each LO.

The evaluation of a data retrieval system [14] consists in measuring user satisfaction rates. To do so, the three following elements are typically on hand: (i) an LO collection; (ii) a set of searches and (iii) a collection of relevant LOs for each search emitted by human judgment.

To evaluate the Lyrebird system, 50 test LO were used from LOD which belonged to different concepts of Robotics such as: Robotics, Sensors, Movement, Exploration, Model and Structure. Collection of searches: Fig. 79.5, presents each search with its results.

This figure expresses the number of recovered and considered relevant LO is 98 % of all stored LO. This is because the system frees the user to sort them and store according to their own criteria of relevance, the same criteria used in the recovery.

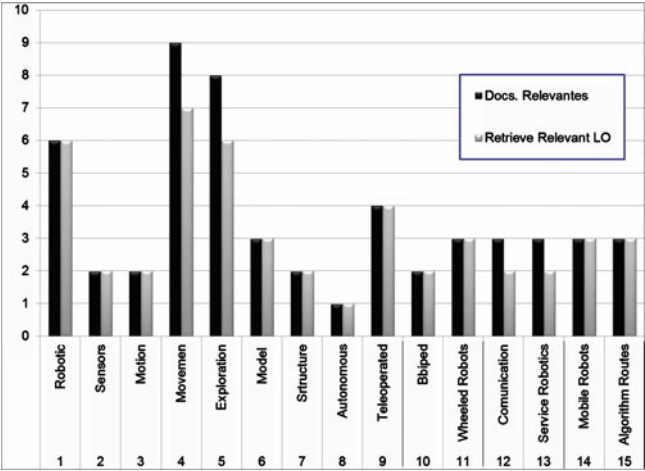


Fig. 79.5 Retrieved relevant LO Vs storage LO

79.6 Conclusions and Future Work

This article developed an LO semantic repository called *Lyrebird*; it allows storing and semantic searches concerning a specific domain taxonomy; The system developed infers on a knowledge base to find relations among concepts defined in the ontology of the system; moreover, it introduces the use of ontologies to specify metadata and the definition of what a *Lyrebird* LO is; also, it is accompanied by a graphic user-friendly interface which enables navigating through a taxonomy offering additional help represented in the visualization of concepts and their meanings The main advantage of this implementation lies in the possibility of empowering storage and search processes not only from a preferent concept, but also from other synonym concepts or even mistaken, which it interprets as the same preferent. Thanks to the inference process. All this formalization manages to solve process participants’ significant heterogeneity (people and software).

As future work, the intention is to include compound LOs represented by an IMS metadata, packaged in SCORM standard, and allow the use of more than two domain ontologies at the same time within ROA and domain ontologies at the same time.

Acknowledgments This work is part of the preliminary results of master’s thesis project “Reactive Planning Model for Dynamic Routing Composition of Learning in Virtual Environments and Uncertain Heterogeneous”

References

1. López C (2007) Los Repositorios de O.A como soporte a un entorno e-learning. http://www.biblioweb.dgsca.unam.mx/libros/reposireposi/objetos_aprendizaje.htm
2. IEEE Standards Department (2002) Draft standard for learning object metadata. IEEE Publication P1484.12.1/D6.4, Marzo
3. Thropp S (2010) Sharable content object reference model 2004: overview. 2a. edición. <http://www.adlnet.org/scorm/index.cfm>. Consultado Junio
4. Campbell K (2004) Effective writing for e-learning environments. Published in the United States of America. ISBN 1-59140-216-6. Information Science Publishing
5. Soto J, Arrion C, García E, Sanchez S (2006) Problemas de Almacenamiento e inferencia sobre grandes conjuntos de metadatos en un repositorio semántico de objetos de aprendizaje. In: Proceedings of the third symposium on objects and designs Pluridisciplinar learning, Oviedo, Spain
6. IMS (2000) IMS learning resource meta-data specification: version 1.1 final specification. IMS global learning consortium. <http://www.imspj.org/metadata/index.html> [Kautz y Selman 1996] Kautz HA, Selman B (1996) Pushing the envelope: planning, propositional logic and stochastic search. In: Proceedings of the national conference on artificial intelligence (AAAI), vol 2, pp 1194–1201
7. Berners-Lee J, Hendler T, Lassila O (2001) The semantic web. Sci Am 284(5):34–43. <http://www.scientificamerican.com/>
8. Gruber TR (1993) A translation approach to portable ontology specifications. Knowl Acquis 5(2):199–220
9. García FJ (2004) Web Semántica y Ontologías. In: García JF, Moreno M (eds) Tendencias en el Desarrollo de Aplicaciones Web. Departamento de Informática y Automática, Universidad de Salamanca, Salamanca, pp 1–23
10. RDF (2011) “RDF semantics”. W3C recommendation, 10 Feb 2004. <http://www.w3.org/TR/rdf-mt/>. Consultado Enero 2011
11. OWL (2009) <http://www.w3.org/TR/owl-features/>
12. Kontopoulos E, Vrakas D (2006) An ontology-based planning system for e-course generation. Expert Syst Appl 37(1):398–406 (Elsevier)
13. SKOS-Core (2007) http://www2.ub.es/bid/consulta_articulos.php?fichero=13perez2.html
14. Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval. Addison-Wesley, New York

Chapter 80

Design Process and Building Simulation

Heitor da Costa Silva, Clarissa SartoriZiebell, Lennart Bertram
Pöhls and Mariana Moura Bagnati

Abstract In the beginning of mankind's architectural history the resulting form was basically distinguished by the style, the material to be found near the construction site and finally the laws of physics—as by try and error the constructive systems had been developed. Over the time, other components have been added; the form finding process of architecture has been changed especially by the advancing technology and the possibility to mathematically foresee the physical outcome. Today simulation is able to very realistically calculate building performances, such as the thermal behaviour. Among many other components, laws and legislative restrictions emerge to cope with the development of societies and their ever-growing need to organize their cities. Nowadays, modern law-making is trying to deal with one of the major concerns for the future of our species: the lack of resources to sustain the world population with its 7 billion people. The conservation of energy sources around the world is playing a major rule, as fossil resources are soon to be exhausted and atomic power generation is in the hand of only a little number of states. Further, the production of electrical energy via carbon or oil is named one of the main responsibilities for air pollution and climate change. In Brazil the concerns started in the 1980s, with the oil crisis, but have been carried on to become part of the legislation only in the 1990s. In February of 2010 a proposal to evaluate the energy efficiency of buildings called Label PROCEL has been published.

H. da Costa Silva (✉) · C. SartoriZiebell · L. B. Pöhls · M. M. Bagnati
Federal University of Rio Grande do Sul Brazil, Sarmento Leite Street 320/215, Porto Alegre, RS, Brazil

80.1 Introduction

One of the major concerns for the future of our species is the lack of resources to sustain the world population with its 7 billion people. The conservation of energy sources around the world is playing a major role, as fossil resources are soon to be exhausted and atomic power generation being, with all its concerns and problems, in the hand of only a little number of states. Further, the production of electrical energy via carbon or oil is pointed at to be responsible for a significant part of today's air pollution and finally for the climate change.

Edification is playing a major rule in this scenario, as buildings use 36 % of the electrical energy consumed worldwide [1]. In developed countries the oil crisis and the buildings sector's high energy consumption led to the implementation of efficiency regulations regarding buildings [2]. However, the authors strongly believe that it lies in the architect's responsibility to integrate energy efficiency into the form finding process.

In February of 2010, in Brazil, a proposal to evaluate the energy efficiency of buildings called Label PROCEL [3] was published and plans to make such evaluation mandatory are in process of assessment. As a consequence of this process, the integration of energy concerns in architectural design education has started only very recently when compared to architects education in other parts of the world. This paper is a first analysis of thermal simulation regarding its use for educational purposes concerning the design of the building's envelope. The simulation with Energy Plus (EP) [4] is executed to examine and compare the necessities of students during the form-finding process guided by the legal regulations and the expected thermal results. To do so, the authors employ a case study of three office building projects, created during a post-graduate course at the Federal University of Rio Grande do Sul (UFRGS), Brazil. The projects' different architectural results, equally guaranteeing optimal results in the proposed legislative's equation, are analysed using thermal simulations.

This paper is structured as follows: In *Case Study*, the equation and further initial information, given to the students during the course, are introduced. The section *Projects* lays out the architectonical results of the case study and the next section describes the performed simulations and their results in more detail, while the conclusions are drawn in the last section.

80.2 Case Study

The case study is based on three projects elaborated during the post-graduate studies in architecture of the Federal University of Rio Grande do Sul (UFRGS). Each project was elaborated by two students and, at the beginning of the course, the regulation, its equation and basic ideas were presented to the students. Further, the studies on the influences of the window design for the thermal comfort in

buildings in southern Brazil [5] were presented the start of the student's work on the project. Basically, this work determines a range of percentages of openings, such as windows, for each solar orientation, aiming at a comfortable inside climate. With this information at hand, the final goal of the course was the creation of an architectural project for a small office building in city-area of Porto Alegre, Brazil. In detail, the building-site as well as the space program were defined and the building regulations of the city had to be applied. Moreover, the project had to attend the highest standard (A) of the Label PROCEL. Nevertheless, the most urging issue for the students was Porto Alegre's humid subtropical climate with four well-defined seasons and a predominant warm season, combined with slightly high level of humidity year around.

The proposed label in its latest version [3], the so-called Label PROCEL, has been defined to be a plan to create the basis for rational energy consumption in buildings throughout Brazil. It defines referential parameters for the verification of the building's efficiency regarding electric energy. It is important to mention that the development of this regulation is basically sponsored by the National Electrical Energy Agency (Eletrobras), and is exclusively focused on the conservation of electrical energy.

This paper aims at analysing the potential use of thermal simulation in education, teaching the process of finding the project's architectural form. Today the same is guided only by the classical or modern form finding methodologies and maybe the technical regulations imposed by the state. As the case study included the Label PROCEL as orientation, students were introduced to the proposed legislative equation.

The general equation is composed of four macro elements, (1) the *building envelope*, (2) the *air-conditioning*, (3) the *artificial lighting* as well as (4) *bonuses for sustainable construction elements*. As the envelope basically represents the building's form, and air-conditioning as well as artificial lighting are considered non-architectonical details, the most logic step is to focus on the first, which impacts with 30 % on the general equation shown in Eq. 80.1 (Total Punctuation composed of : (a) Building Envelope, (b) Artificial Lighting, (c) Air-conditioning, (d) Bonuses). The same percentage is valid for artificial lighting while air-conditioning is weighted with 40 %.

$$\begin{aligned}
 PT = & 0.30 * \left\{ \left(EqNumEnv * \frac{AC}{AU} \right) + \left(\frac{APT}{AU} * 5 + \frac{ANC}{AU} * EqNumV \right) \right\}^{(a)} \\
 & + 0.30(EqNumDPI)^{(b)} \\
 & + 0.40 \left\{ \left(EqNumCA * \frac{AC}{AU} \right) + \left(\frac{APT}{AU} * 5 + \frac{ANC}{AU} * EqNumV \right) \right\}^{(c)} + b_0^{(d)}
 \end{aligned}
 \tag{80.1}$$

When looking at the part concerning the envelope (a), the regulation proposes 16 equations, two for each climate zone defined in the eight Brazilian climatic

regions. Which of the two is to be used depends on the size of the building to be evaluated. In more detail, one or the other equation is to choose, if the building's projected area is larger or less than 500 m². The part of the equation concerning the building envelope is depicted in Eq. 80.2 (Equation for the Building Envelope in the Brazilian Climatic Zones 2 and 3 for buildings with more than 500 m²).

$$IC_{env} = -14.14 * FA - 113.94 * FF + 50.82 * PAF_T + 4.86 * FS \\ - 0.32 * AVS + 0.26 * AHS - \frac{35.75}{FF} - 0.54 * PAF_T * AHS + 277.98 \quad (80.2)$$

In detail the variables are combinations of the following 5 elements:

- Factor Form (FF)
- Factor Height (FA)
- Horizontal Shading Angle (AHS)
- Vertical Shading Angle (AVS)
- Percentage of Transparent Elements in the Total Facade (PAF_T).

It is important to highlight that a lower numerical result of the equation corresponds to a better evaluation of the building. In other words, a building with a lower result is supposed to use less electrical energy.

The students elaborated the projects considering these five variables, which affect the building envelope. However, it was difficult for students to know the exact consequences of their design decisions related to the thermal comfort and energy efficiency. In this case, the building simulation would be very useful since it makes it possible to test each decision rapidly. The factors from Eq. 80.2 with the upmost importance for this paper are: the *horizontal shading angle* (AHS), *vertical shading angle* (AVS) and *percentage of transparent elements in the total facade* (PAF_T), since those variables will be differentiated during the simulations to be performed. As will be shown later, another variable will be involved in the simulations: the *location*, which is treated by Label PROCEL throughout the distinction of the before mentioned sixteen equations, each related to a different climate zone.

80.3 Projects

This section presents the visual and architectural results elaborated by the student groups. Looking at the three projects shown in Fig. 80.1 we can identify different solutions exploring distinctive architectural strategies. The floor plans of the three projects are shown in Fig. 80.2.

Despite their unique first appearance, the three approaches present certain common points in their approach to deal with the given tasks. All make reduced use of transparent constructions, such as windows, in the western facade as well as

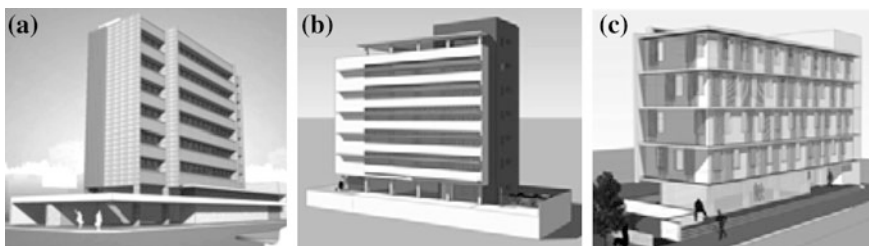


Fig. 80.1 3D views of the three projects as produced by the student groups. **a** Project A by G. Maia and R. Zampieri, **b** Project B by A. Feitosa and F. Pasquali, **c** Project C by L. Poehls and C. Ziebell. *Supervising Professor H. Silva*

all employ horizontal shading elements northwards and vertical shading elements eastwards. The students assumed the bar-type form as a result of the program's space necessities on a relatively reduced ground available for construction.

The presented work investigates the influence of simulation performed simultaneously to the form-finding process. Thereby the authors intend to demonstrate the possibilities of using thermal simulation as a part of the form-finding process and how to possibly integrate it into such workflow.

80.4 Simulations

In the following the thermal simulations performed with Energy Plus (EP) are presented. In more detail, EP's software modelling and the simulation parameters are described. The simulation's results for different situations are presented at the end of this section. It is important to highlight, that besides the investigation of the results from the regulation's point of view, the results of the thermal simulation are supposed to be part of the future workflow of educational architectural projects.

EP is developed by the United States Department of Energy and distributed without fees. Its main part consists of an energy analysis and thermal load simulation program. Based on a model of a building from the perspective of the building's physical make-up and associated mechanical and other systems, EP calculates heating and cooling loads necessary to maintain thermal control set points, conditions throughout a secondary HVAC system and coil loads as well as the energy consumption of primary plant equipment. Simultaneous integration of these details verifies that the simulation performs as the real building would [4]. The authors highlight that the quantity of parameters used is intentionally reduced and the models used to simulate the three projects have been intentionally simplified as far as possible to render the results easy to compare and quick to obtain. These are basic needs for the use of simulation during the design phase of a project, as more realistic simulation would cost too much time and effort.

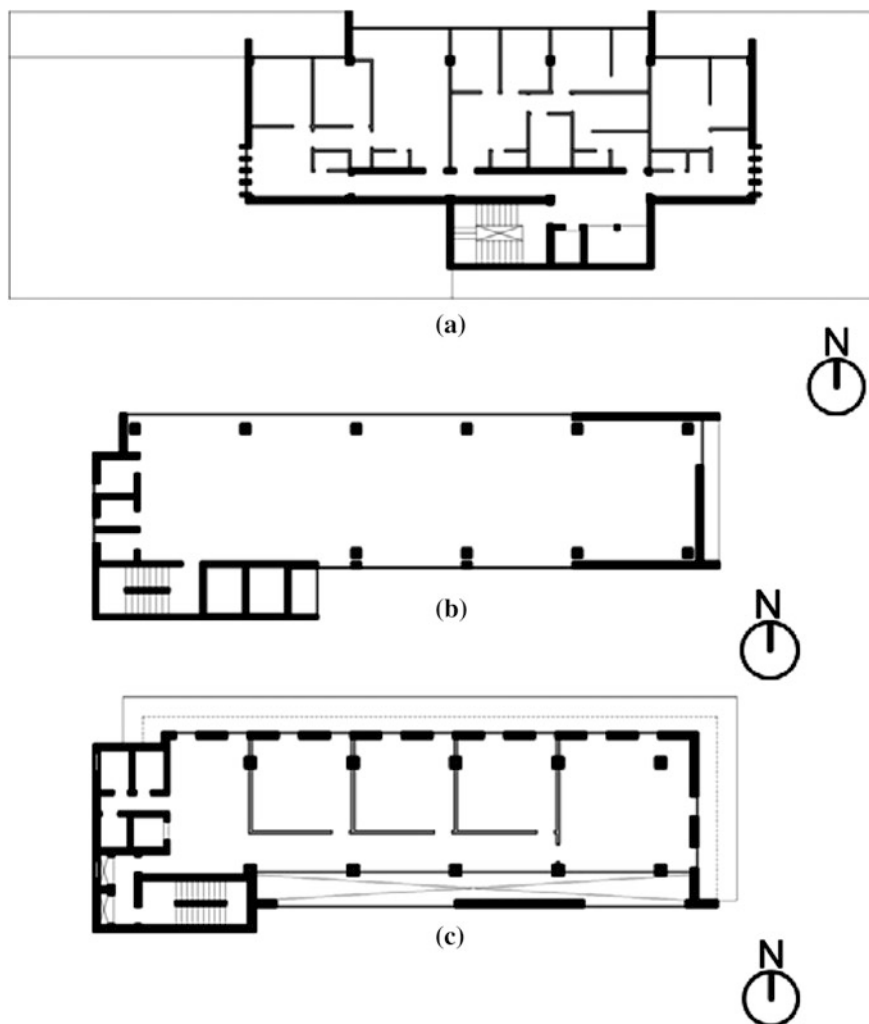


Fig. 80.2 Floor plans of the typical floors of: **a** Project A, **b** Project B, **c** Project C

To further help with the creation of the geometrical and mathematical models of the buildings a freeware plug-in of EP for Google's Sketch Up [6] called Open Studio [7] was used. It not only allows generating the basic geometry of the building volumes, thermal zones, windows and shading devices, but also defining the structural components such as outer or inner walls in the graphical interface of Sketch Up. This model's file is then imported into the simulation program and further parameters such as the weather data, setpoints, internal loads and schedules are set to obtain the output data. To use the simulation for the development of the building envelope, the students focused on simulation parameters which help to adjust design decisions. The results aim to compare the size of the external

| Field | Units | Obj1 | Obj2 | Obj3 | Obj4 | Obj5 |
|--------------------|--------|-----------------|-------------------|-------------------|-----------------|-------------------|
| Name | | F07 25mm stucco | F08 Metal surface | F16 Acoustic tile | G01 16mm gypsum | M11 100mm lightwe |
| Roughness | | Smooth | Smooth | MediumSmooth | MediumSmooth | MediumRough |
| Thickness | m | 0.0254 | 0.0008 | 0.0191 | 0.0159 | 0.1016 |
| Conductivity | W/m-K | 0.72 | 45.28 | 0.06 | 0.16 | 0.53 |
| Density | kg/m3 | 1856 | 7824 | 368 | 800 | 1280 |
| Specific Heat | J/kg-K | 840 | 500 | 590 | 1090 | 840 |
| Thermal Absorbance | | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 |
| Solar Absorbance | | 0.3 | 0.4 | 0.3 | 0.3 | 0.2 |
| Visible Absorbance | | 0.3 | 0.4 | 0.3 | 0.3 | 0.2 |

| Obj12 | Obj13 | Obj14 | Obj15 |
|---------------------------|------------------------------------|-----------------|--|
| I01 25mm insulation board | Concrete Block: Low Mass Aggregate | F11 Wood siding | Concrete: Sand and gravel or stone aggregate |
| MediumRough | MediumRough | MediumSmooth | MediumRough |
| 0.0254 | 0.15 | 0.0127 | 0.102 |
| 0.03 | 0.33 | 0.09 | 1.95 |
| 43 | 1390 | 592 | 2240 |
| 1210 | 880 | 1170 | 900 |
| | | | 0.9 |
| | | | 0.6 |
| | | | 0.6 |

| Obj6 | Obj7 | Obj8 | Obj9 | Obj10 | Obj11 |
|-----------------|-----------------|------------------|-----------------|-------------------------|--------------------|
| M15 200mm heavy | M01 300mm brick | F18 Terrazzo 1cm | F07 50mm stucco | insulation polyurethane | F08a Metal surface |
| MediumRough | MediumRough | Rough | Smooth | Rough | Smooth |
| 0.2032 | 0.3 | 0.01 | 0.05 | 0.05 | 0.0008 |
| 1.95 | 0.89 | 1.8 | 0.72 | 0.0245 | 45.28 |
| 2240 | 1920 | 2560 | 1856 | 24 | 7824 |
| 900 | 790 | 790 | 840 | 1590 | 500 |
| | | 0.9 | 0.9 | | 0.9 |
| | | 0.4 | 0.3 | | 0.2 |
| | | 0.4 | 0.3 | | 0.2 |

Fig. 80.3 EP simulation parameters: constructions

openings of the offices at the northern facade, the solar orientation of the building as a whole, as well as the influence of the climate on the envelopes form -all variables that the Label PROCEL’s equation intends to cover. For each of these parameters a series of tests has been developed, simulations have been performed and results, measured in hours of comfort, discomfort caused by heat as well as discomfort caused by cold, have been analysed for all the three case study’s projects. The setpoints for comfort are defined by a dry bulb temperature of 18.5 °C for the lower limit and 26.5 °C for the upper.

It is important to highlight that for the sake of better comparison some architectonical qualities expressed and shown in the final projects are ignored in order to generate simplified models. To further facilitate the implementation of the program as an educational tool, the models of the projects were designed with the minimum information necessary for their simulation. Likewise, the construction elements, such as walls, slabs, and openings, were determined equally for all three projects, disregarding materials or colours indicated in the projects’ renderings. The regarding menu of EP is shown in Fig. 80.3.

Furthermore, it was determined that only the areas of prolonged presence of persons, such as shops and offices, were to be artificially conditioned by *Packaged Terminal Heat Pumps* (PTHP).



Fig. 80.4 3D model: Project A

To correctly measure the air conditioning system, EP needs the user to define the so-called project days. These days are actually calculated by the simulation software when the weather data is given as input and the user can choose the design days to be considered for the simulation's output. The weather files used for the presented simulations were obtained from [8].¹

In the following, three series to analyse the parameters are described in more detail. In the first series the influence of the relation of transparent openings to opaque wall construction in the northern facade is analysed. It is important to note that in Brazil the northern and western facades are the most critical when regarding the thermal impact by solar radiation. In the first simulation the window size of the north facade is raised by 50 %, in the second simulation the windows are reduced by 50 % with respect to their original size.

The second series concerns the solar orientation of the building. To outline its importance the authors performed a total of three simulations, one for each 90° of rotation, exposing the original northern facade to the west, south and east.

To obtain information about the climate one last simulation has been performed. As the original simulation has taken place at the original location of the project, at Porto Alegre (30°S 51°W), the simulation for this series was situated in a very different climate. The authors chose Brasilia (15°S 47°W) in the Federal District located in the central-west of Brazil. In opposition to Porto Alegre's humid four seasons, Brasilia represents a desert-like highland with relatively constant dry and hot climate.

As mentioned above, the created models are reduced to the necessary elements. The performed simulation do not aim at highly realistic results, rather they should serve as orientation about the effects of changes in the architectural design on the thermal comfort. In more detail, in the model of project A the ground floor is divided into three zones: hall, vertical circulation, commercial room and garage. The typical floors are divided into five zones: the vertical circulation, the horizontal circulation, the east offices, west offices and the central group of offices.

¹ The Design Conditions used were obtained from Energy Plus database.

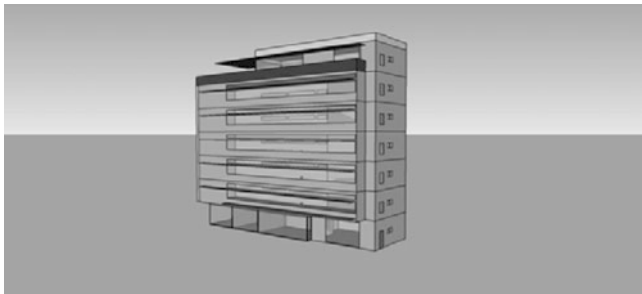


Fig. 80.5 3D model: Project B

Further, the central offices were grouped in a single zone, whereas they all have the same thermal characteristics. In more detail, they have contact with the exterior only at the northern facade. The top floor that corresponds to the engine room and water reservoir was established as a single zone. The resulting 3D-model is shown in Fig. 80.4.

The ground floor of project B's model is divided into four zones: the hall and vertical circulation as well as the three offices. The typical floors are divided into two zones: the vertical circulation and commercial area. The highest floor consists of two zones: the vertical circulation and the restaurant. The model generated in Sketch Up for this project can be seen in Fig. 80.5.

Project C's ground floor is divided into four zones: hall, vertical circulation, and three rooms. The typical floors are divided into five zones: the hall, the vertical and horizontal circulation, an area corresponding to an open space, the east room, and a group of central rooms. These last rooms were grouped into one unique zone, whereas they all have the same thermal characteristics. The top floor, the area corresponding to the engine room and water reservoir, was defined as a single zone. The visualization of project C is depicted in Fig. 80.6.

The results have been obtained by calculating the arithmetic mean value of all simulated thermal zones of the building. Vertical and horizontal circulations as well as other zones of transit have been simulated with a constant temperature and are not included in the calculation regarding the hours of comfort.

All others zones have been simulated with internal thermal loads according to their specified purpose. Table 80.1 also shows the hours of comfort (C) or discomfort for heat (DH) and cold (DC) during the hours of occupation. As the simulation regards a building with mainly office use, the authors simulated a five-day week and 8-h workday with 1 h of lunch-break at noon. It is important to highlight that the simulation considered a small number of persons outside these work hours to represent the safety and cleaning personal. Therefore, the authors calculated the occupied hours based on all hours that persons are in the building. Table 80.4 represents the total energy consumed in Megawatts per year for the projects of the case study.

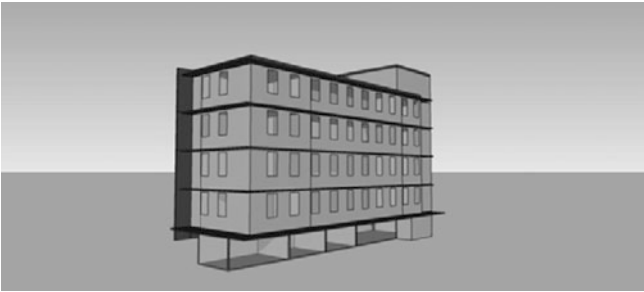


Fig. 80.6 3D model: Project C

Table 80.1 Thermal simulation results for Project A

| Project A | Occupied hours (h) | | | Total hours (h) | | |
|------------------------------|--------------------|-------|-------|-----------------|-------|-------|
| | C(h) | DH(h) | DC(h) | C(h) | DH(h) | DC(h) |
| <i>(a) Original version</i> | | | | | | |
| No changes | 4963 | 193 | 15 | 7016 | 1476 | 268 |
| <i>(b) Window size</i> | | | | | | |
| Plus 40 % ^a | 4966 | 193 | 13 | 6998 | 1517 | 245 |
| Less 50 % | 4952 | 206 | 14 | 6947 | 1545 | 269 |
| <i>(c) Solar orientation</i> | | | | | | |
| Rotate 90° | 4964 | 188 | 19 | 6923 | 1500 | 337 |
| Rotate 180° | 4969 | 182 | 20 | 6988 | 1393 | 379 |
| Rotate 270° | 4935 | 219 | 17 | 6844 | 1604 | 311 |
| <i>(d) Location</i> | | | | | | |
| Brasilia | 4915 | 256 | 0 | 6972 | 1784 | 3 |

^a It was not possible increase 50 % of window area
Subtitle Comfort (C), Discomfort for Heat (DH), and Discomfort for Cold (DC)

The results for each building from the Case Study are divided in the four parameters: (a) original version, (b) window size, (c) solar orientation and (d) location, with one, two, three and one simulation result(s) respectively. Equivalent observations can be made for the results regarding the projects B and C, which are depicted in Tables 80.2 and 80.3 respectively.

Table 80.4 shows the total energy consumed in each situation previously described.

Table 80.2 Thermal simulation results for Project B

| Project B | Occupied hours (h) | | | Total hours (h) | | |
|------------------------------|--------------------|-------|-------|-----------------|-------|-------|
| Changes | C(h) | DH(h) | DC(h) | C(h) | DH(h) | DC(h) |
| <i>(a) Original version</i> | | | | | | |
| No changes | 5191 | 411 | 27 | 7052 | 1443 | 265 |
| <i>(b) Window size</i> | | | | | | |
| Plus 50 % | 5194 | 408 | 27 | 7052 | 1436 | 272 |
| Less 50 % | 5186 | 416 | 27 | 7044 | 1455 | 260 |
| <i>(c) Solar orientation</i> | | | | | | |
| Rotate 90° | 5145 | 443 | 41 | 6699 | 1701 | 361 |
| Rotate 180° | 5183 | 413 | 33 | 7060 | 1355 | 345 |
| Rotate 270° | 5113 | 488 | 28 | 6673 | 1792 | 295 |
| <i>(d) Location</i> | | | | | | |
| Brasilia | 5096 | 533 | 0 | 6901 | 1836 | 24 |

Subtitle Comfort (C), Discomfort for Heat (DH), and Discomfort for Cold (DC)

Table 80.3 Thermal simulation results for Project C

| Project C | Occupied hours (h) | | | Total hours (h) | | |
|------------------------------|--------------------|-------|-------|-----------------|-------|-------|
| Changes | C(h) | DH(h) | DC(h) | C(h) | DH(h) | DC(h) |
| <i>(a) Original version</i> | | | | | | |
| No changes | 5033 | 225 | 6 | 7213 | 1351 | 196 |
| <i>(b) Window size</i> | | | | | | |
| Plus 50 % | 5034 | 224 | 6 | 7174 | 1393 | 193 |
| Less 50 % | 5035 | 222 | 7 | 7264 | 1268 | 227 |
| <i>(c) Solar orientation</i> | | | | | | |
| Rotate 90° | 5021 | 229 | 14 | 7068 | 1368 | 324 |
| Rotate 180° | 5028 | 219 | 16 | 7146 | 1259 | 355 |
| Rotate 270° | 5001 | 254 | 10 | 7054 | 1463 | 243 |
| <i>(d) Location</i> | | | | | | |
| Brasilia | 4961 | 303 | 0 | 7163 | 1595 | 3 |

Subtitle Comfort (C), Discomfort for Heat (DH), and Discomfort for Cold (DC)

80.5 Conclusion

The conclusions from the experience with the student sand the results obtained throughout thermal simulation of the three projects elaborated during the post-graduate course are presented in the following paragraphs.

This paper is seen as a first analysis of the use of simplified thermal simulations during the educational architectural process and draws its conclusions about the results from the thermal simulation with EP. Moreover, by using the described

Table 80.4 Total energy consumed (Megawatts/year)for Projects A, B and C

| Changes | Total energy consumed (Megawatts/year) | | |
|------------------------------|--|-----------|-----------|
| | Project A | Project B | Project C |
| <i>(a) Original version</i> | | | |
| No changes | 185528 | 149030 | 80647 |
| <i>(b) Window size</i> | | | |
| Plus 50 % ^a | 192697 | 154131 | 85005 |
| Less 50 % | 179929 | 145311 | 78062 |
| <i>(c) Solar orientation</i> | | | |
| Rotate 90° | 190053 | 162909 | 82083 |
| Rotate 180° | 182389 | 145641 | 79643 |
| Rotate 270° | 195276 | 167274 | 83562 |
| <i>(d) Location</i> | | | |
| Brasilia | 188835 | 150961 | 81289 |

^a The Project A had an increase of 40 % of its window area

seven scenarios, students aim at retrieving conclusions about the influence and potentials of simulation, regarding the importance of the fenestration area, the solar orientation as well as the location of the construction site. Martinez [9] explained the design process with the idea that “different projects can be developed based on the same sketch”. He also writes: “the process consists of passing from the steps of bigger generality and lower definition to steps of greater definition”. The software of EP, making quick and relatively simple thermal simulations possible, could be an tool to help students in this process.

It is important to emphasize that all software and most of the references of this paper are available on the Internet. As a result, the accessibility of the described simulation process is easy and quick. This facilitates the work of students, who want to use simulation as a tool during the design process.

The authors have to point out that the thermal simulations have not been performed by the project’s groups, but by the authors after the projects had been finalized. Therefore, the architectonical form of the presented projects does not reflect the influence of any experience with thermal simulations. Nonetheless, the authors strongly believe, that the simulations carried out have been realized in a way that could be easily included into the design process by post-graduate or even graduate students.

Regarding the different situations (window size, solar orientation and climate) the results of the thermal simulation show differentiated results. The results are depicted in the Tables 80.1, 80.2, 80.3 and 80.4.

All projects react with higher annual energy consumption when placed in the climate of the city of Brasilia instead of their original position. In more detail, all projects showed a reduction of the hours outside the comfort zone caused by low temperatures inside the building. This can be clearly attributed to the lack of low temperatures from the outside climate of Brasilia. On the contrary, the hours of discomfort caused by high temperatures are elevated significantly in respect to the

original project location. Therefore, the simulation not only indicates the negative effect on the energy consumption, but also indicates that its reason is the energy used for cooling down the inside of the building.

When looking at the results regarding the window size, the results of the thermal simulation do not show very distinguished results. Anyhow, the slight changes of about 5 % do indicate the expected results. In more detail, all projects are using less energy when the window size is reduced, which is logical, because the negative effects of insolation during hot periods and the heat loss during cold periods are reduced. It has to be taken into consideration though, that the comfort guaranteed by natural illumination is not regarded in such statements.

The simulation of the different solar orientations also results in a correct detection of the problematic situations by EP. All projects obtain their worst results for 90 and 270°, while all projects are consuming less when turned by 180°. This can be explained by the fact, that the little percentage of openings in the original southern façade is equally adapted to reduce negative effects of insulation on the northern facade. It is interesting to observe that the equation presented by the Label PROCEL does not consider the solar orientation. As a conclusion, the authors are convinced that EP can help with the form finding process; it distinguishes adapted solutions from not adapted ones for decisions like the facade orientations chosen for the envelope.

80.5.1 Final Conclusions

Finally, the simulation proofs to give all the correct indications and can even deliver detailed information about the reasons of good or bad outcomes regarding the energy consumption. The authors conclude that the various possibilities of evaluation, that this case study has explored only in a very simplified form, already proofed to be a guideline for the students. In future works, the authors of this paper plan to include the simulation into the educational experience. This will result in more detailed results, but the authors foresee that the simulation will continue to be an useful tool to help and to guide students along the non-linear process of design, it may even help to obtain better results in the equation for the Label PROCEL.

References

1. Statens energimyndighet (2010) Energy in Sweden—facts and figures. Energimyndighet, Sweden
2. Lamberts R, Dutra L (2004) Fernando Oscar Ruttkay Pereira. Eficiência Energética na Arquitetura, 2nd edn. ProLivros, São Paulo
3. Lobão E et al (2011) Etiquetagem de Eficiência Energética de Edificações. Brasília, Brazil
4. U.S. Department of Energy (2011) Energy plus energy simulation software. <http://apps1.eere.energy.gov/buildings/energyplus>. Accessed April 2011

5. Da Costa Silva H (1994) Window design for thermal comfort in domestic buildings in Southern Brazil. The Architectural Association School of Architecture School Environment and Energy Studies Programme, United Kingdom
6. Google Sketchup. <http://sketchup.google.com/>. Accessed April 2011
7. U.S. Department of Energy. http://apps1.eere.energy.gov/buildings/energyplus/openstudio_suite.cfm. Accessed April 2011
8. Laboratório de Eficiência Energética em Edificações. <http://www.labeee.ufsc.br/>. Accessed August 2010
9. Martinez AC (2000) Ensaio Sobre o Projeto. Publisher University of Brasilia (UNB), Brasilia

Chapter 81

Behavioral Models with Alternative Alphabets

Mohammed Lafi and Jackson Carvalho

Abstract Scenarios and properties are used to describe requirements specifications. Behavioral models can be synthesized from these scenarios and properties. New scenarios and properties are usually added to requirements specification after the creation of the behavioral models. Alphabet is the set of actions/events used to describe scenarios and properties. Alternative alphabets arise when a new scenario or property is added to the requirements with a new alphabet. Modifying behavioral models in the case of alternative alphabets need more considerations because existing synthesis algorithms must be rerun with the new alphabet. We propose additional steps that modify the behavioral model of properties to reflect new actions/events discovered in scenarios (or properties). Similarly, we propose additional steps that modify the behavioral model of scenarios to reflect new events (actions) discovered in properties.

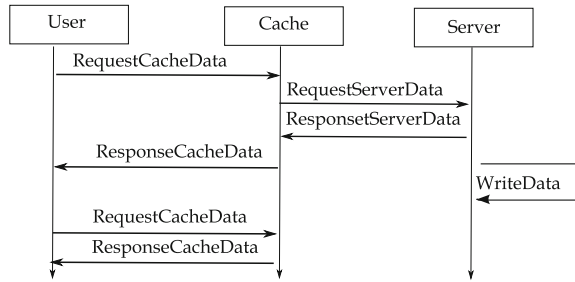
81.1 Introduction

Scenarios are written descriptions used to organize the information generated during requirements analysis [1]. These descriptions are typically in *story* form giving a sequence of events. Representing user requirements as a set of scenarios is an established activity in software engineering [2, 3] and its use makes the

M. Lafi (✉) · J. Carvalho
Electrical Engineering and Computer Science Department, University of Toledo,
Toledo, OH 43606-3390, USA
e-mail: Mohammed.Lafi@utoledo.edu

J. Carvalho
e-mail: Jackson.Carvalho@utoledo.edu

Fig. 81.1 A scenario, represented as an MSC, that shows the interactions between user, cache and server



communication between end users and software developers more efficient [4]. To assist with representing scenarios in intuitive and clear forms, *sequence diagrams*, such as the Unified Modeling Diagrams (UML) [1] and Message Sequence Charts (MSC) [5], are used. For instance, Fig. 81.1 [6] shows the MSC corresponding to the interactions between user, cache and server.

In addition to scenarios, *properties* can be used to add constraints and/or conditions that limit the behavior of a system [7, 8]. Properties are used to ensure that the system will not exhibit incorrect or disallowed behavior. Properties are often represented in forms such as the Object Constraint Language (OCL) [9] and Fluent Linear Temporal Logic (FLTL) [8]. For instance, Fig. 81.2 show the FLTL representation of the properties for the scenario shown in Fig. 81.1, respectively.

A *behavioral model* is a formalism that describes the behavior of the system using a sequence of state transitions [4]. It also provides a graphical representation for the new system to be developed [4]. Scenarios and properties can be used to create a behavioral model for the system to be developed [4] and such a model can be used to validate the user requirements [10, 11]. Similarly, scenarios may also be used to validate both system analysis and design [12].

Uchitel et al. [7] introduce an algorithm to synthesize a behavioral model from scenarios and properties. This algorithm constructs the partial behavioral model in three steps. Figure 81.3 illustrates these steps. First, it creates two separate behavioral models one for the scenarios and another for the properties. The resulting behavioral models are represented as Label Transition System (LTS).¹ This step will be referred to, in this work, as *creation*. Next, the algorithm converts the behavioral models to partial behavioral models represented as Model Transition System (MTS).² This step will be referred to, in this work, as *conversion*. Finally, the algorithm merges the partial behavioral models of both scenarios and properties. This step will be referred to, in this work, as *merging*.

Algorithms, that synthesize behavioral models from scenarios and properties, such as Uchitel et al.’s approach [7] and Krka et al.’s approach [6], do not consider alternative alphabets [7] if the known set of events before and after the creation of the behavioral models are different. When alternative alphabets occur,

¹ Label Transition System (LTS) is a formalism used to describe the behavioral model.

² Model Transition System (MTS) is a formalism used to describe the partial behavioral model.

Fig. 81.2 Constraints represented as an OCL on the scenario shown in Fig. 81.1

```

Fluent requestPending = < requestCacheData, response-
CacheData > initially false
Fluent cached = < resposeServeData, WriteData > ini-
tially false

```

these algorithms must be reapplied because they assume that the system's alphabet will not be affected by the addition of new requirements. However, this assumption is not practical because the requirements specification is usually built incrementally. New scenarios and new properties are expected to appear during the requirements specification process. Both new scenarios and new properties are likely to have new alphabets. A synthesis algorithm supporting alternative alphabets would offer a significant improvement over the mentioned algorithms because it will eliminate the re-run of the creation step.

In the case of alternative alphabets, the known alphabet is vital in the *conversion* step during the generation of the partial behavioral model [7]. This characteristic is because during the *conversion* step, the synthesis algorithm assumes the system's alphabet will not be affected by the addition of a new requirement. Then, it creates transitions based on this assumption for the generation of the transitions for its behavioral model. However, in the case of alternative alphabets, such alphabet will be affected by the addition of a new requirement. This means the generated behavioral models are not valid for modification and they need to be reconstructed.

A re-run of the *creation* step can be avoided by modifying the generated behavioral models to incorporate an alternative alphabet. Our decision to modify the behavioral models instead of reconstructing them is because the *creation* step is known to be the most expensive component of the synthesis algorithm [13].

In the following sections, we show how to modify both the partial behavioral model of the scenarios and the partial behavioral model of the properties. The solution proposed in this work is applied to the models generated during the *creation* step and/or *conversion* step. Our approach is composed of two fundamental components. One to address the new actions discovered in the new properties and another to address the new actions discovered in the new scenarios. This approach enhances the capabilities of the existing algorithms for synthesizing partial behavioral models by addressing the alternative alphabets.

This chapter is organized as follows. [Section 81.2](#) defines the notation of alternative alphabets and illustrates the modification of behavioral models to reflect newly discovered events. In [Sects. 81.3](#) and [81.4](#), we show how to modify the partial behavioral model of the properties, and partial behavioral model of scenarios to reflect newly discovered events, respectively. Conclusions are given in [Sect. 81.5](#).

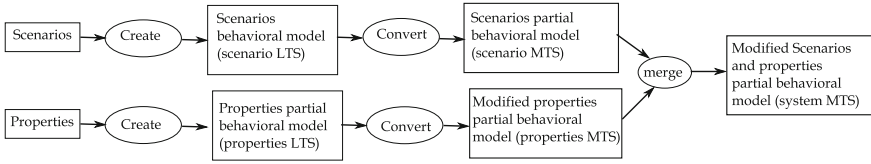


Fig. 81.3 The synthesis algorithm generates a partial behavioral model in three steps: creates behavioral model, converts it to partial behavioral model, and merges both the scenarios and properties partial behavioral models

81.2 Modifying Behavioral Models to Reflect Newly Discovered Events

In this section, we explain the meaning of two terms, *alphabet* and *alternative alphabets*. Next, we explain the effect of alternative alphabet in the synthesis algorithm. We also explain the meaning of *communicating alphabet* and *superset alphabet*. Finally, we show the different cases which arise when a new alphabet is discovered.

Definitions 1 and 2 give the definitions of alphabet and alternative alphabets, respectively.

Definition 1 (*Alphabet*) An alphabet is a set of actions that is used in both scenarios and properties to describe a system.

Definition 2 (*Alternative alphabets*) Let α_1 be the alphabet that was used during the creation of a partial behavioral model and let α_2 be the alphabet α_1 union the new actions discovered when a new scenario or property is added. We call α_1 and α_2 alternative alphabets.

Alternative alphabets arise when new actions in a new scenario (or a new property) are discovered and need to be included. Modifying the partial behavioral model, in the case of alternative alphabets, presents a problem,³ and in this case, there are two options for solutions. We could either (1) restart the synthesis algorithm from the beginning, creating a new model or (2) modify the existing model [7]. Alternative (2) avoids the exponential growth results of the computation time of synthesizing partial behavioral model from a larger number of scenarios and properties with a larger alphabet [14].

Each scenario or property exhibits a set of actions which may be described as a *communicating alphabet*. We will consider the set of actions recognized by all scenarios (and properties) and call it the *superset alphabet*. Definitions 3 and 4 conclude the definitions of the communicating alphabet and superset alphabet, respectively.

³ The problem is that the synthesis algorithm assumes the alphabet is fixed. However, in the case of alternative alphabet such characteristic does not hold.

Definition 3 (*Communicating alphabet*) A communicating alphabet, for a scenario (or a property), is the set of events that are exhibited by a scenario (or a property).

Definition 4 (*Superset alphabet*) A superset alphabet is the set of all events that are recognized by all scenarios and properties.

Similar scenarios (or properties) may be described by different alphabets originating from different stakeholders. This characteristic is because individual viewpoints and vocabularies can be expected to differ [7]. Usually a synthesis algorithm is used to build a partial behavioral model from a set of scenarios and a set of properties incrementally. As a consequence, it is common that new scenarios and new properties are added during the requirements engineering and model building stages [7]. Often, these new scenarios and new properties have new actions that were not considered in a previous application of the synthesis algorithm. Therefore, it is important to find a way to incorporate the newly discovered actions into the set of the known actions [7].

Any new event is either originated from a scenario or a property. Also, two types of behavioral models are generated for the system requirements. The first type is for the scenarios and the other type is for properties.⁴ Therefore, when a new event is added, the designer may be faced with the following semantics.

- (1) Modifying the partial behavioral model of scenarios to reflect newly discovered events in scenarios. New scenario is added to describe a new behavior.
- (2) Modifying the partial behavioral model of properties to reflect newly discovered events in scenarios. New scenario is added to describe a new behavior.
- (3) Modifying the partial behavioral model of scenarios to reflect newly discovered events in properties. New properties are needed to state the restrictions associated with scenarios.
- (4) Modifying the partial behavioral model properties to reflect newly discovered events in properties. Additional properties were identified.

Cases (2) and (4) are discussed in [Sect. 81.3](#). Cases (1) and (3) are discussed in [Sect. 81.4](#).

⁴ Each property has its own behavioral model, on the other hand each entity object in scenarios has its own behavioral model. Also, the behavioral model of all properties is the parallel composition of their behavioral models. Similarly, the behavioral model of all scenarios is the parallel composition of the behavioral models of the entities (objects).

81.3 Modifying the Partial Behavioral Model of the Properties to Reflect Newly Discovered Events

The modification that is needed to be applied to the partial behavioral model of properties when the new actions are originated from a new property is the same as the modification that is needed when the new actions are originated from a new scenario. Therefore, we will explain both Cases (2) and (4) in this section.

The current approaches, such as Uchitel et al.'s [7] and Krka et al.'s [6], that consider converting properties into behavioral models, do not consider two issues: the definition of the alphabet used in properties and the usage of the temporal logic used in FLTL formalism. We start by explaining the definition of alphabet in properties. Next, we compare two types of temporal logic used in FLTL formalism. After that, we explain the proposed modification for the behavioral model of properties. Finally, we provide an example.

81.3.1 The Definition of The Alphabet Used in Properties

The property's alphabet is the *actual* set of events used by the property [13] which we call the communicating alphabet (see Definition 3). This means it is possible (and more likely) that a property will have its own (different) communicating alphabet. Therefore, alternative alphabets for the set of properties exist in many cases.

Proposition 1 [13] *Let a_1 and a_2 be safety properties in FLTL. If a_1 and a_2 are closed under stuttering⁵ then for all traces $tr \models (a_1 \wedge a_2)$ if and only if tr is accepted by $\text{constraint}_{a_1} \parallel \text{constraint}_{a_2}$.*

In [13], the Proposition 1 is introduced to describe the relationships between the composition of properties and the composition of their behavioral models. Proposition 1 implies that the property's communicating alphabet includes only the events used by that property and cannot be the universe of all actions (the superset alphabet). If one supposes that the property's communicating alphabet is *the universe* of all events in all properties (the superset alphabet) then the use of Proposition 1 will not be applicable. As a consequence, the creation of behavioral model of multiple properties will not be possible.

⁵ This will be explain in the [Sect. 81.3.2.2](#)

81.3.2 Comparing Two Types of Temporal Logic

In this section, we describe two types of temporal logic used in FLTL formalism. We explain the *closed under stuttering* notation. We show which temporal logic is more appropriate to be used in the FLTL formalism to describe the system's properties.

81.3.2.1 Types of Temporal Logic

In FLTL formalism, two types of temporal logic can be used to describe the system requirements as properties: Lamport temporal logic [15–17] and the extended temporal logic [8]. The temporal logic described by Lamport [15–17] uses the boolean operators : and \wedge , or \vee , not \neg , and implication \rightarrow and the temporal operators: always \square , and eventually \diamond [17].

Extended temporal logic is defined using the standard boolean operators: and \wedge , or \vee , not \neg , and implication \rightarrow and the temporal operators: next X , strong until U , weak until W , eventually \diamond , and always \square [8].

The main difference between Lamport temporal logic and extended temporal logic is the usage of the X operator. The effect of the usage of this operator will be discussed later in Sect. 81.3.2.3.

81.3.2.2 Closed (Invariant) Under Stuttering Definition

Suppose, for instance, that we have the following two sequences:

$\sigma_1: S_0 \xrightarrow{\alpha_1} S_1 \xrightarrow{\alpha_2} S_2 \rightarrow \alpha_3 \dots$ and $\sigma_2: S_0 \xrightarrow{\alpha_1} S_1^1 \xrightarrow{\alpha_2^1} S_1^2 \xrightarrow{\alpha_2^2} S_1^3 \xrightarrow{\alpha_2^3} \dots \xrightarrow{\alpha_2^n} S_2 \rightarrow \dots$ where S_i represents a state, and α_i is an event that belongs to the property communicating alphabet. If the temporal logic can not distinguish between σ_1 and σ_2 , then the temporal logic is closed (invariant) under stuttering. Otherwise, it is not closed (variant) under stuttering [15]. In the other words, if σ_1 and σ_2 always have the same truth values, then the temporal logic is closed (invariant) under stuttering. On the other hand, if σ_1 and σ_2 may have different truth values, then the temporal logic is not closed (variant) under stuttering [15].

81.3.2.3 What Types of Temporal Logic Are More Appropriate to Use to Describe Properties (The Effect of Using the Next Operator)

The temporal logic operator *next*, X , is used to express that an event occurs in the next state. However, using the *next* operator, X , in the event-based temporal logic⁶ makes it not closed (variant) under stuttering [15].

⁶ The temporal logic that we mean here is synchronous temporal logic and not asynchronous temporal logic.

The extended temporal logic uses the *next* operator, X , which makes it not closed (variant) under stuttering. As a consequence, the creation of the compositional behavioral model for a set of properties is not possible. The reason for this is because Proposition 1 [13, 15] is not applicable. However, under some assumptions the above condition, *closed under stuttering*, can be satisfied such as in [17].⁷ As a consequence, the creation of the compositional behavioral model for a set of properties is possible because Proposition 1 is applicable [13, 15].

The use of the *next* operator X , makes the temporal logic more expressive, but this makes it not closed (variant) under stuttering. The ability provided by the addition of the *next* operator X can be substituted by the use of other operators [15]. This means the elimination of the *next* operator, X , is possible without much loss of expressiveness or preciseness of the temporal logic [15].

We propose the use of Lamport temporal logic [15–17]. Our choice is because this logic is closed under stuttering and therefore, allows us to use the parallel composition of the behavioral model of the individual properties to create the behavioral model of all properties.

81.3.2.4 The Proposed Algorithms

The creation of the behavioral model of the properties, represented as an LTS model, is affected by its communicating alphabet and the difference between the superset alphabet and its communicating alphabet. If we add an extra event, say, for instance, other-events,⁸ to the superset alphabet, then we can use it as a reference for any new alphabet discovered during the elaboration of the system requirements. Therefore, any addition of a new alphabet will not affect the created behavioral model of properties. Exception to this is when we refer to the difference between the superset alphabet and the property's communicating alphabet. In this case the extra event, other-events, can be used as a reference to update the behavioral model.

Our idea is to keep the behavioral model of properties. Then, we make the modification that we propose in Fig. 81.5. This can be applied for each individual behavioral model of a property. Such modification can also be done for the conjunction of behavioral models of properties represented as an LTS model. The reason for this is because building a behavioral model for the conjunction of properties is equivalent to the conjunction of the behavioral model of properties as Proposition 1 states [7]. Algorithm [7], shown in Fig. 81.4, creates a partial behavioral model for properties without considering alternative alphabets. To

⁷ This is done by assuming that asynchronous temporal logic is used. However, in event-based model synchronous temporal logic should be used. On the other hand, Lamport temporal logic does not use the *next* operator, X , which makes it closed (invariant) under stuttering.

⁸ This is similar to the term *anon* used in [8] where is used to represent other events (actions) that the system may use but is not used in defining properties.

Fig. 81.4 Creation of a behavioral model for properties without considering alternative alphabets

```

1: for all properties  $\varphi$  do
2:   Construct a Buchi automaton.
3:   Remove transitions that correspond to a finite trace.
4:   Remove all unreachable states.
5: end for

```

modify this algorithm to take into consideration the alternative alphabets, we propose the following: (1) redefine the alphabet to be the alphabet union a new term, *other-events*. (2) in the case of alternative alphabets, modify the behavioral model that resulted from the *creation* step (instead of re-applying the *creation* step) (see Fig. 81.3) by applying the step shown in Fig. 81.5 on the existing partial behavioral model of properties. This can also be done on the partial behavioral model that resulted from the *conversion step*.

81.3.2.5 Example

Suppose, for instance, a television system which blanks its screen while it tunes to a new channel [8]. The properties for such a system are described in Fig. 81.6.

In the following, we show the effect of adding a new alphabet to an already created behavioral model of properties. We modify the alphabet for the system in Fig. 81.6 and examine the resulting effects both with and without the use of the solution this paper proposes.

81.3.2.6 The Effect of Adding a New Alphabet on the Behavioral Model of the Properties Without Using the Proposed Algorithm

The alphabet of the properties in Fig. 81.6 is $\alpha_1 = \{\text{blank, unblank, endtune, tune}\}$. Figure 81.7a shows the behavioral model of the above properties. Now suppose, for instance, that new events (actions), *Press-sound-up-key*, *Press-sound-down-key*, are discovered in a new scenario or property. The new alphabet for this case is $\alpha_2 = \alpha_1 \cup \{\text{Press-sound-up-key, Press-sound-down-key}\}$ or $\alpha_2 = \{\text{blank, unblank, endtune, tune, Press-sound-up-key, Press-sound-down-key}\}$. As it can be seen the result of this operation created the alternative alphabet α_2 . As a consequence, the behavioral model must be re-created from scratch according to the new alphabet α_2 and the conversion algorithm needs to be reapplied also according this new alphabet. The new behavioral model obtained by using the new alphabet, α_2 , is shown in Fig. 81.7b.

Fig. 81.5 Modify the partial behavioral model of properties to incorporate the new actions that are discovered

- 1: Let new-actions be the new actions discovered in new properties or new scenarios
- 2: **for all** state S_1 such that $S_1 \xrightarrow{\text{other-events}} S_2$ **do**
- 3: Add new transition $S_1 \xrightarrow{\text{newactions}} S_2$
- 4: **end for**

Fig. 81.6 Properties that describes a television system which blanks its screen while it tunes to a new channel [8]

fluent BLANKED = < blank, unblank >
 fluent TUNING = < tune, endtune >
 $NOARTIFACTS = \square(TUNING \Rightarrow BLANKED)$

81.3.2.7 The Effect of Adding a New Alphabet on the Behavioral Model of the Properties Using the Proposed Algorithm

If we use the proposed algorithm, then $\alpha_1 = \{\text{blank, unblank, endtune, tune, other-events}\}$. The behavioral model of this property is shown in Fig. 81.8a. Now suppose, for instance, that new events (actions), *Press-sound-up-key*, *Press-sound-down-key*, are discovered in a new scenario or property. The new alphabet for this case is $\alpha_2 = \alpha_1 \cup \text{newEvent}$ or $\alpha_2 = \{\text{blank, unblank, endtune, tune, other-events, Press-sound-up-key, Press-sound-down-key}\}$. The advantage of the proposed algorithm is that we do not need to re-create the behavioral model. Instead, we can add the new events to each transition that has *other-events* as a label. This process will save the need of re-applying of the costly step for the creation of the behavioral model of the property. The modified behavioral model according to the new alphabet, α_2 , is shown in Fig. 81.8b.

If the event is originated from a scenario, then no other modifications are needed to the behavioral model of the properties. This is because the behavioral model of properties will not be affected of a new scenario. On the other hand, if the event is originated from a property, we need an extra step to generate a behavioral model for the new property and merge it with the existing model. This is because the behavioral model of properties need to include the new behavior introduced by the new property.

81.4 Modifying the Partial Behavioral Model of the Properties to Reflect Newly Discovered Events

The modification of the behavioral model when new events are originated from scenarios, Case (1),⁹ is different than the modification when new events are originated from properties, Case (3). Therefore, we will discuss each case

⁹ See Sect. 81.4 for both Cases (1) and (3)

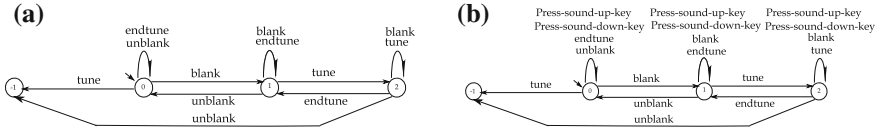


Fig. 81.7 The behavioral model of the set of properties without using the proposed algorithm. **a** Before the inclusion of the new alphabet; **b** after the inclusion of the new alphabet

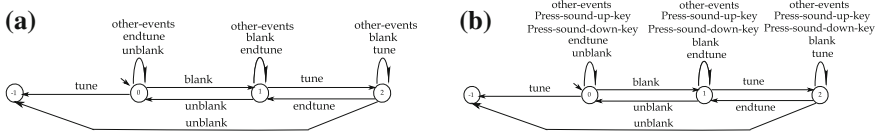


Fig. 81.8 Using the proposed algorithm, **a** shows the behavioral model of the set of properties before the inclusion of the new alphabet, and **b** shows the behavioral model of the set of properties after the inclusion of the new alphabet

separately. First, in [Sect. 81.4.1](#), we discuss Case (1). Then, in [Sect. 81.4.2](#), we discuss Case (3).

81.4.1 Modifying the Partial Behavioral Model of the Scenarios to Reflect Newly Discovered Events in Properties

To modify the partial behavioral model of scenarios to take into account the effect of new actions that are discovered in new properties, we propose the algorithm shown in [Fig. 81.9](#). In the algorithm shown in [Fig. 81.9](#), *sink* is a special state in the partial behavioral model of scenarios, and Y is an event that belongs to the superset alphabet. This means $Y \in \text{superset alphabet}$. Suppose, for instance, that we have the partial behavioral model shown in [Fig. 81.10a](#). Also, suppose that e is a new action discovered in a new property. Then, according to the algorithm shown in [Fig. 81.9](#), we modify the partial behavioral model shown in [Fig. 81.10a](#). This modification introduces the transitions: $S_1 \xrightarrow{e?} S_0$, $S_2 \xrightarrow{e?} S_0$, and $S_3 \xrightarrow{e?} S_0$, as shown in [Fig. 81.10b](#).

81.4.2 Modifying the Partial Behavioral Model of the Scenarios to Reflect Newly Discovered Events in Scenarios

Although it is possible that a new scenario is added after the creation of the partial behavioral model, the existing algorithms such as [\[10, 18, 19\]](#) do not support adding the effect of a new scenario (even if it has the same alphabet) to the partial

Fig. 81.9 Modify the partial behavioral model of scenarios to incorporate the new actions are discovered in properties

- 1: Let new-actions be the new actions discovered in the new properties
- 2: **for all** state s_1 such that there is a transition $S1 \xrightarrow{Y} sink$
do
- 3: Add new transition $S1 \xrightarrow{new-actions?} sink$
- 4: **end for**

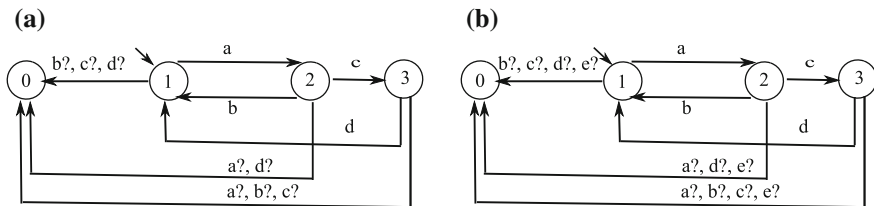


Fig. 81.10 A partial behavioral model for a set of scenarios before addition of new alphabet, and a partial behavioral model for a set of scenarios after addition of new alphabet “e”. **a**, and **b** illustrate the first and second model respectively

behavioral models. This situation arises because the creation of the behavioral model depends on the way the scenarios are related to each other.¹⁰ When adding a new scenario, it is difficult to determine how the new scenario is related to other scenarios. Even if it is known how the new scenario is related to the other scenarios, it is difficult to modify the partial behavioral model to reflect the effect of the new scenario. Therefore, a possible way to add a new scenario is to re-apply the steps from one to three (see Fig. 81.3).

81.5 Conclusion

The behavioral model of scenarios can be modified using the proposed algorithm if the actions are originated from properties instead of re-constructing it. Alternative alphabets cause the algorithms, that assume fixed alphabet to re-run again by re-generating the behavioral model of both scenarios and properties. Our proposed algorithm eliminates the need of the re-run of the creation step of the behavioral models of both scenarios and properties. Instead, our proposed algorithm modify the already created behavioral models. The elimination of the re-run of the creation step is important because this step is known to be a costly step.

¹⁰ Each scenario contains interactions for all entities in the scenario. On the other hand, the behavioral model is created for each entity.

References

1. Jacobson I, Booch G, Rumbaugh J (1999) The unified software development process. Addison-Wesley Longman Publishing, Boston
2. Weidenhaupt K, Pohl K, Jarke M, Haumer P (1998) Scenarios in system development: current practice. *Softw IEEE* 15(2):34–45
3. Weidenhaupt K, Pohl K, Jarke M, Haumer P (1998) Scenario usage in system development: a report on current practice. In: Proceedings of third international conference on requirements engineering 1998
4. van Lamsweerde A (2009) Requirements engineering: from system goals to UML models to software specifications. Wiley, New York
5. Haugen (ed) (1999) Recommendation Z.120: message sequence chart (MSC). International Telecommunication Union (ITU), Geneva
6. Krka I, Brun Y, Edwards G, Medvidovic N (2009) Synthesizing partial component-level behavior models from system specifications. In: ESEC/FSE' 09: proceedings of the 7th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on the foundations of software engineering. ACM, New York, pp 305–314
7. Uchitel S, Brunet G, Chechik M (2009) Synthesis of partial behavior models from properties and scenarios. *IEEE Trans Softw Eng* 35(3):384–406
8. Giannakopoulou D, Magee J (2003) Fluent model checking for eventbased systems. In: ESEC / SIGSOFT FSE. ACM, New York, pp 257–266
9. OMG (2005) Object constraint language specification, version 2.0. Object Modeling Group, June 2005. <http://www.fparreiras/papers/OCLSpec.pdf>
10. Uchitel S, Kramer J, Magee J (2003) Synthesis of behavioral models from scenarios. *IEEE Trans Softw Eng* 29(2):99–115
11. Haumer P, Pohl K, Weidenhaupt K (1998) Requirements elicitation and validation with real world scenes. *IEEE Trans Softw Eng* 24:1036–1054
12. Dzida W, Freitag R (1998) Making use of scenarios for validating analysis and design. *IEEE Trans Softw Eng* 24:1182–1196
13. Letier E, Kramer J, Magee J, Uchitel S (2008) Deriving event-based transition systems from goal-oriented requirements models. *Autom Softw Eng* 15(2):175–206
14. Bontemps Y, Heymans P, Schobbens PY (2005) From live sequence charts to state machines and back: a guided tour. *IEEE Trans Softw Eng* 31(12):999–1014
15. Lamport L (1983) What good is temporal logic? In: IFIP congress, pp 657–668
16. Lamport L (1983) Specifying concurrent program modules. *ACM Trans Program Lang Syst* 5:190–222. [10.1145/69624.357207](https://doi.org/10.1145/69624.357207)
17. Lamport L (1994) The temporal logic of actions. *ACM Trans Program Lang Syst* 16:872–923. [10.1145/177492.177726](https://doi.org/10.1145/177492.177726)
18. Whittle J, Schumann J (2000) Generating statechart designs from scenarios. In: ICSE '00: proceedings of the 22nd international conference on software engineering. ACM, New York, pp 314–323
19. Damas C, Lambeau B, Dupont P, van Lamsweerde A (2005) Generating annotated behavior models from end-user scenarios. *IEEE Trans Softw Eng* 31(12):1056–1073

Chapter 82

Watermark Singular-Values Encryption and Embedding in the Frequency Domain

Chady El Moucary and Bachar El Hassan

Abstract Protecting digital assets has become not only a taken-for-granted practice in most electronic applications for content and ownership authentication, but robustness and efficiency are becoming decisive attributes when exercising watermarking. A highly efficient approach is suggested and which integrates encryption to a hybrid watermarking algorithm based on the mathematical tools of the well-known singular vector decomposition (SVD) and the discrete wavelet transform (DWT). The innovative attribute of this approach stems from the fact that the suggested method achieves the encryption of the data prior to watermarking in the frequency domain. The encryption algorithm is based on a random key source which entitles the user to have a selected key-length and guarantee a significantly advanced level of intruders' unawareness. The encryption program is applied to the singular values of the watermark. The watermarking algorithm is based on the combination of the SVD and DWT. This watermarking method has high robustness against geometric transformations but not efficient for malign attacks. The combination of encryption and SVD-DWT algorithms overcomes this last weakness while keeping all advantages. As shown by the simulation results, a high level of robustness and security is attained in the sense that the concealed watermark is indestructible and almost undetectable in many cases. Moreover, the capacity of insertion is sustained at a satisfactory level while securing a high signal-to-noise ratio.

C. E. Moucary (✉)

North Lebanon Campus, Notre Dame University, Louaize, P.O. Box 87Tripoli
Municipality Street, Barsa, El Koura, Lebanon
e-mail: celmoucary@ndu.edu.lb

B. E. Hassan

Laboratory of Electronic Systems, Telecommunications, and Networking (LASTRE),
Lebanese University, Al Arz Street, Tripoli, El Kobbah, Lebanon
e-mail: bachar_elhassan@ul.edu.lb

82.1 Introduction

Due to the sharp evolution and expansion in Technology and electronic Business, intensive research has been carried out to outdo the drawbacks of such inevitable and imposing neo society. Indeed, piracy and illegal reproduction of digital assets have rendered both consumers and owners helpless in trusting various, sometimes incredulous, transactions that they have to deal with around the clock and across the Globe. Additionally, the development of monstrous processors and greedy data storage containers as well as the innovation of astoundingly sophisticated communication networks and software, have accentuated both the ability and repercussions of intrusions and hacking. Globalization and open-source assets only added to the problem and made it more urgent to develop a therapy. In fact, the dawn of Digital World has indubitably risen and highly efficient solutions have to be carried out to counter-balance the drawbacks of this era. The purpose of such solution would not be to abolish counterfeiting or reproduction, but to realistically hinder and/or deter illegal manipulations of digital resources.

Efficient solutions have been presented by researchers' endeavors based on Digital Watermarking [1]. This science/approach has attracted the attention of proprietors and has gained trust and reliability with the apparition of new competing techniques that would keep up with the rhythm of e-raids. In the same way that antivirus firmware are releasing new signatures and updates, the industry of watermarking shall need to maintain an adequately pace in supplying the e-community with overthrowing pertinent solutions to deter illegitimate operations.

Digital Watermarking techniques inundated the scientific community and many strategic approaches have been presented to satisfy various applications and requirements' specifications on many strata [18, 23]. The mathematical background has been abundantly elaborated and improved. Additionally, various aspects of the requirements have been highlighted and specific criteria have been laid out for evaluation of the performance of these techniques. Finally, emerging algorithms have overlapped with other security areas of expertise such as encryption and data hiding procedures to engender more potent watermarking tools.

Digital Watermarking is currently deployed in almost every application that needs to be shared over any communication channel or that presents any digital feature, thus embracing multimedia in its general gist, software industry, military, medical, forensics, national security and other "sensitive" applications under the umbrella of Data Hiding or Steganography, where incursion might have disastrous impacts [18, 26–29]. In fact, Digital Watermarking is not only used to identify ownership by insertion of a copyright segment, but has been and continues to be extensively used in data hiding for various purposes. Finally, Digital Watermarking is also employed in less critical applications from a security viewpoint such as land consolidation and, recently in some technical approaches with a "purely" scientific background [17, 30].

In this paper, we will present an innovative application of Digital Watermarking based on *robustly* encrypting the watermark prior to insertion. The watermarking

algorithm will be based on a hybrid method which makes use of two well-known algorithms, namely the singular value decomposition (SVD) and the DWT. The watermark will be added using various approaches for an improved level of security in the frame of specified applications. Particularly, multiple-insertion of the watermark will be studied and implemented on multiple levels of the still-image layers in the transform domain. Simulation results have been carried out to show and demonstrate the effectiveness of the algorithms elaborated in terms of diversified criteria such as robustness, transparency, capacity and peak signal-to-noise ratio (PSNR).

The paper will be organized as follows: in the subsequent section, Digital Watermarking will be recalled and SVD and DWT will be discussed in the context of the application. In Sect. 82.3, the aforementioned algorithms will be elaborated and simulation results will be demonstrated. Finally, some conclusions and perspectives will be drawn in Sect. 82.4 to underline the overall effectiveness of the suggested watermarking scheme.

82.2 Digital Watermarking Using a Hybrid Method Based on Combining the Features of SVD and DWT

82.2.1 Digital Watermarking of Still Images

Watermarking originated as the *art* of stamping a product with a *visible* logo to deter falsifications and, later on, for matching ownership of non-digital items. This routine has been exercised since ancient History, mainly for *concealed* communication of information and was rather labeled under “data hiding”. With the advent of technology, watermarking embraced invisible messages and thus, became a powerful technique for steganography. Nowadays, Watermarking, Steganography, and Data Hiding became interchangeably used. Indeed, they differ only in the purpose they serve but not in the *science* behind. Additionally, a discipline called cryptography has been scrounged to provide the entire structure with an efficient protection against information retrieval by intruders. Digital Watermarking can be classified from much diversified perspectives.

(1) Casting Method

Two broad categories can be underlined: the spatial domain where the cover image’s (host) pixels are directly modified to accommodate the message to be concealed. Data encryption and random distribution are needed here in order to deter intruders from reading the watermark. The second category of watermarking techniques is achieved in a transform domain where the coefficients’ magnitude is modified to reflect the hidden message. To name a few, discrete cosine transform (DCT), DWT, Fast Fourier Transform, Hadamard Transform, Haar Wavelet, Hough Codes, Neural Networks, singular value decomposition (SVD), Multiwavelet, and

other mathematical operations are examples of such types of watermarking [16, 19–22].

(2) *CompetenceCriteria*

The aforementioned categories essentially differ according to three major criteria: robustness, transparency and capacity of insertion. It is understood that these are somehow contradicting criteria since enhancing the watermark according to one of them, yields worsening the remaining ones. While capacity, sometimes referred to as data payload, refers to the size of the watermark that could be withheld by the host, transparency measures the visual impact that the watermark leaves on the host for watermarking is eventually an alteration of the cover image to a certain extent by modifying its luminosity, boundaries/edges, brightness/texture, layers, coefficients, etc.

Robustness against attacks reveals a critical criterion when referring to data hiding in general. Indeed, if an intruder is able to break the oyster and reveal and/or destroy the hidden watermark, the consequences are highly costly in most of the applications, namely in the ones requiring confidentiality or security. It is understood that these attacks might be benign or malign, thus watermarking should provide the capacity to resist geometric distortions in general and targeted raid aiming at intentionally harm the watermark as previously described.

(3) *“Blindness” Level*

Not all watermarking techniques can secure watermark retrieval with minimum *information*. Some techniques require possession of the original host image and/or the secret key by the intended receiver. This classification does not reveal of the performance of the watermarking technique since in some critical application, having a completely non-blind algorithm is of utmost importance for an enhanced level of security.

(4) *The use of a “Secret” Key*

The enormous innovation in the industry of computer science has led to the emergence of sophisticated and powerful software which makes it feasible for expert hackers to locate the watermark and eventually decode or alter it. The use of a key refers to the encryption phase and is implemented to augment the level of robustness against attacks since an encrypted watermark would present an additional front that the intruder has to overcome. There exist two different configurations: the symmetric and asymmetric configurations. In the first one, the same key is used for both encryption and decryption. The main drawback of this configuration is that the number of keys increases as the square of the number of network members. This quandary was solved by W. Diffie and M. Hellman who invented the notion of public key, which is referred to as the asymmetric key cryptography [24]. In this context, two keys are used, one private and the other one is public; they are interrelated but unachievable one from the other. Furthermore, C. Shannon and W. Weaver established that it is possible to create a perfectly indissoluble key if three conditions are met [25]. First, the key should emanate

from a random source; second, the key must not be reused; and finally, the key length should be at least equal to the one of the concealed data.

In this paper, the approach for key generation is fundamentally different; it relies on the idea that how matter astutely a key is generated by an algorithm, some professional hackers would be able to decode it using sophisticated software in a reasonable amount of time; otherwise, the encryption phase might suffer from a lack of new keys after many applications. In this paper we suggest an innovative approach for adopting a key that has no logical bonds amongst its bits and with a random length that can check with the watermark's length.

82.2.2 Singular Value Decomposition

For a real matrix A of size (l, c) , the Singular Value Decomposition, also known as SVD, is a linear-algebra factorization, which finds many applications in signal processing. Applying SVD to A allows decomposing it into a product of three matrices U , V , and S where U and V are orthogonal matrices with $UU^T = I$, $VV^T = I$, and S is a diagonal matrix (I is the identity matrix). The diagonal entries of S are called the singular values of A , whereas the columns of U and the columns of V are called the left and right singular vectors of A , respectively. U is of size (l, l) whereas the size of V is (m, m) . The left and right singular values of A also represent the eigenvectors of AA^T and A^TA , respectively.

The decomposition of a matrix A of rank r can be expressed using (82.1):

$$A = U_1 S_1 V_1^T + U_2 S_2 V_2^T + \dots + U_T S_T V_T^T \quad (82.1)$$

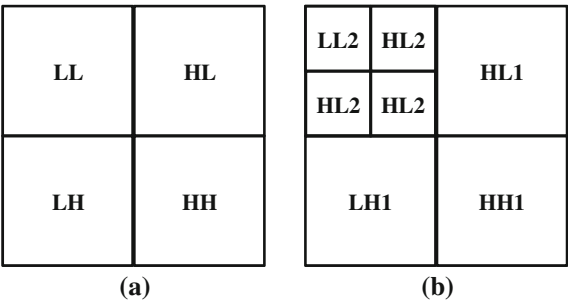
When A represents a still image, the singular values characterize the luminance while the corresponding pair of singular vectors specifies its geometry.

Various watermarking techniques are based on SVD; data is embedded via the modification of the singular values but they result in degradation in the quality of the retrieved watermark. Furthermore, this approach usually presents drawbacks related to geometrical and morphological attacks. In this paper we will propose applying SVD watermarking after transferring the host image into the DWT domain to enhance both robustness and the quality of the retrieved message at the receiver end as shown in Sect. 82.3. This combination will allow for a tool that would compensate for the drawbacks of each domain and thus, realize a beneficial blend [7, 9, 10, 12].

82.2.3 Discrete Wavelet Transform

There exists a wide variety of frequency-domain transforms used in watermarking such as the DCT, discrete fourier transform (DFT), and Hadamard Transform [6, 8, 13–15]. Nonetheless, the DWT represents superiority when it comes to spatial

Fig. 82.1 One-level and two-level 2D DWT of a still image



localization and multi-resolution characteristics [Ali Al Haj] which are similar to the Human Visual System (HVS). Furthermore, DWT is also one of the most useful tools in image compression such as JPEG2000.

The application of a two-dimensional DWT to an image results in the decomposition of the image into 4 sub-bands as shown in the Fig. 82.1 below:

- LL—Lower resolution version of the image;
- LH—Horizontal edge;
- HL—Vertical edge;
- HH—Diagonal edge.

Watermarking is performed via additive modification of the coefficients of the sub-bands selected to be higher than a threshold. The threshold is determined based on the correlation between the coefficients of the watermarked image and the watermark. The appropriate watermark would be the one that surpasses the threshold.

Most watermarking techniques based on DWT perform the insertion of data in the three latter sub-bands. In fact, altering the coefficients in the low-frequency LL sub-band results in a significant impact on the HVS of the image and defies the purpose of transparency for most of the energy of the image is present in this sub-band.

Furthermore, it is understood that embedding data in the edge layers of the image would result in the least visually perceptible changes in the cover watermarked image. Consequently, images with more frequent edges will allow for an increased capacity of insertion. DWT watermarking schemes are known for their robustness against many benign attacks, namely geometrical ones such as JPEG compression, Gaussian noise, and filtering. Additionally, embedding solely in the mid- and high- frequencies of the image has a relatively negative impact on the robustness level against low-pass filtering, and rotation. As we will show in the subsequent sections, using a hybrid method based on the combination of SVD and DWT allows embedding in all the frequencies of the image and achieves a twofold objective: lessen the drawbacks of each of these watermarking schemes and emphasize the robustness of each, thus yielding a more comprehensively better performance.

82.3 Hybrid Scheme: Combining SVD and DWT with Encryption for a More Robust Watermark

82.3.1 Secret Key for a Very Robust Encryption

The secret key used in decryption is of vital importance and plays a key role in protecting the hidden watermark if the intruder was able to locate it. It is the oyster that embraces the information that should be kept inaccessible to hackers. As it has been previously stated, if the key is used only once and has a length equal to the one of the watermark, it can reveal indestructible in a computationally reasonable time [24, 25]. However, with the growth of software industry yielded potent deciphering tools.

In this frame, we will employ a sort of randomly generated keys that do not exhibit any kind of mathematical or logical interrelation. Indeed, the key adopted is not generated by any function or algorithm that uses mathematical processing. Rather, the key for ciphering/deciphering will be picked almost randomly and can stem from any *source* that the user would select. It is understood that one constraint is to be respected, that is the key should be processed by a computer. Therefore, one could choose a song, a piece of image or a text as a source of cyphering bits.

This endows our scheme with superior maneuvering margins and entitles the watermark with a challenging caliber of security while being a user-friendly interface.

82.3.2 Encrypted Watermarking Algorithm

The encrypted watermarking scheme can be described by the block diagram of Fig. 82.2 below. It is based on the well-known SVD-DWT hybrid algorithm [2–5, 11].

In more details, the procedure of embedding the watermark can be summarized by the steps below:

- Apply DWT to the cover/host image. This can be achieved either using one-level or two-level approaches.
- Apply SVD to the watermark images. Two cases can be considered. If the watermark is of utmost importance and need be recovered at any cost, the same watermark can be embedded in all frequencies, thus guaranteeing its existence. Otherwise, if security is more important, we can use multiple (up to four) watermarks to distract the intruder from the wanted one.
- Encrypt the singular values of each watermark using the generated secret key. This key is obtained as discussed in the previous section.

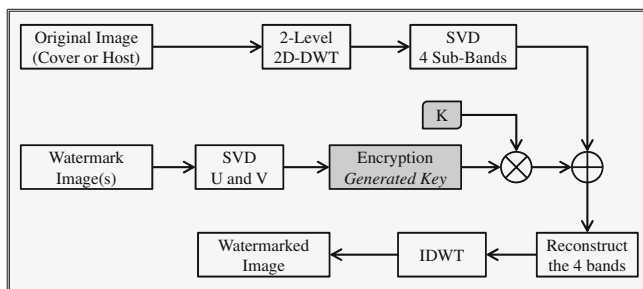


Fig. 82.2 Hybrid and encrypted watermarking using combined SVD and DWT

- Apply SVD to each sub-band on a selected level of the DWT block; all frequencies will be involved.
- Modify the diagonal of each band in an additive manner with the coefficients of the watermark singular values multiplied by a constant K .
- Reconstruct the four bands using the left and right values with the modified diagonal relevant to each band.
- Calculate the Inverse-DWT to obtain the watermarked image.

It should be noted that for watermark retrieval, the inverse path is straightforwardly followed.

82.3.3 Test Results

As an example of application of this algorithm, we will insert four watermark images in all frequencies of a host image. Since attacks usually are of a certain frequency range, one could retrieve the watermark even if the image is altered. Furthermore, if the intruder attempts at abolishing the watermark, the cover image should be almost completely destroyed in order to remove the watermarking from the entire image's frequencies. Finally, if the intruder attempts retrieving the watermark, the encryption stage would significantly resist reading correctly the hidden message since the key is used in a way to verify the indestructibility condition.

It should be noted that at least one watermark could be retrieved under severe "raids" and which is the one hidden in the low-frequency sub-band for the values of its coefficients vary slightly.

Our algorithm is mainly based on a non-blind watermarking scheme and requires that the watermarks have half size, both in rows and columns, compared to the cover image for proper watermarking in the four sub-bands.

Table 82.1 below summarizes the result for different cases of watermarking using the suggested algorithm.

Table 82.1 PSNR for different images and values of K

| Cover images | PSNR | PSNR | Size |
|--------------|--|--|---------|
| | K = 0.085 for LL K = 0.02 otherwise | K = 0.05 for LL K = 0.005 otherwise | |
| Rice | 25.5082 | 30.8407 | 256*256 |
| Cameraman | 26.5953 | 31.9173 | 256*256 |
| Lifting body | 27.3736 | 32.6001 | 512*512 |
| Tire | 22.6749 | 27.9365 | 232*205 |
| Pout | 25.0787 | 30.3642 | 240*291 |
| Circuit | 23.0139 | 28.3510 | 272*280 |

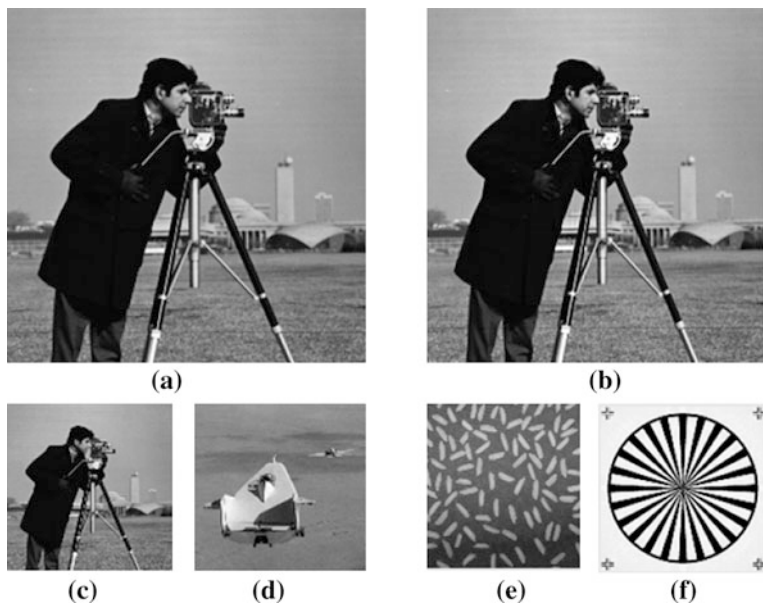


Fig. 82.3 Watermarking four different images in a cover image (**a** cover image, **b** watermarked images, **c–f** extracted watermarks from LL, HL, LH, and HH, respectively)

Figure 82.3 below demonstrates simulation results when four different watermarks were embedded in all frequencies of a host image. The PSNR obtained is approximately 26.5.

82.4 Conclusion and Perspectives

In this paper we referred to the hybrid watermarking algorithm based on SVD and DWT to achieve watermarking in all the frequencies of the cover image. The innovative part of this part relates to applying encryption of the watermark prior to

insertion. This encryption is performed to the singular values of the watermarks and then the ones of the cover image transformed in the DWT domain are “modulated” using an additive manner.

The proposed scheme achieves the well-known attributes of such hybrid combination of the SVD and DWT, i.e. having a watermark that is robust to a wide range of benign attacks, namely the geometrical ones. DWT is recalled to be the basis of the JPEG2000 compression algorithm.

The use of the key was presented in accordance with the criteria for an indestructible one. Rather than generating the key from a computer program, which can be deciphered using sophisticated software, the key was extracted from random sources selected by the user. This approach leaves the scheme of security extremely versatile and endows it with a very high level of robustness against malign attacks.

The simulation results demonstrate the effectiveness of the suggested scheme in terms of an acceptable PSNR, HVS transparency, and recovered watermarks. Furthermore, the capacity of insertion is beyond satisfactory as stated by the aforementioned constraints. Finally, this paper has utilized the benefits of the SVD-DWT in terms of robustness against geometrical attacks and exalts the confidence and privacy of the hidden data to a higher rank worth of “security-sensitive” applications. In fact, both attempts from an intruder to either abolish or decipher the watermark will need extremely demanding computational time and modus operandi.

References

1. Gonzalez RC, Woods RE (2008) Digital image processing, 3rd ed., Pearson Education, Singapore
2. Prasad MR, Koliwad S (2009) A comprehensive survey of contemporary researches in watermarking for copyright protection of digital images. *Int J Comput Sci Netw Secur* 9(4):91–107
3. Chen B, Wornell G (2001) Quantization index modulation: a class of probably good methods for digital watermarking and information embedding. *IEEE Trans Inf Theory* 47(4):1423–1443
4. Judge JC (2003) Steganography: past, present, future. SANS Institute, Information Security Reading Room, Genoa
5. Kessler GC (2004) An overview of steganography for the computer forensics examiner. *Forensics Sci Commun* 6(3):1–29
6. Hosmer C, Hyde C (2003) Discovering covert digital evidence. In: Digital Forensic Research Workshop (DFRWS), August 2003
7. Fang W-P (2006) Combining copyright protection and data hiding—a sensitive transformation approach. In: Proceedings of the 6th WSEAS international conference on signal processing, computational geometry & artificial vision, Greece, August 21–23, pp 45–49
8. Li L, Zhang C, Liang M, Li D (2009) Data Protection for Land Consolidation with Distortion Tolerable LSB Watermarking. *WSEAS Trans Inf Sci Appl* 6(3):427–436

9. Kumar S, Balasubramanian R (2009) An optimally robust digital watermarking algorithm for stereo image coding. In: Grgic M et al. (eds) *Studies in computational intelligence*, Rec. Advan In Mult Sig Process and Commun, SCI 231, pp 467–493
10. Por LY, Lai WK, Alireza Z, Ang TF, Su MT, Delina B (2008) StegCure: a comprehensive steganographic tool using enhanced LSB scheme. *WSEAS Trans Comput* 7(8):1309–1318
11. Saryazdi S, Nezamabadi-pour H (2005) A blind digital watermark in hadamard domain. *World Acad Sci Eng Technol* 3:126–129
12. Wang Y, Pearmain A (2004) AC estimation-based image watermarking method. In: *Proceedings of WIAMIS*, April 2004, pp 21–23
13. Veeraswamy K, Srinivas Kumar S (2008) Adaptive AC-coefficient prediction for image compression and blind watermarking. *J Multimed* 3(1):16–22
14. Cox IJ, Kilian J, Leighton T, Shamoon T (1997) Secure spread spectrum watermarking for multimedia. *IEEE Trans Image Process* 6(12):1673–1678
15. Diffie W, Hellman M (1976) Multi-user cryptographic techniques. In: *AFIPS Proceedings*, vol 45, June 1976, pp 109–112
16. Shannon C, Weaver W (1963) *The mathematical theory of communication*, University of Illinois Press, New York
17. Zhou B, Chen J (2004) A geometric distortion resilient image watermarking algorithm based on SVD. *Chin J Image Graph* 9:506–512
18. Andrews HC, Patterson CL (1976) Singular value decomposition (SVD) image coding. *IEEE Trans Commun* 24(4):425–432
19. Liu R, Tan T (2002) A SVD-based watermarking scheme for protecting rightful ownership. *IEEE Trans Multimed* 4(1):121–128
20. Ganic E, Eskicioglu AM (2004) Secure DWT-SVD domain image watermarking: embedding data in all frequencies. In: *ACM Multimedia and Security Workshop*, Magdeburg, Germany, September 20–21, pp 166–174
21. Chan PW, Lyu MR (2003) A DWT-based digital video watermarking scheme with error correcting code. In: *Proceedings fifth international conference on information and communications security*, pp 202–213
22. Mehul R, Priti R (2003) Discrete wavelet transform based multiple watermarking scheme. In: *Proceedings of IEEE region 10 technical conference on convergent technologies for the asia-pacific*, India, October 14–17
23. Mallat S (1989) A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Pattern Anal Mach Intell* 11(7):674–693
24. Shikha T, Nishanth R, Bernito a, Neeraj KJ (2010) A DWT based dual image watermarking technique for authenticity and watermark protection. *Signal Image Process Int J (SIPIJ)* 1(2):33–45
25. Daren H, Jiufen L, Jiwu H, Hongmei L (2001) A DWT-based image watermarking algorithm. In: *IEEE international conference on multimedia and expo (ICME'01)*, Japan, August 2001, pp 80–86
26. Al-Haj A (2007) Combined DWT-DCT digital image watermarking. *J Comput Sci* 3(9):740–746
27. Ganic E, Eskicioglu AM (2005) Robust embedding of visual watermarks using DWT-SVD. *J Electron Imaging* 14(4):1–13
28. Ganic E, Zubair N, Eskicioglu AM (2003) An optimal watermarking scheme based on singular value decomposition. In: *Proceedings of the IASTED international conference on communication, network, and information security (CNIS)*, Uniondale, pp 85–90
29. Mansouri A, Aznavah AM, Azar FT (2009) SVD-based digital image watermarking using complex wavelet transform. *Sadhana* 34(3):393–406
30. Mathivadhani D, Meena C (2011) Performance evaluation of key for watermarking using 2-D wavelet transformations. *Int J Eng Sci Tech (IJEST)* 3(3):1959–1966

Chapter 83

Business Intelligence Made Simple

Vasso Stylianou, Andreas Savva and Spyros Spyrou

Abstract Business Intelligence (BI) refers to a very broad category of applications which assist executive business users to improve their decision making and strategic planning by collecting, storing and analysing a usually large volume of historically collected data and providing access to powerful and dynamic query results. Business Intelligence Systems (BIS) are also sometimes referred to as Decision Support Systems due to the fact that they provide support to users and organisations concerning their decision making. Without a doubt, BI is extremely vital for all organisations in today's competitive business environment where even the smallest detail could affect the future of an organization. The study of BI is of interest to Computer Science (CS) and possibly the related Computer Engineering majors as well as Management Information Systems (MIS) and in broad to Business Studies majors. A course on BI is not usually a major requirement in any one of these curriculums, and it may not even appear as a major elective but rather be included in some other related course. For example, in the CS undergraduate university curriculum the topic of Business Intelligence may be touched upon briefly in a course on Database Management and in the graduate CS curriculum it usually becomes part of a Knowledge Management course. In the undergraduate (MIS) curriculum some additional mentioning is usually made in an Information Systems course. The brief coverage of the topic and the absence of simple and inexpensive educational tools that could easily explore the role of BI to students do not allow the educator to stress the magnitude of BI. This paper addresses the need for an educational case

V. Stylianou (✉) · A. Savva · S. Spyrou
University of Nicosia, 46 Makedonitissas Avenue, P.O. Box 24005, 1700 Nicosia, Cyprus
e-mail: stylianou.v@unic.ac.cy

A. Savva
e-mail: savva.a@unic.ac.cy

S. Spyrou
e-mail: Spyrou.S4@student.unic.ac.cy

which will demonstrate to the students the process used to create a BI application and the subsequent use of the BI tool in a simple but realistic way.

83.1 Introduction

83.1.1 *What is Business Intelligence?*

Business Intelligence refers to a very broad category of applications and technologies which serve the purpose of collecting, storing, analysing and providing access to data with the purpose of helping business users to make better business decisions [1].

In modern businesses, the emergence of standards in computing and technologies has meant that businesses now possess vast amounts of electronic data. In addition, businesses require that this data provides, with further analysis, even more information concerning the environment in which they operate. In this process of analysing data there are various technologies used and the information obtained from such an analysis will serve the business in its decision making and its future strategies [2].

BI can also be described as the business process which involves a series of activities used to gather and analyze data and distribute the results to those requiring them in order that decision-making is improved [3].

83.1.2 *The Role of Business Intelligence in an Organisation*

BI is extremely vital to organisations and their employees as they need access to timely information and they also have the need for analysis of that information [4].

In order to pursue more strategic BI, organisations must be able to effectively manage all the data at their disposal but also all the information that is retrieved must be of very high quality standards. If the information retrieved from the data is not of high quality standards then this effectively means that organisations will be limited to a less strategic BI approach. To be successful in its data analysis a business must possess fast, relatively cheap data warehouses required to effectively support BI [4]. A data warehouse refers to a database system that includes data, programs, and the necessary personnel who specialise in the preparation of data for BI processing.

Figure 83.1 below illustrates the components of a data warehouse. Data are read from operational databases by the Extract, Transform and Load System which is referred to as ETL in the figure. This system will then clean and prepare the data in order to be processed for BI. The extracted data will be stored in a data warehouse using a data warehouse Database Management System (DBMS). Additionally,

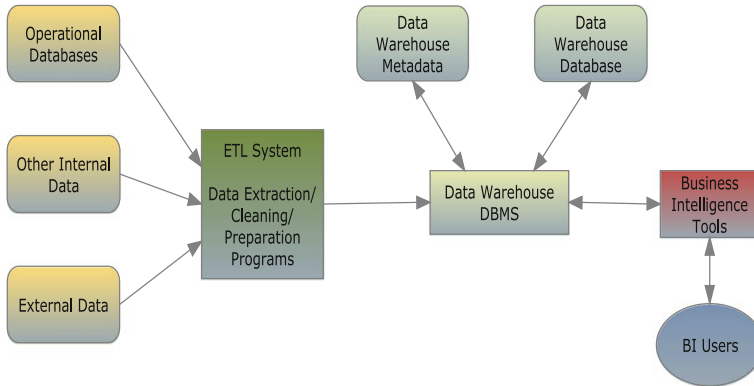


Fig. 83.1 Data warehouse components (Revised—Kroenke and Auer [4])

metadata concerning the data's source, format assumptions, etc., is maintained. The data warehouse DBMS will then extract data and forward them to BI Tools with which BI users interact [5].

BI systems are also referred to as Decision Support Systems since they assist organisations with their decision making [5]. They are used in analysing present and past activities of the organisation but also in accurately predicting future events. BI systems do not support operational activities such as processing or recording of orders. Instead, they support managers in their planning, analysis, control, and decision-making tasks.

Figure 83.2 below shows the role of BIS in decision making. Business Intelligence Systems provide the tools for the transformation of data into information and knowledge which in turn assists the organisation in its decision making. If the decisions undertaken by the organisation are successful, it is expected that the organisation will witness an important improvement in its competitiveness [6].

Business Intelligence Systems should make it possible for the users to set precise objectives and to execute these objectives. Furthermore, they provide a basis for decision making and allow the optimisation of future actions by acting upon various aspects of the company's performance. Ultimately, they help enterprises to realise their strategic objectives more effectively [6].

83.1.3 Categories of Business Intelligence Systems

Business Intelligence Systems fall under two, very broad categories. These categories are Reporting Systems and Data Mining Applications [5].

- *Reporting Systems*

Reporting Systems group, filter and sort data, while they also perform simple data calculations. Any reporting analyses can and are usually performed using

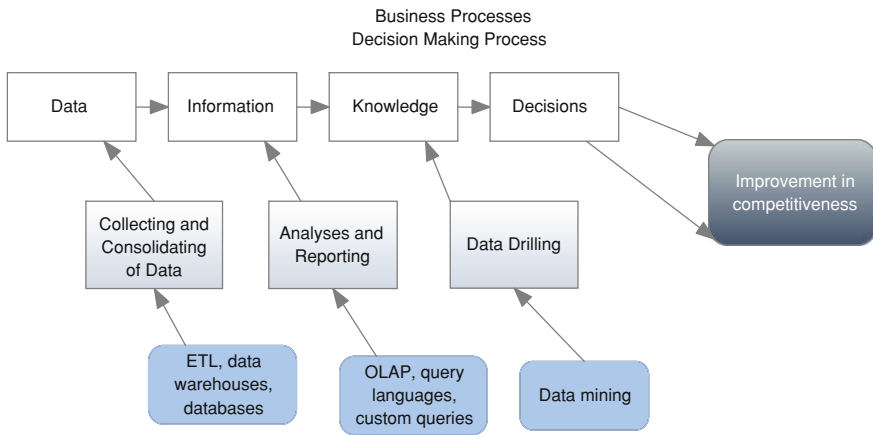


Fig. 83.2 The role of Business Intelligence Systems in decision making (Revised from Olszak and Ziemia [6])

standard Structured Query Language (SQL). Such systems offer the capability of collecting and summarising the current activities of the organisation, and allow the comparison between current and past activities when required. They may also be assisting in predicting any future activities of the organisation. The timely delivery of reports to the appropriate users is vital and the presentation of data in the appropriate format(s) is necessary [5].

- *Data Mining Systems*

Data Mining Applications use mathematical and statistical techniques to perform what-if analysis, make future predictions, and also assist in decision making [5]. Usually they are operated by relatively few people within an organisation and these are people who possess very sophisticated computer skills. Data Mining Applications are designed with the purpose of analysing large amounts of data and search for specific patterns or relationships if they exist, but are also used to identify unusual patterns as well. An unusual pattern for example could relate to credit card usage. If the card owner does not normally use his/her credit card much, a very rapid increase in usage could possibly mean that the card was stolen. Such an application could very likely help detect credit card fraud by monitoring credit card usage for each owner [7].

83.1.4 Business Intelligence Tools

A number of software tools are available for BI analysis which in the hands of the knowledge user can become very valuable for running a business successfully. This knowledge user can be defined as the business executive who has good computer skills or the computer professional who is commissioned to develop a BI

application or employed as a business analyst to perform the task of BI within the organisation. Examples of these tools are:

- SPSS [8]—A powerful general-purpose statistical package offering advanced options for numerical analysis and diverse modes of presentation via graphical and other means.
- Excel—A simplified commercial spreadsheet application software with built-in functionality for numerical analysis and presentation and a very easy-to-use GUI.
- TARGIT [9]—A Business Intelligence Suite offering a wide range of tools for data analyses and in-depth reports and to create dashboards, perform what-if analyses, simulation and more.

Below is a list of a few more proprietary BI tools [10]:

- Microstrategy
- IBM
 - Applix
 - Cognos
- InetSoft
- Informatica
- Information builders
- Microsoft
 - SQL Server Reporting Services
 - SQL Server Analysis Services
 - PerformancePoint Server 2007
- Proclarity
- Oracle corporation
 - Hyperion solutions Corp.
 - Oracle business
- Intelligence
 - Suite enterprise edition
- SAP Business Information warehouse
 - Business Objects
 - OutlookSoft
- Sybase IQ

Some open source BI tools are also available for example BIRT, Jasper, etc.

83.2 Aims and Objectives of the Study

The study of BI is of interest to Computer Science (CS) and possibly Computer Engineering majors as well as Management Information Systems (MIS) and in broad to Business Studies majors. A course on BI is not usually a major requirement in any one of these curriculums, and it may not even appear as a major elective but rather be included in some other related course. For example, in the CS undergraduate university curriculum the topic of Business Intelligence may be touched upon briefly in a course on Database Management and in the graduate CS curriculum it usually becomes part of a Knowledge Management course. In the undergraduate (MIS) curriculum some additional mentioning is usually made in an Information Systems course. The brief coverage of the topic and the absence of simple and inexpensive educational tools that could easily explore the role of BI to students do not allow the educator to stress the magnitude of BI.

This paper addresses the need for an educational case which will demonstrate to the students the process used to create a BI Reporting System and its subsequent use as a powerful Decision Support System. The specific objectives of the study include:

- Select an appropriate development methodology for the BI application.
- Explain the development of the BI application by following the activities of the development methodology.
- Select and explain BI suitable development/implementation tools.
- Explain the design of BI Dimensional Database.
- Present and explain the BI Graphical User Interface.
- Provide examples to demonstrate the potentials of a BI Reporting System to support the decision maker.

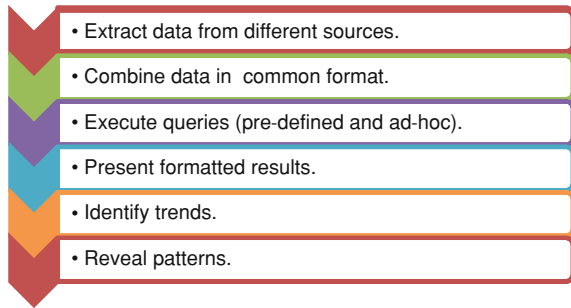
83.3 A Business Intelligence Case Study

(The following case study is based on a hypothetical scenario).

ABC Inc. is a business which trades sporting goods. The business maintains a number of retail stores. The overall sales of the organisation have been in decline in the last few quarter years which has resulted in top-level management being concerned about this situation.

Each retail store that the company owns has its own, individual databases which keep, among other, sales data for all of its retail products. At the top-level of management only summarised sales figures are reported. Management identified a need for more advanced and detailed reports in order to have a better understanding of the company's performance and its customers' behaviour. Examples of such detailed reports are Sales by Product Item, such as specifically Brand A and

Fig. 83.3 BI reporting system functionality



Model B of running shoes, or Sales by Product Category, such as running shoes. These reports will be used by management to make more accurate decisions on which product lines to stop supporting in order to reduce costs of maintaining them in inventory.

Business Intelligence Reporting System (BIRS)

A development team consisting of BI specialists has been commissioned to develop a BIRS by considering the needs of different user groups within the organisation. This BIRS will be capable of (Fig. 83.3):

1. Extracting all the required data from any database necessary.
2. Combining these data in a common format.
3. Executing multiple queries based on management requirements.
4. Presenting the results in a useful format output.
5. Identifying any underlying trends regarding sales or costs or any other aspect of the organisation and its operations (optional).
6. Revealing any existing customer buying patterns (optional).

Management will be able to use the reports of the BI system to take improved decisions. One example of a buying pattern would be that customers tend to buy more products in the first few days of each new month, or they tend to buy more at the end of each week. In such a case, management might decide to offer better deals at the start of each month. Thus, taking advantage of the knowledge it now possesses concerning customer behaviour it takes action in order to increase the revenues of the business.

83.3.1 Development Methodology Used

The methodology selected for the development of the BI application takes into consideration the role of the end-users and begins by drafting out organisation and user requirements for the successful implementation of the system.

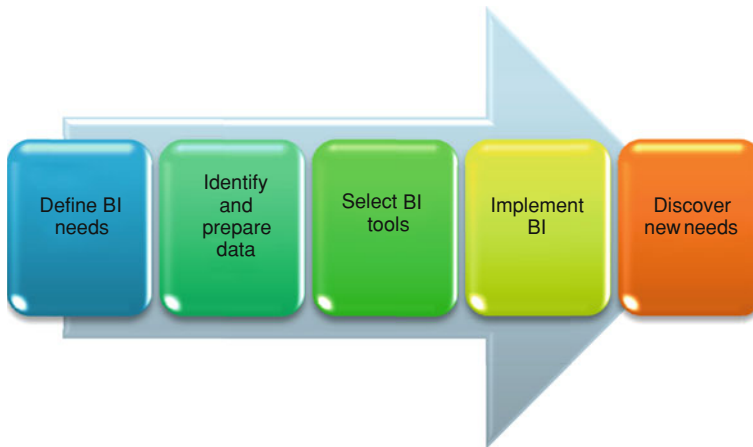


Fig. 83.4 BI development process

Organisation requirements include that the organisation should have a certain culture and certain experience of working in the past with information technologies [6]. These and some other requirements are also imposed on the end users.

The selected development methodology follows this sequence of activities [6] (Fig. 83.4):

- Defining BI undertaking. This step involves specifying the informational needs of the organisation, the desired solution and the requirements for its development.
- Identifying and preparing source data.
- Selecting BI tools.
- Implementing BI.
- Discovering and exploring new informational needs.

83.3.2 BI Development Tools

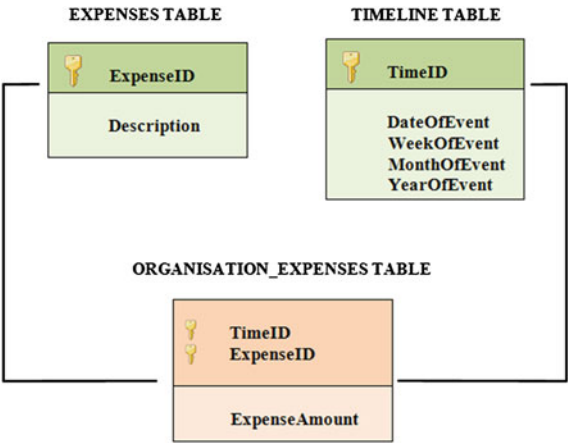
Visual Basic (VB) was used to design the GUI of the BI application and to implement the various functionalities and features required by the organisation.

A Microsoft Access 2007 [11] database was chosen to store the data to be used by the BIS. Microsoft Access 2007 offered an inexpensive and simple solution as the data to be stored was not of a very complex nature.

The selected BI tool was Microsoft Excel 2007, which although not purely a BI tool, offered sufficient functionalities for the case organisation.

Through the VB GUI the user may select from various pre-defined or custom-built queries. In the execution of these queries, data will be selected from the dimensional Access database using SQL commands. The results will be displayed in an Excel spreadsheet.

Fig. 83.5 A part of the dimensional database



83.3.3 BI Dimensional Database

The Dimensional Database was deduced from the Operational Database used by the organisation to record all its daily activities. A Dimensional Database differs from an Operational Database in that it is designed for efficient data analysis and for performing queries.

Within the Dimensional Database, the various tables were arranged using a Star Schema. In a Star Schema, Fact Tables are connected to Dimension Tables. Examples of the Dimension tables created are an EXPENSES and a TIMELINE table which are connected to a Fact table called ORGANISATION_EXPENSES using a Star Schema as shown in Fig. 83.5 above.

83.3.4 BI Graphical User Interface

Access to the BI application is restricted to valid users granted a username and password. The main screen of the application is as shown in Fig. 83.6 below.

By selecting the Business Intelligence Tools option the user can perform a number of queries mainly grouped in two categories. The Automated Queries which are pre-defined queries available without any additional input and the Custom Queries which are built based on the criteria specified by the user. Fig. 83.7 shows the output of an Automated Query of “Employee Sales for the year 2009” and Fig. 83.8 which follows shows the screen from which the user inputs the criteria for a specific Custom Query.

The example Custom Query of Fig. 83.8 is one in which the user has requested to find specific employee sales in a selected time period. The selection included Employee = “Emp003” and sales made between 22/04/2009 14:32:25 and 01/01/

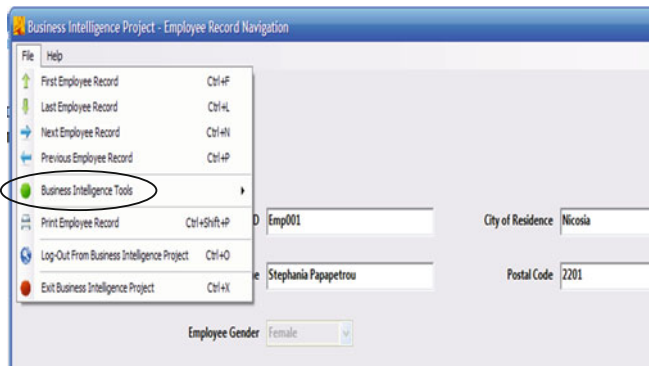


Fig. 83.6 The Initial screen of the BI Application

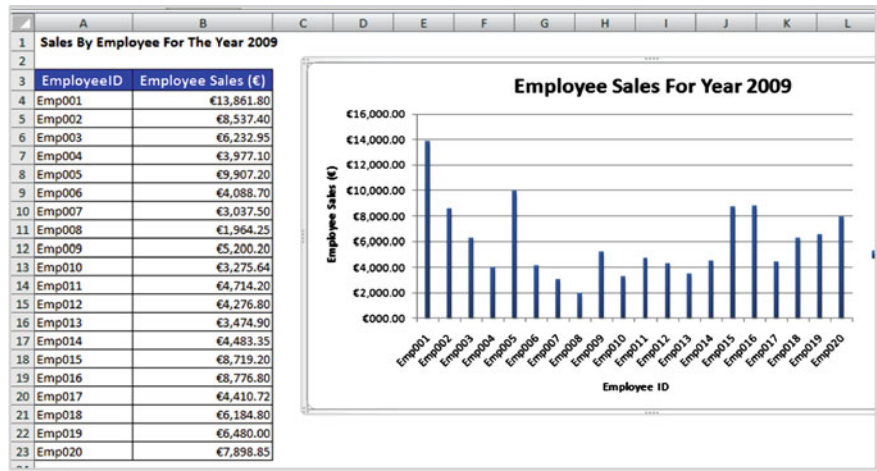


Fig. 83.7 “Employee Sales for the year 2009”; an automated query

2010 11:27:35. The results of this Custom Query shown both in a numerical and graphical format are illustrated by Fig. 83.9.

The criteria available for Custom Queries, as shown in Fig. 83.8, are:

- Time frame of interest, specified as a From–To date, or in week(s) (From–To), or month(s) (From–To) or year.
- Employee ID.
- Store or stores.
- City or cities.

The above criteria can be combined in many ways in order that company sales are analyzed as best as possible. Such sales analysis constitutes valuable Business

Fig. 83.8 The custom queries input form of the BI application

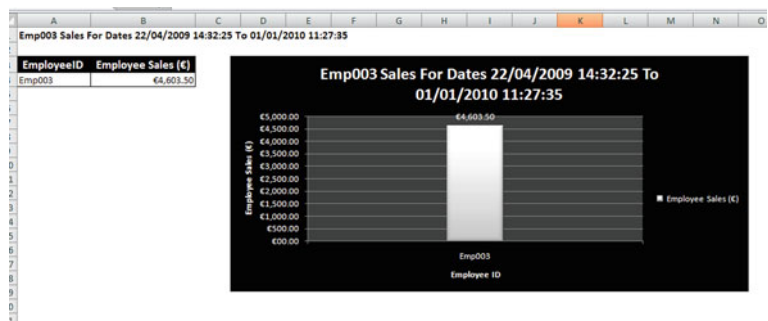


Fig. 83.9 The resulting excel worksheet of the custom query of Fig. 83.8

Intelligence information available in the hands of the business analysts and decision-makers.

83.4 Summary and Conclusions

This paper presents an educational case and tool which can be used to demonstrate in a simple but realistic way the use of BI to undergraduate CS, MIS and business students. The main purpose of the case and accompanying software is to demonstrate, with the aid of a business case that outlines some quite realistic needs of a business which strives to complete in today’s competitive marketplace, the need for Business Intelligence and the usefulness of its output to management. Alongside the case describes and follows through the process of developing such a BI Reporting System application. Such process can be replicated as needed and

thus the case can also be useful as a step-by-step example of BI application development.

References

1. Rossetti L (2006) What is business intelligence (BI)? <http://searchdatamanagement.techtarget.com/definition/business-intelligence>. (Last accessed Oct 2011)
2. Connolly TM, Begg CE (2009) Database systems: a practical approach to design, implementation and management, 5th edn. Pearson Education (US), New Jersey, pp 133–135, 1145
3. Misner S (2009) Microsoft Corporation, Business intelligence: planning your first BI solution. <http://technet.microsoft.com/en-us/magazine/gg413261.aspx>. (Last accessed Oct 2011)
4. Smith M (2010) BI is serious business. <http://businessintelligence.com/research/302>
5. Kroenke DM, Auer DJ (2009) Database processing: fundamentals, design, and implementation, 11th International Education. Pearson Education (US), New Jersey, pp 134–136, 162–163, 556–571, 584–585
6. Olszak CM, Ziemba E (2007) Approach to building and implementing business intelligence systems. *Interdiscipl J Inf Knowl Manag* 2:135–148
7. Elmasri RA, Navathe SB (2010) Database systems: models, languages, design, and application programming. 6th Global Education, Pearson Education (US), New Jersey, pp 3–4, 23, 309
8. IBM Corporation, IBM SPSS statistics products (2010) <http://www.spss.com/software/statistics/products>. (Last accessed Oct 2011)
9. TARGIT A/S (2010) TARGIT BI suite—important features. http://www.targit.com/en/Products/TARGIT_BI_Suite/Features. (Last accessed Oct 2011)
10. Evans P (2010) Business intelligence is a growing field. <http://www.databasejournal.com/features/article.php/3878566/Business-Intelligence-is-a-Growing-Field.htm>. (Last accessed Oct 2011)
11. Microsoft Corp (2007). Access 2007 specifications. <http://office.microsoft.com/en-us/access-help/access-2007-specifications-HA010030739.aspx>. (Last accessed Oct 2011)

Chapter 84

A Process Model for Supporting the Management of Distance Learning Courses Through an Agile Approach

Amélia Acácia M. Batista, Zair Abdelouahab, Denivaldo Lopes
and Pedro Santos Neto

Abstract Agile principles can be used appropriately in management of general projects, outside the object of study of computer science and allow a quick adaptation to new realities. In this work we propose a management process model of distance learning courses within the context of the Open University of Brazil (Universidade Aberta do Brasil—UAB) in order to manage the flow of activities in the construction of a distance learning course. This research work is based on principles of agile management, model driven engineering and management models for supporting distance learning courses. We provide a domain-specific language based on agile method adapted to UAB context. We propose a prototype of our approach based on models.

84.1 Introduction

Open University of Brazil (Universidade Aberta do Brasil—UAB) is an integrated system for public universities that offer college-level courses using distance learning to population groups with difficulties of access to university education [1].

A. A. M. Batista (✉) · Z. Abdelouahab · D. Lopes
Federal University of Maranhão—UFMA, São Luís-MA, 65080-040 Brazil
e-mail: ameliacaciamb@gmail.com

Z. Abdelouahab
e-mail: zair@dee.ufma.br

D. Lopes
e-mail: dlopes@dee.ufma.br

P. S. Neto
Federal University of Piauí—UFPI, Teresina-PI, Brazil
e-mail: epasn@ufpi.edu.br

This system was created by Brazil government in 2005 and it extends to all regions of the country. The characterization of this environment has allowed us to observe and ascertain the difficulties of managing and controlling the flow of activities that involve the construction of a distance learning course.

Within this context, we propose a process model for the management of distance learning course, here called Agile-UAB Management Process Model, in order to achieve the process control and, as a result, to ensure a better quality product.

This research work started with a bibliographic survey, including authors as Pressman (2006), Kniberg (2009), Mellor (2005), Dhamer (2006) and Rodrigues (1998). This theoretical basis allowed us to identify shortcomings and propose improvements to UAB current management process. In the next step, agile method kanban was adapted to Agile-UAB Management Model (process tailoring). The method Kanban is an approach that introduces changes in a project management methodology [2]. Besides, it is extremely adaptive and uses a visual control mechanism to follow the work flow, giving transparency to the management process. Once completed this process tailoring stage, it was achieved the framework meta modeling, which resulted in a domain specific language (DSL) to specify the process model for distance learning courses.

This paper is organized in six sections described this way: in the first section a short explanation about the goals and motivation of this work; in the second section we present a general view on the main areas involved in this project; in third section we deal with the UAB management process model proposal, its flow and modeling in terms of Kanban; in the fourth section, we discuss about related works. Final considerations are reported in the fifth section.

84.2 Overview

84.2.1 Agile Project Management

Historically, the subject project management is in a new stage, in which techniques and methods, called traditional, started to be questioned [3]. The signing of the Agile Manifesto [4] was one of the landmarks for the movement that criticizes the traditional techniques of project management. Of course, there is no doubt about the good results produced in the classical approaches. However, the agile approaches propose a new methodology that deals with projects based on knowledge and creativity.

Among several agile modalities available in the literature, we found Kanban, a method that uses a visual control mechanism to follow the work flow, giving transparency to the management process. Compared to other agile methodologies, such as Scrum and RUP, it is less prescriptive and easier to implant. Such characteristics elected Kanban as a tool to support the activity flow of Agile-UAB Management Model.

84.2.2 Distance Learning

At the first moment, projects for distance learning may seem simple and easy to manage. However, its multidisciplinary characteristic associated with the need for change in didactic-pedagogic behavior of the teacher and the absence of an efficient technological infrastructure may result in failures or unsuccessful distance learning courses. Besides, in Education there is a considerable number of teachers who are not familiar with computer, software and Internet. Thus, the use of Information Technology (IT) can be barriers to teachers aiming transmit knowledge to students. In this case, training is a fundamental step. All these characteristics, if not properly evaluated and managed, may endanger the project success.

Tools such as RedMine, OpenProj and MSProject could be used to manage EaD projects. However, the project complexity level is measured by its particularities, features and the environment in which it will be inserted and, therefore, nothing more appropriate than designing a specific domain tool. Distance Learning projects fit in this context. Moodle (Modular Object-Oriented Dynamic Learning Environment), TelEduc and AMADEUS are learning management systems to manage educational activities for creating online communities, in virtual environment oriented for collaborative learning. The choice of this kind of tool defines the modality of distance learning adopted by the institution, e.g. UAB uses Moodle as its virtual learning platform.

84.2.3 Model Driven Engineering

Model driven engineering (MDE) Models consist of sets of elements that describe some physical, abstract or hypothetical reality. Good models work as means of communication [5]. MDE is an approach where models are at the center of software artifact development [6]. Model driven architecture (MDA) is an example of MDE application. The development of a basic MDA framework is related to the transformation process of a source model, defined as Platform-independent model (PIM) for a target model called platform-specific model (PSM). As benefits provided by MDA we have such an improvement in productivity, portability guarantee and system interoperability, besides an easier maintenance process with a better quality documentation. Figure 84.1 represents a basic MDA framework [6].

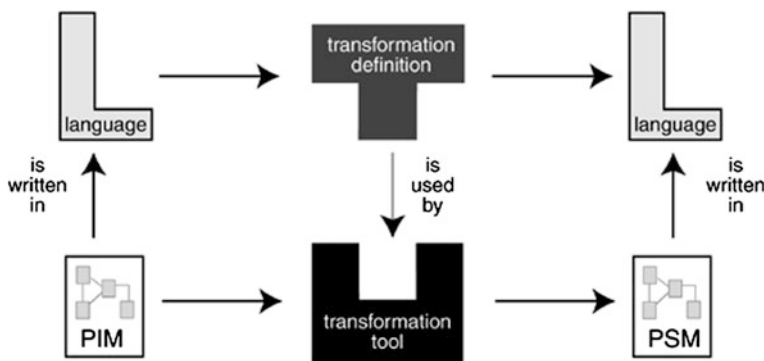


Fig. 84.1 A basic MDA framework [6]

84.3 Agile-UAB Management Model

84.3.1 Proposed Model

The management of distance learning course at UAB is based on the model proposed by Rodrigues [1998], which was appropriate for the UAB system operating mechanism. However, during the UAB operating environment characterization, we observed some inconsistencies between what was supposed to be done and what really happens in practice, what certainly leads the process to management problems. Our proposal is to adapt and extend the current UAB's management model in order to introduce positive aspects provided by the management model designed by Dhamer [2006]. Among the existent problems in UAB model, we highlight: low quality of learning materials and difficulty to produce and deploy a pedagogical academic program. Figure 84.2 represents our proposal of activity flow of Agile-UAB management model.

The model is made up by five stages: requirement analysis, planning, implementation, course evaluation and knowledge acquisition. Each one of these stages is described as follows:

- **Requirement analysis:** for distance learning course projects, it is considered as requirement the profile of the student, the teacher, the tutor, and the course itself. Student is our target; teachers and tutors will interact directly with the construction and execution of the project. Besides, the definition of these profiles will be the base for us to determine what kind of media and pedagogical strategy will be used in the course arrangement. At this stage, the course managers will be able to reevaluate their requirements according to the results that they have obtained during the course evaluation stage, in order to improve its structure and management;
- **Planning:** this stage consists of five activities, according to UAB's context. In the Course Project activity, the course model is created by the coordinator, supported

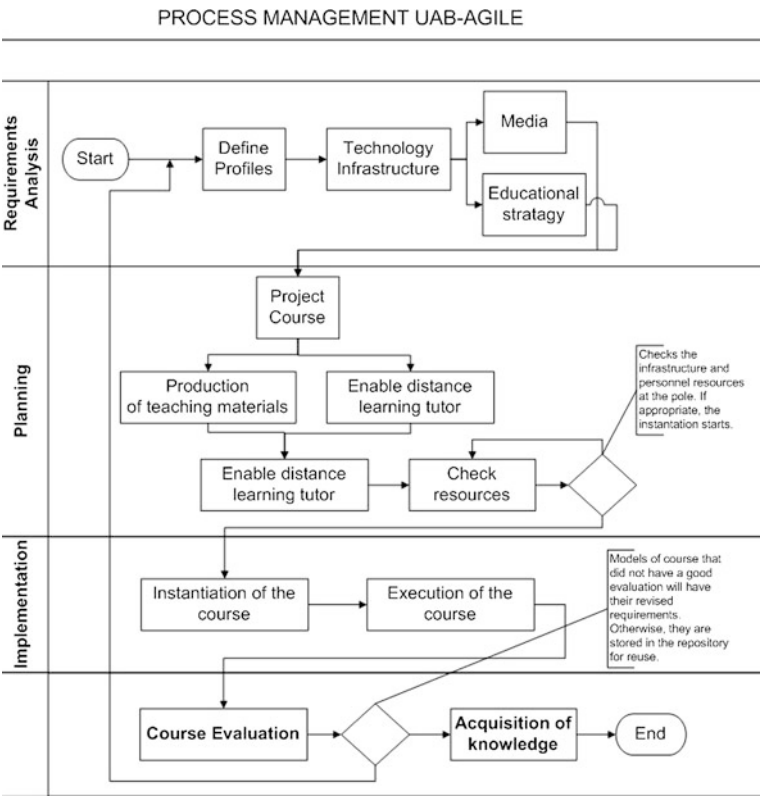


Fig. 84.2 Activity flow of the agile-UAB management model

- by a teaching staff. The development of the course model may be improved through the reuse of projects evaluated positively, which are stored in a database originated from the stage of Knowledge Acquisition, it is as the last step of the model proposed in this paper and will be described below. Didactic Material Production consists in the selection, preparation and review of teaching material, and such activity is achieved by a multidisciplinary staff, and in parallel with this activity occurs the Tutors' Capacity, i.e. enable distance learning tutor; completed this activity, a new training starts with tutors using the teaching material already produced; thereafter, the basic resources for implementation of the course are checked and if appropriate will enable the next step, instantiation, is initiated;
- **Implementation:** means structuring the course in a physical place according to available infrastructure conditions. At this stage, two activities occur, the instantiation of the course followed by the execution of the course. In the first activity we have a relative possibility of customization of the course model, if this has been reused. In this stage, students, teachers and tutors put into practice everything was planned and documented in the course project activity in the Planning phase;

- **Course Evaluation:** during the course execution activity, evaluation mechanisms will be applied, both for students and teachers who can generate data that allow a qualitative analysis in several levels, from the quality of material used to students' to teachers' performance. Moreover, some evaluation elements can be extracted from the learning management system (LMS), e.g. Moodle. These data will be employed to elaborate a report and this will form the basis for the reevaluation of requirements in order to improve the quality of the projects;
- **Knowledge acquisition:** in this stage, all course project models that were considered as good quality ones are stored, according to reports obtained at the evaluation stage. These models can be reused during the Course Project activity in the Planning stage and, consequently, it will make the process quicker, with the reduction of effort and time dedicated for creating the course model.

84.3.2 Adaptation of the Agile-UAB Management Model to Kanban Method

Tailoring, in the Software Engineering context, is described as the process of customization and combination of methodologies for software development. This process is used to adjust the development methodology to a project in particular, taking its context into account and it can be achieved in two levels: organizational and development level [9]. The Agile-UAB management model was adapted to be conform to Kanban agile method, considering the organizational level, since this level is related to the practices that conduct the activities of planning and control, focus of the project management of this work. First, the roles, phases, tasks and artifacts present in the proposed method were defined. To organize this task, we made use of the Praxis Nomenclature [10], a software development process based on unified process (UP), which is an object oriented notation similar to UML and to IEEE standards of the Software Engineering. Once this definition was achieved, the elements were inserted and accommodated to Kanban context. The sequence of Figs. 84.3, 84.4 and 84.5 shows an example of this mechanism regarding Requirement Analysis phase.

Consider AT-ER1, AT-ER2, AT-ER3 and AT-ER4 as activities of the Requirement Analysis phase. For each phase defined in the model, a specific Kanban board was building. The output of a board is configured as the inputs necessary to start the execution of the next board. Next, the description of these activities and their respective artifacts are as follows:

- **AT-ER1:** obtains the requirements from the students, the teachers, the tutors and the course profile definition, besides the basic infrastructure;
- **AT-ER2:** organizes the requirements and defines the relationship among them;
- **AT-ER3:** identifies the media and the appropriate pedagogical strategy;

Fig. 84.3 Table according to Kanban presenting the flow of activities on the 1st phase of the management model UAB

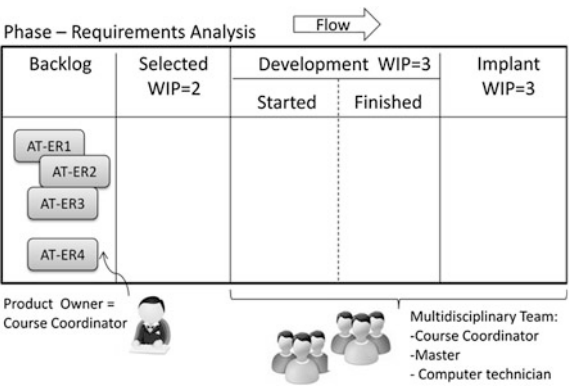


Fig. 84.4 Selection of the first activities to initiate the process

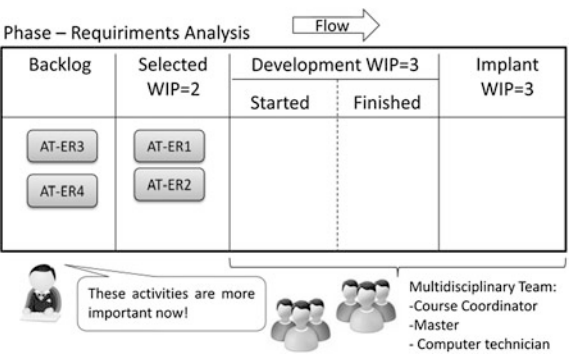
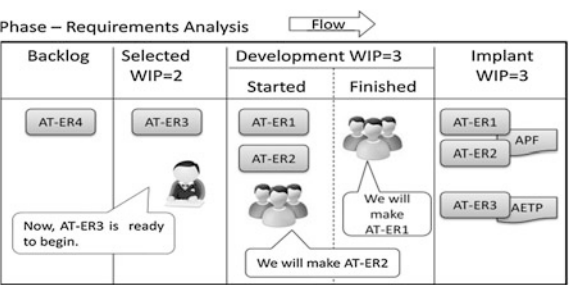


Fig. 84.5 Termination of flow activities AT-ER1, AT-ER2 and AT-ER3



- AT-ER4: achieves informal check of the course management process based on the course evaluation reports, performance of its collaborators and quality of the material.

The artifacts generated with the completion of the activities AT-ER1, AT-ER2 and AT-ER3 are a document describing the characteristics of the course, target audience and its collaborators (teachers and tutors) that will interact directly with the virtual learning environment, identified by APF and the document that tells the media and pedagogical strategy to be addressed in the pedagogical project following

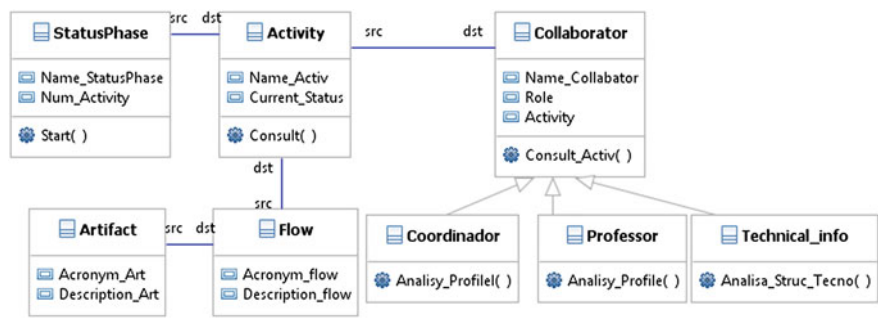


Fig. 84.6 Business model: requirements analysis phase

the guidelines outlined in the APF, this called AETP. AT-ER4 activity generates a document that presents an analysis of the reports generated in the evaluation phase about the course model and its identification is AAMC.

On Kanban board, that represents the requirement analysis phase, four states of the process flow were defined. Activities distributed without a predetermined order of priority are in the Backlog. It is the initial state of the process represented in Fig. 84.3, from which take part the Course Coordinator, assuming the role of Product Owner, and the multidisciplinary team, made up by teachers, computer technicians and the course coordinator himself, who continuously monitors the flow.

In a second stage, here represented by Fig. 84.4, the product owner selects and prioritizes some activities, considering the Work-in-Progress (WIP) of the current state, in the WIP = 2 example. The flow is continuous, always considering this criterion.

In Fig. 84.5, the multidisciplinary team is organized to implement the initiated activities, releasing the desktop of the state Selected, and this enables the product owner to determine the next activity to be started in the flow.

The flow is finished when all activities have gone through the state Implanted and its corresponding artifacts (APF, AETP) will be inputs to the beginning of the next flow related to planning phase.

After the Agile-UAB management model was adapted to Kanban method, it was built a metamodel for each phase of the process. These metamodels are not simply restricted to be used to create models and support model transformation definitions. Its goal is wider: managing the complexity and maintenance of the information in a distance learning course. The addition of a new activity at a particular stage, for example, will not be done in terms of code, but directly on the metamodel. Figures 84.6 and 84.7 represent, respectively, the business model (PIM) and the metamodel of the Requirement Analysis phase of the Agile-UAB management model.

Eclipse tool, Galileo version, was used to support the creation of the PIM model (see Fig. 84.6) according to the proposed metamodel (see Fig. 84.7). The model proposed at this paper, as it is mentioned at Sect. 84.2, presents five stages and for each one of them it was implemented a metamodel that, integrated, they constitute

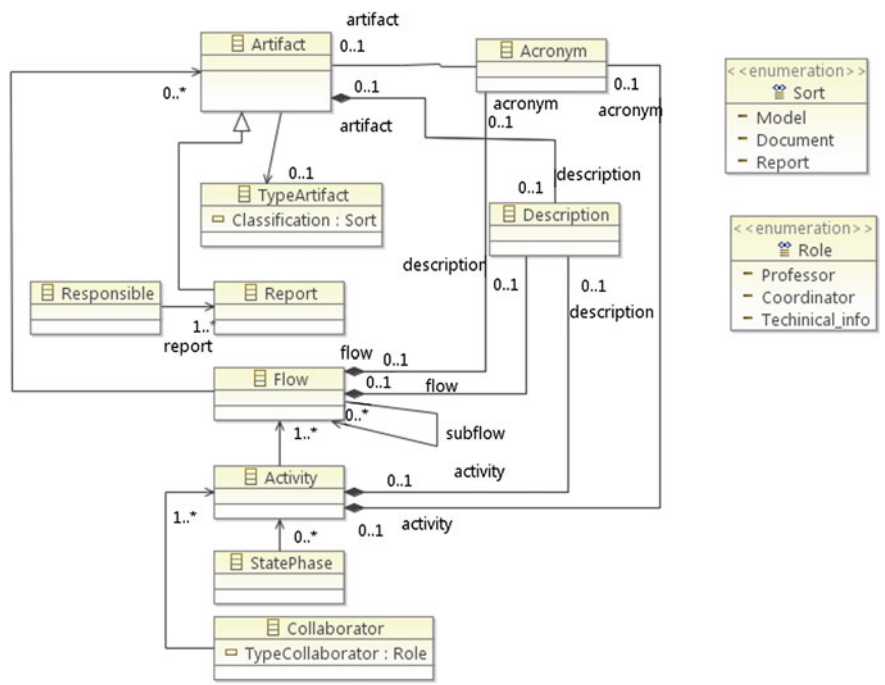


Fig. 84.7 Metamodel requirements analysis phase

the metamodel that represents the system as a whole. We chose the Requirement Analysis stage to demonstrate the execution of the process, which mechanism is applied to all other phases.

In the PIM model of the Requirement Analysis stage, we defined the Status-Phase class that represents the states in which an activity can be found when selected by the product owner for the beginning of its execution. This class is directly related to Activity class, which can contain one or more flows that may generate artifacts. All activities are performed by one or more Collaborators and these ones can be either a Professor, a Coordinator or a Technical_info.

Figure 84.7 represents the metamodel of the Requirement Analysis stage that allows the construction of the PIM model. Artifact, Flow and Activities classes contain Description and Acronym. Two Enumeration were built so that we could define two necessary types, Role and Sort. Internally at a flow we can have a subflow. At the Requirement Analysis phase, existing flows do not generate subflows, but at the Planning phase, for example, during the Training of Tutors activity we have the managing flow that is called Process Engineering and, internally to it, the Training Management subflow. Remembering that we are using the nomenclature based on Praxis [10].

A prerequisite for developing a DSL is mature domain knowledge, thus we performed the characterization of the environment UAB and a detailed survey of

its requirements obtained from managers and teachers of distance courses. DSLs are usually declarative languages and can therefore be seen as specification languages, not just as a programming language [15].

84.4 Related Works

The management of a distance course at UAB is based on the model proposed by Rodrigues [7]. This model is made up by four stages: planning, production of material, implementation and evaluation.

Some critical points that influenced the project and management of distance courses were perceived during the characterization of the UAB environment. First, the restructuring of the Pedagogical Project of the classroom Course in the pattern of distance learning course. In practice, what is observed is the reuse of the Pedagogical Project, which implies a risky procedure, since the methodology used in classroom courses differs considerably from the teaching methodology adopted in distance learning courses. Secondly, the course evaluation system. Moodle platform used in the UAB environment has reports that indicate the amount of access and operations carried out by both the teacher and the student, and through these reports, the course coordinators can make its monitoring. In the collaborators' statements who took part of the system characterization process (coordinator, teacher), the lack of an evaluation mechanism that allows, for example, the quality checking of the teaching material produced was quite clear. Even with the course coordinators' monitoring in this production process, there are still difficulties in material quality control. Besides, the lack of skilled human resource, in many centers, to perform the several roles described in the UAB system (tutor-teacher, contents-teacher, subject-teacher, etc.) is an aggravating factor, because many professionals overload of activities and eventually jeopardize the course functioning. The figure of the instructional designer, for example, is almost nonexistent.

In the face of this setting, researches advanced in search of an alternative to solve these critical points, but at least minimize their occurrences. As a result, we structured a distance learning approach to support the UAB's context, which includes features of the current model (based on Rodrigues [7]), as well as the process model designed by Dhamer [8]. Table 84.1 presents a comparative analysis among these models and the proposed Agile-UAB management model.

Rodrigues's and Dhamer's models are represented respectively by Figs. 84.8 and 84.9.

Table 84.1 Comparative analysis of different models of case management distance learning courses

| Model | Rodrigues [7] | UAB-current (2005) ^a | Dhamer [8] | Agile-UAB Management Model |
|---------------------------------------|---|---|--|--|
| Requirement analysis | It considers the students profile | Student profile, technological infrastructure | Student profile, máster, tutor, course and infrastructure technological. | Student profile, master, tutor course and infra structure technological. |
| Material production | Prepared by multidisciplinary team | Prepared by multidisciplinary team | Does not consider this activity | Prepared by multidisciplinary team |
| Reuse | No | No | Yes | Yes |
| Application | Small courses, few subjects | Courses in public universities—UAB-System | Distance Learning Courses | Courses in public universities—UAB-System |
| Pedagogical project of the course—PPC | Based on PPC classroom | Based on PPC Classroom | Establishes new PPC | Establishes new PPC |
| Media | Print, video conferencing, teleconferencing | Platform for on-line learning | Platform for on-line learning | Platform for on-line learning |
| Number phases | 4 | 4 | 6 | 5 |

^a The characteristics of this model were related to the characterization of the environment as UAB in one of its regional headquarters

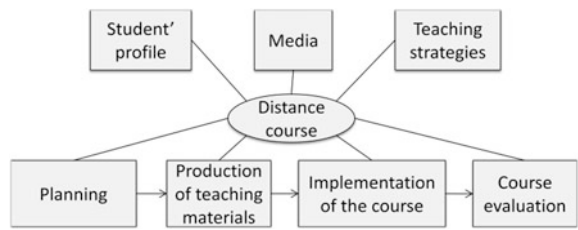


Fig. 84.8 Model for producing distance learning courses [7]

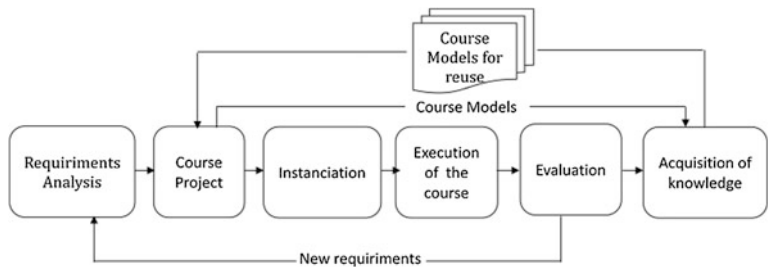


Fig. 84.9 Process model course [8]

84.5 Conclusion

The work discussed in this paper was motivated by the representative growth in recent years of the undergraduate courses focusing on the distance learning methodology, as well as the perceived management weakness of these courses in the Open University of Brazil context (UAB context). Thus, we propose an Agile-UAB management model to describe activity flow control based on Kanban.

Academically, the proposal contributes to the generation of a DSL for managing the projects of distance learning courses. The proposed metamodel is an adaptation of agile-UAB management model to Kanban.

In next steps of this research work, we aim to develop a prototype using the eclipse modeling framework (EMF) [11] and graphical modeling framework (GMF). The implementation of the prototype will automate the process and facilitate the verification and validation at the test environment. It is expected, with these contributions, to become the UAB more productivity, more efficient and improve the quality of courses produced.

Acknowledgments This research work was sponsored by Maranhão State Government through FAPEMA and Federal Government of Brazil through CAPES.

References

1. Brasil. Decree n° 5.800, from June 8 2006. Establishes the Open University of Brazil System for the development of distance education modality, in order to expand and take to the countryside the offer of courses and programs of higher education in the Country. Official Gazette. Brasília, June 8 2006
2. Kniberg H, Skarin M (2011) Kanban and Scrum: getting the best of both. Available at: <http://www.infoq.com.br/minibooks/kanban-scrum-minibook>. Access on Jan 15 2011
3. Conforto EC (2011) Project agile management: method proposal and evaluation for scope and time management. Available at: <http://www.cascavel.cpd.ufsm.br/tede/tde.busca/arquivo.php?codArquivo=3060>. Access on 22 Jan 2011
4. Beck K et al (2010) Manifest for agile software development. 2011. Available at: <http://agilemanifesto.org>. Access on 28 Dec 2010
5. Mellor SJ et al (2005) Distilled MDA architecture principles guided by models. Ciência Moderna Publishing House, Rio de Janeiro
6. Lopes D (2011) MDA: Model driven architecture. Available at: <http://www.leserc.dee.br>. Access on 02 Oct 2011
7. Rodrigues SR (1998) Evaluation model for distance learning courses. 185 f. Dissertation (Master's Degree in Production Engineering) Universidade Federal de Santa Catarina, Florianópolis
8. Dhamer A (2006) A model for course process 132 f. Thesis (Ph.D. in Computer Science) Universidade Federal do Rio Grande do Sul, Porto Alegre
9. Segal J (2005) When software engineers met research scientists: a case study. *Empir Softw Eng* 10(4):403–404
10. Filho WPP (2009) Software engineering: fundamentals, methods and standards. 3edn. LTC
11. Steinberger D et al (2009) EMF: eclipse modeling framework. 2edn. Addison–Wesley professional
12. Pressman R (2006) Software engineering. 6 edn GrawHill, São Paulo
13. Sommerville I (2007) Software engineering. 8edn. Pearson, São Paulo
14. Molhanec M (2007) The Agile methods: an innovative approach in the project management. In: 30th International Spring Seminar on Electronics technology, pp 304–307, 9–13 May 2007
15. Deursen AV, Klint P, Visser J (2000) Domain-specific languages: an annotated bibliography. Newsletter ACM SIGPLAN Notices. vol 35 Issue 6, New York, USA. Available at: <http://dl.acm.org/citation.cfm?doid=352029.352035>

Chapter 85

Numerical Modeling of Electromagnetic Induction Heating Process Using an Inductor with Constant Step Between Turns

Mihaela Novac, Ovidiu Novac, Mircea Gordan
and Cornelia Gordan

Abstract This paper presents a numerical modeling for an induction heating system with axial symmetry. Our induction heating system use cylindrical pieces and one particular property of the inductor is the constant step between the turns of coils. We use a simulation program that will give us a lot of information that are experimentally inaccessible. To obtain a high precision numerical modeling we use a performant software (Flux2D).

85.1 Introduction

The numerical modeling of the heating process through electromagnetic induction implies the calculus of the eddy currents, which the basically the cause of the electromagnetic losses due to the development of heat in the piece. While at the beginning, most numerical methods were based on Finite Differences Method (FDM), today, the vast majority of applications are modeled using induction Finite Element Method (FEM). However, we must know that the early development of Finite Element Method (FEM) dates from 1960, but not for electromagnetic problems. With

M. Novac (✉) · M. Gordan

Department of Electrical Engineering, Faculty of Electrical Engineering and Information Technology, University of Oradea, 1 Universităţii Str, 410087 Oradea, Romania
e-mail: mnovac@uoradea.ro

O. Novac

Department of Visual Arts, Faculty of Arts, University of Oradea,
1 Universităţii Str, 410087 Oradea, Romania

C. Gordan

Department of Electronics, Faculty of Electrical Engineering and Information Technology,
University of Oradea, 1 Universităţii Str, 410087 Oradea, Romania

current trends of computing techniques development and advanced software for electrical engineering domain, numerical calculation of induction heating process has become very common [1]. An induction heating process generally involves the coupling of electromagnetic phenomenon with thermal one. In addition to power systems, modeling and designing of inductor coil is a difficult task because both aspects, electromagnetic and thermal, of the process and their coupling must be taken into account in the designing process. Initially, the inductor design was based on trials and errors that led to several empirical design formulas. These analytical models are suitable only for simple geometries such as round bars, tubes, plates, etc. In the case of complex geometries, such as toothed wheel, is very difficult, even impossible, to model the heating application. Moreover, the problem is very complicated due to nonlinear material properties. The development of numerical methods [e.g. Finite Element Method (FEM)] and of numerical simulation programs, have made it possible, inductor behavior simulation, before prototype building and avoiding the design mistakes [2]. This is a remarkable advantage for inductor coil designers because most of inductor design work can be done using computer simulation, thereby reducing the designing effort and reducing occurrence time of other new developments, because the simulation program gives us a whole series of experimentally inaccessible information. Software to numerically model induction processes has advanced to the point where, today, several very sophisticated packages are available commercially and are able to take into account almost all important factors in both 2D and 3D.

With the development of numerical modeling programs for induction heating, were also developed optimization programs. First works on this subject have appeared in 1995 and since then, a variety of classical optimization techniques have appeared and have been applied to induction heating processes [3, 4].

These advantages of computer simulation, contribute to the development of innovative products and processes.

85.2 Mathematical Model of Numerical Modeling

The mathematical models of electromagnetic and thermal fields, based on the specific laws of this phenomenon are described by partial differential equation. In this paper, for the coupled electromagnetic and thermal field, the general equation used are, [5–7]:

$$\nabla \times \mathbf{E} = - \frac{\partial \mathbf{B}}{\partial t} \quad (85.1)$$

$$\nabla \times \mathbf{H} = \mathbf{J} \quad (85.2)$$

The material equations are:

$$\mathbf{H} = \nu \mathbf{B} \quad (85.3)$$

$$\mathbf{J} = \sigma \mathbf{E} \quad (85.4)$$

The magnetic flux density \mathbf{B} , using the magnetic vector potential \mathbf{A} , is:

$$\mathbf{B} = \nabla \times \mathbf{A} \quad (85.5)$$

where ϕ is the electric scalar potential. So we obtain:

$$\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t} - \nabla \Phi \quad (85.6)$$

$$\nabla \times (\nu \nabla \times \mathbf{A}) = \mathbf{J} \quad (85.7)$$

The current density is given by the relation

$$\mathbf{J} = -\sigma \frac{\partial \mathbf{A}}{\partial t} - \sigma \nabla \Phi \quad (85.8)$$

It satisfied the continuity equation

$$\nabla \cdot \mathbf{J} = 0 \quad (85.9)$$

By substituting Eq. (85.8) in relation (85.7) and (85.9), the equations for the vector and scalar potentials become:

$$\nabla \times (\nu \nabla \times \mathbf{A}) + \sigma \frac{\partial \mathbf{A}}{\partial t} + \sigma \nabla \Phi = 0 \quad (85.9a)$$

After transformations the Eq. (85.9) takes the following form:

$$\nabla \times (\nu \nabla \times \mathbf{A}) + j\omega \sigma \mathbf{A} = \underline{\mathbf{J}}_s \quad (85.10)$$

where:

$$\mathbf{A} = \mathbf{A}(\mathbf{P}, t) = \mathbf{A}(\mathbf{P})e^{j\omega t}$$

In the solution of the field equations, the boundary and initial values of the vector potential must be known.

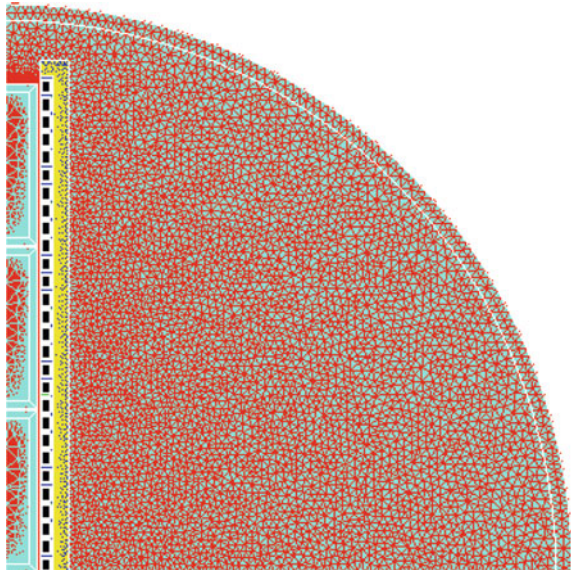
The thermal field is modeled by:

$$\gamma C \frac{\partial T}{\partial t} - \nabla \lambda \nabla T - \lambda \nabla^2 T = J^2 \varphi(T) \quad (85.11)$$

In the heat transfer process there are three kinds of boundary conditions commonly used.

The first boundary condition (Dirichlet boundary condition), corresponds to a surface at a fixed temperature T_s , [4],

$$T(t) = T_s \quad (85.12)$$

Fig. 85.1 Mesh

The second condition (Neumann boundary condition), corresponds to the existence of a fixed or constant heat flux at the surface. This heat flux is related to the temperature gradient at the surface by Fourier's law:

$$g'' = \lambda h \frac{\partial T}{\partial n} \quad (85.13)$$

The third boundary condition corresponds to the thermal convection at the surface:

$$-\lambda_h \frac{\partial e}{\partial n} = \alpha_s (T_s - T_a) \quad (85.14)$$

Thus the complete third boundary condition is:

$$-\lambda h \frac{\partial T}{\partial n} = \alpha_s (T_s - T_a) + C_s (T_s^4 - T_a^4) \quad (85.15)$$

where c_s is the radiation coefficient and α_s

represent the convections. For a surface I the radiation coefficient c_s is given by:

$$C_{s,i} = \sigma h \cdot \varepsilon_i \cdot F_{ij} \cdot A_i \quad (85.16)$$

where A_i is the area of surface i, F_{ij} is the view factor from surface I to surface j, ε is the emissivity and σ_h is the Stefan–Boltzman constant [8].

Fig. 85.2 **a** Border conditions for electromagnetic field problem
b Border conditions for thermal field problem

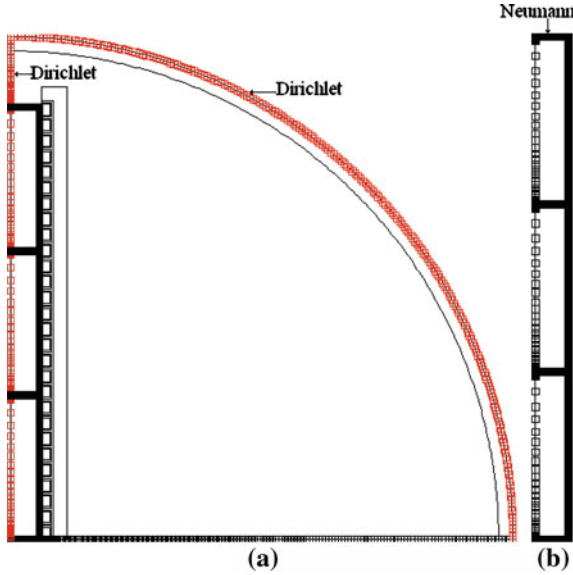
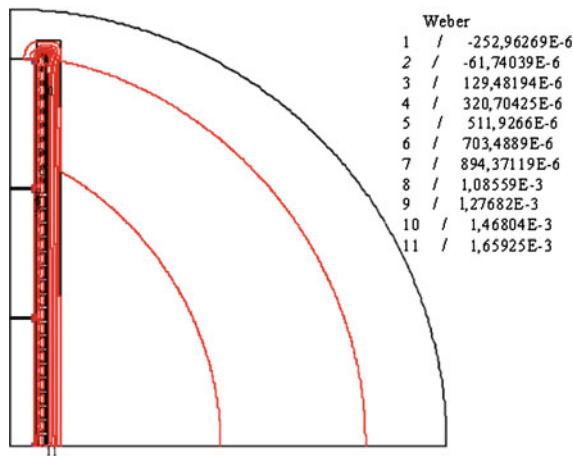


Fig. 85.3 Magnetic field lines at the start of heating process



85.3 Numerical Modeling of Induction Heating Process Results

This paper analyses induction heating system with axial symmetry. The length of inductor is $L_i = 1,240$ mm, and one particular property of the inductor is the constant step between the turns of coils. Also the shape of inductor turns is rectangular.

Fig. 85.4 Magnetic field lines at the end of heating process

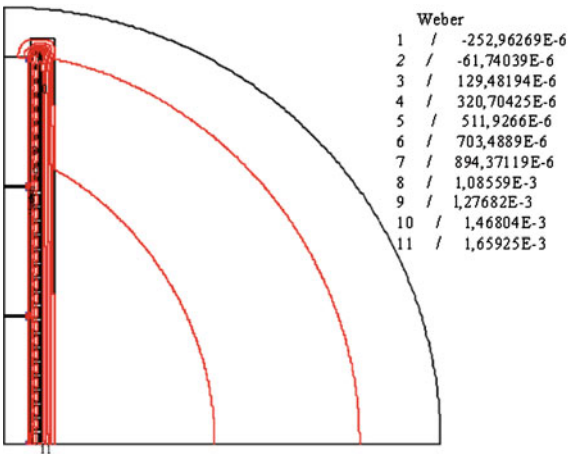
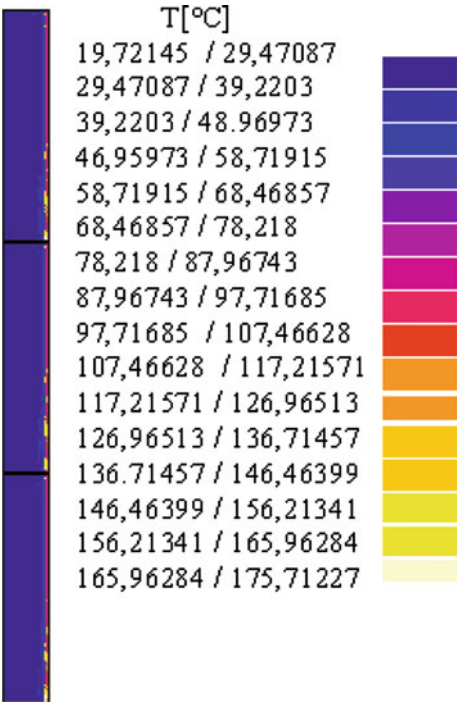


Fig. 85.5 Temperature map in piece at $t = 0.2$ s



We have obtained the following results from calculus current density $J_i = 31 \text{ A/mm}^2$, frequency $f = 2,500 \text{ Hz}$; inductor length $L_i = 1,240 \text{ mm}$, piece length $a_2 = 200 \text{ mm}$, piece diameter $d_2 = 80 \text{ mm}$, number of turns $N_{sp} = 56$.

In Fig. 85.1 we present the mesh of induction heating device.

The mesh for Finite Elements Method, of the study domain consist of triangles, the maxim density correspond to the maximal interest's area.

Fig. 85.6 Temperature map
in piece at $t = 1.527$ s

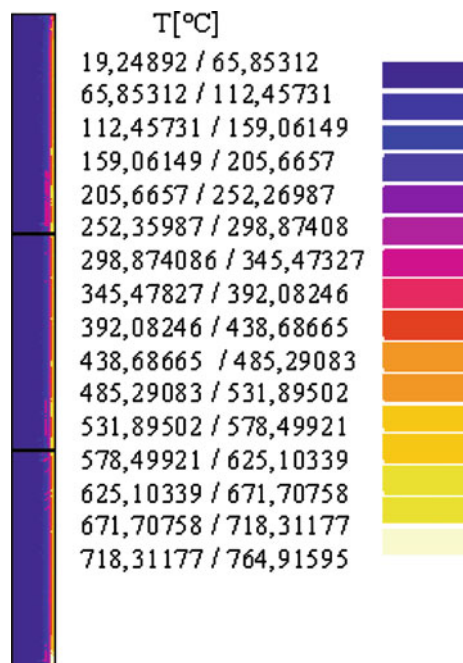


Fig. 85.7 Temperature map
in piece at $t = 5$ s

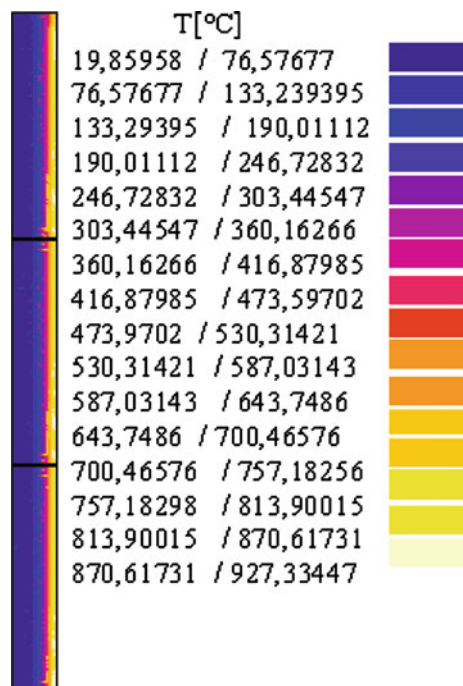


Fig. 85.8 Temperature map in piece at $t = 9.99$ s

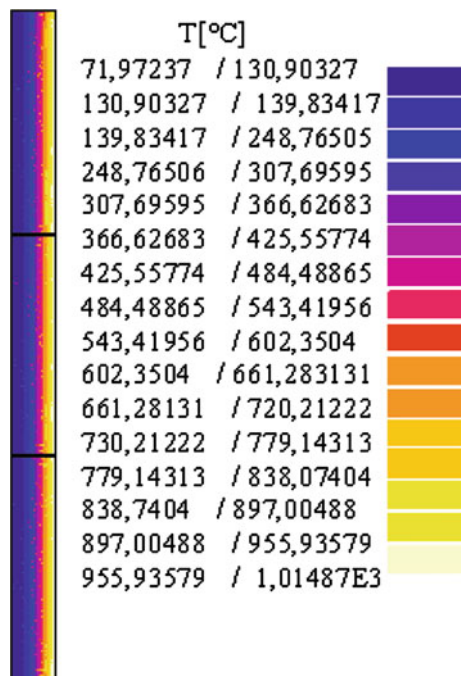


Fig. 85.9 Temperature map in piece at $t = 20$ s

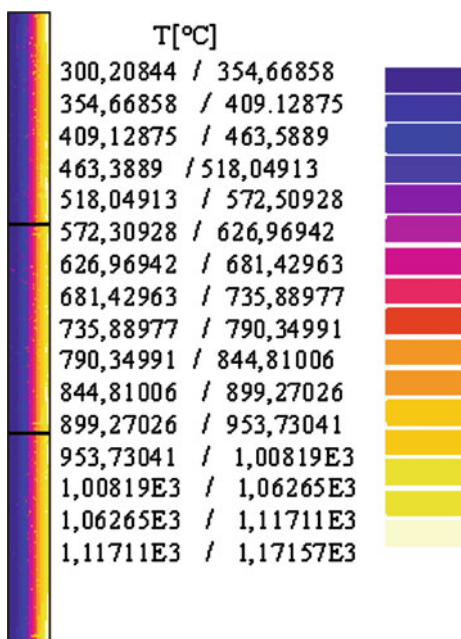


Fig. 85.10 Temperature map in piece at $t = 25.8$ s

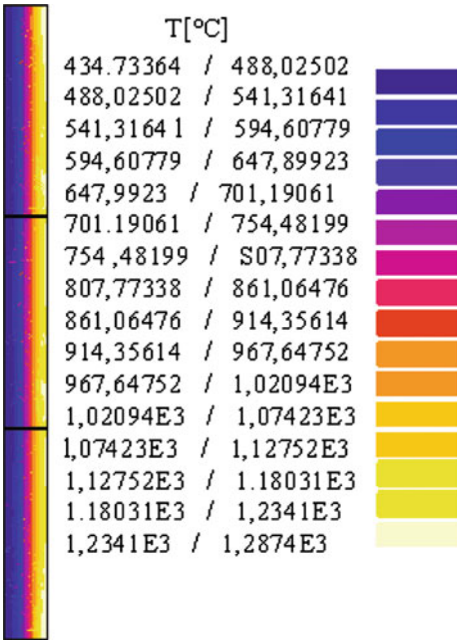


Fig. 85.11 Position of points 1 and 2 where we will make the analysis of temperature field

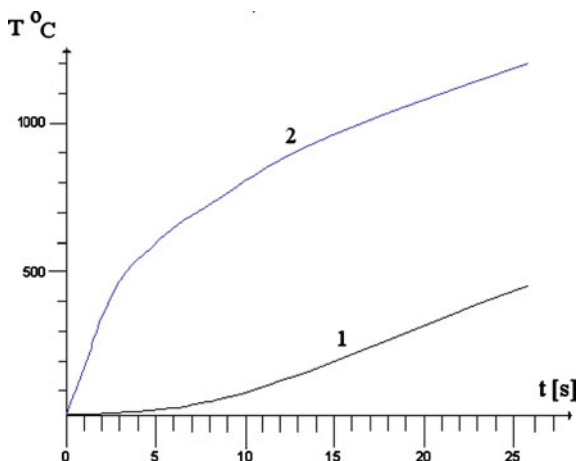


In Fig. 85.2 a is presented the border conditions for electromagnetic field problem and in Fig. 85.2 b the border conditions for thermal field problem.

In Fig. 85.3 we have presented the magnetic field lines, at the start of induction heating process and in Fig. 85.4 we have presented the magnetic field lines at the end of induction heating process.

Temperature maps in piece for different times are presented in Figs. 85.5, 85.6 and 85.7, 85.8, 85.9, 85.10.

Fig. 85.12 Curves that presents the temperature evolution in points 1 and 2



In Figs. 85.11 and 85.12 we present temperature evolution in points: 1-(5 mm, 600 mm), 2-(35 mm, 600 mm) in pieces, where we will make the analysis of temperature field.

The point 1 is situated in the middle on the piece and the point 2 is situated at the surface of the piece.

85.4 Conclusions

In this paper we have presented some results of induction heating process of cylindrical shapes pieces with constant step between two consecutive turns.

The development of FEM and of software programs for numerical simulation are very important tasks, since these make possible the simulation of inductor behavior before the construction, and so we can obtain optimal solutions for the studied applications.

This is an remarkable advantage for designers, because the designer work is done now using computer simulation programs, reducing the costs and reducing the time of coming out on the market for other new developments.

The simulation program can give a lot of information that are experimentally inaccessible. Another advantage is high precision numerical modeling, using a performant software (Flux2D) in the evaluation stage.

Another conclusion is that at the end of heating process, the pieces are uniformly heated because the distance between two consecutive turns of inductor is constant.

The disadvantage associated with this method consists in the fact that the electromagnetic design problems are very demanding in terms of computing resources, by requiring the resolution of a complex electromagnetic problem for each evaluation.

References

1. Lavers JD (2007) State of the art of numerical modeling for induction processes, HES 07, Italy
2. Tsukerman I (2005) Electromagnetic applications of a new finite-difference calculus. *IEEE Trans Magn* 41:2206–2225
3. Biro O, Preis K (1990) Finite element analysis of eddy currents. *IEEE Trans Magn* 26(2): 418–423
4. Ciric I, Hantila F (2007) An efficient harmonic method for solving nonlinear time-periodic Eddy-current problems. *IEEE Trans Magn* 43(4):1185–1188
5. Antero A (1987) Analysis of induction motors based on the numerical solution of the magnetic field and circuit equations. *Acta Polytechnica Scandinavica, Electrical Engineering Series*, No. 59, Helsinki
6. Hănțilă F, Preda G, Vasiliu M, Leuca T, Della Giacomo E (2001) Calculul numeric al curenților turbionari. Editura ICPE. ISBN:973–8067–31–6
7. Hănțilă F, Vasiliu M (2005) Câmpul electromagnetic variabil în timp, Editura Electra, București
8. Fireșteanu V, Popa M, Tudorache T (2004) Modele numerice în studiul și concepția dispozitivelor electrotehnice. Editura Matrix Rom București

Chapter 86

Satisficing-Based Approach to Resolve Feature Interactions in Control Systems

Jan Corfixen Sørensen and Bo Nørregaard Jørgensen

Abstract To handle the complexity of modern control systems there is an urgent need to develop features as independently developed units of extension. However, when independently developed features are later composed they become coupled through the shared environment resources. As a consequence, the system requirements may no longer be entailed when independent features try to control the same shared environment. Malfunctioning behavior as a consequence of feature interference is known in the literature as the feature interaction problem. This paper presents an approach that uses design-time specification of independent requirements, in combination with a runtime arbitrator that searches for feature interaction-free programs which entail the system requirements. In case of conflicting requirements that can't be satisfied simultaneously, the mechanism supports explanation of the interactions as a context sharing problem. We demonstrate our approach in a real-life control system for industrial pot plant cultivation in greenhouses and show that solutions are found for compatible requirements and that conflicts are identified and explained for incompatible requirements.

J. C. Sørensen (✉) · B. N. Jørgensen
The Maersk Mc-Kinney Møller Institute, University of Southern Denmark,
Campusvej 55, 5230, Odense, Denmark
e-mail: jcs@mmmi.sdu.dk

B. N. Jørgensen
e-mail: bnj@mmmi.sdu.dk

86.1 Introduction

Today, control systems are becoming more complex as a result of increasing demand for more advanced functionality. To cope with this increasing complexity, it is desirable that control systems are open to new extensions and can be composed from independently developed features that are reusable in different control environments. Composing control systems from independent features is however fragile to feature interactions that emerge when features interact through the systems environment in unexpected ways. Basically, independence of work implies that features developed by different independent vendors may require different conflicting requirements. If the span between requirements is narrow, it may be doable to combine the corresponding features. However, if the span is broad, the requirements are most likely conflicting, making their feature combination infeasible.

The prevailing implementation approach for control systems is to manually combine the specifications of individual features into a centralized specification governing the collective behavior of the whole control system. However this approach requires profound domain knowledge in order to find the correct control trade-offs for conflicting requirements. Hence, the correctness of the resulting control system depends on the developer's ability to identify all potential feature interactions and make the right trade-offs in each case. Consequently, whenever a new feature is added or an existing feature is modified, the developer has to reconsider all possible feature interactions and their respective trade-offs. A task which complexity increases exponentially as the number of features goes up. Additionally, control systems are built and configured differently which may require different variants of more features. When using a centralized specification approach, it therefore becomes necessary to maintain multiple versions of the central specification, each implementing a variant of the collected features. Obviously, it would be preferable, if such variations could be handled in isolation without the need of creating and maintaining several versions of the same central specification. That is, the use of a centralized specification is undesirable as it requires a global integrity check to verify that no feature interactions will emerge as a result of adding new features or changing the control system configuration.

Instead of trying to combine individual features by merging their specifications, we opt for a decentralized approach which keeps specifications of individual features separate. In order to keep the specifications of individual features separate, they have to be specified as independent units of composition that can be developed and deployed on the same system independently of each other. Systems supporting this property is said to be independently extensible [1, 2]. However, separation of specifications in independent units of composition does not by itself solve the feature interaction problem, it merely transforms the problem of merging the program-logic of individual features into the problem of coordinating their respective effects on shared controlled environment.

Generally, when combining independent units of composition they become implicitly interrelated through sharing of resources in their environment, which is recognized as a typically source for causing conflicts between requirements of individual program features [3–5]. In control systems each feature has specific requirements which may be compromised when combined with other features sharing the same environment. Conflicts caused by interference of program logic are generally referred to as the feature interaction problem [6]. Feature interactions occur whenever the modification or addition of a system feature interferes with the correctness of other system features. In the worse-case scenario such feature interactions can compromise the correctness of the overall system behavior and cause unexpected runtime faults that may lead to system failure. Hence, creation of an independently extensible control system, in which dysfunction due to feature interactions does not happen, requires an implementation approach that is capable of coordinating the effects of independent developed features on shared controlled environment in such a way that the requirements of all features are satisfied.

In this paper, we show: (1) How requirements R in natural language can be described in separate independently developed feature specifications S , (2) How feature interactions can be identified by analysing resource conflicts in the system environment, and (3) How the control programs P can be automatically adapted using a runtime arbitrator to avoid feature interactions when specifications S are composed. To our knowledge, no approach exist that uses domain specific requirement specifications S of the environment W in combination with an runtime arbitrator to find a control program P that entails the specifications S .

The remainder of this paper is structured as follows. [Section 86.2](#) describes related work that perceives the feature interaction problem as a context sharing problem. [Section 86.3](#) describes preliminaries to understand what is meant about features and feature interactions. [Section 86.4](#) introduces a real-life control example for industrial pot plant cultivation in greenhouses. [Section 86.5](#) presents the different elements of our approach based on the running example. Last, [Sect. 86.6](#) presents a short discussion about the how approach could be improved. Finally, we conclude our work in [Sect. 86.7](#).

86.2 Related Work

Our work is inspired by [3–5, 7, 8] which differ from other prevalent approaches by perceiving the feature interaction problem as a resource-sharing problem rather than a feature-behavior problem. The idea behind the approaches is based on the assumption that feature interactions emerge as a consequence of features sharing resources. The argument for focusing on resources instead of feature behaviors is that resources are simpler to model and analyse than the implicit semantics of feature behaviors. To manage feature interactions, the approach requires specification of the domain properties based solely on knowledge about the environment resources.

Jackson invented the Problem Frame Approach used to gather requirements and creating specifications for computer software [9–12]. The philosophy behind the approach is that all software development problems consist of a machine, a world and requirements. The task of the developer is to construct a machine that interacts with the world in such a way that the requirements are satisfied. An important observation behind Jackson's work is that the specification of the world is equally as important as specifying the machine functionality. The concept of problem frames is similar to the concept of design patterns in object oriented systems [13]. Problem frames capture recurring problem classes in the problem space while design patterns capture recurring design idioms in the solution space.

Jackson's approach suggests five artifacts for system development: domain properties W that describe how the environment is expected to behave, requirements R that describe what the system should do, specifications S that describe how the system should behave to achieve the requirements, programs P that rely on program abstractions C to satisfy the specifications. Finally, the approach provides three set of diagrams to describe and categorize problems in the problem space (1) The context diagram to depict the context of the problem, (2) The problem diagram for describing the specific problems, and (3) The problem frame diagram to categorise classes of problems in the problem space.

Armstrong et al. [3] elaborate on Jackson's work and provide a definition of the feature interaction problem as a context sharing problem. In particular, Armstrong et al. argue that the feature interaction problem arises from sharing environment resources, hence that features should be described in a notation that makes the resource context explicit and separate of the requirements.

Godskesen [14] provides a formal framework that captures feature interactions at three different levels: requirements level, specification level and implementation level. It follows from the framework, that feature interactions may be inherited from requirement level to specification level and from specification level to implementation level and that interactions should be considered at the level they belong to. The consequence of the directed inheritance of interactions, is that interactions belonging to one level may be detected at subsequent levels but can not be detected at levels preceding the one it belongs to. For that reason, Godskesen propose a technique to test for absence of feature interactions at the implementation level.

Bisbal and Cheng [4] contribute with a resource-oriented approach to detect and handle feature interactions in component-based software at runtime. Their requirement specification declares the resource goals of a feature. Furthermore, the domain properties are described in terms of the environment resources and their relationship with the components. Both the requirement and domain property specifications are declared at design time and used by a runtime technique to address feature interactions in resource-aware systems.

Liu and Meier [8] contribute with resource-aware contracts and address resource-based feature interactions in dynamic adaptable systems. Resource-aware contracts capture the resource usage patterns of features as well as the inherent constraints of the resources. Feature interactions are detected as resource usage

patterns (domain properties) that compromises the goals of the features or the resource constraints. In contrast to the contracts suggested by Bisbal and Cheng, resource-aware contracts support both fixed-capacity, varying-capacity, exclusive and shared resources and can be applied at both component and component-assembly level.

Zambrano et al. [7] focus on aspect interactions and use metadata annotations to specify the resource requirements of each aspect. Zambrano et al. provide a detection and a resolution strategy that can detect and avoid feature interaction in resource-aware systems, without compromising obliviousness of the aspects. The resource specification is declared as metadata on the aspects at design time in the form of semantic annotations. The interactions between aspects are avoided at runtime by a coordinator aspect. The coordinator aspect is augmented with a list of user-defined conflict situations declared at design time as conflict rules. To avoid interactions, the coordinator aspect can deactivate the conflicting aspects as declared in the action part of the triggered conflict rule.

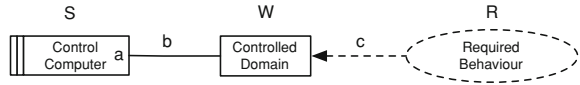
We provide a way to specify designtime specifications S of requirements, in the control domain, which map to natural language requirements R . By using a simple representation of the program P in the control domain, as a set of actuator set-points, it allows us to implement an arbitrator that uses a genetic search algorithm to find a set of set-point values which satisfy the specifications S . The idea of using an arbitrator, is to avoid rewriting the program P each time the system's requirements and thereby its specifications changes. In case no solutions can be found to resolve the feature interaction, the mentioned approaches do not support that the user can redefine the requirements of the conflicting features to find alternative solutions without rewriting the program P . Additionally, redefinition of conflicting requirements to resolve feature interactions requires explanation of what caused the feature interaction. To our knowledge none of the mentioned approaches support such detailed *explanation* of the cause of feature interactions.

86.3 Preliminaries

The relationship between Jackson's suggested artifacts are described using the logical entailment operator \vdash as follows: " $W, S \vdash R$ " and " $C, P \vdash S$." The first entailment relation states that a feature satisfies its requirements R if its specifications S combined with domain properties W hold. Similarly, the second entailment relation states that programs P relies on program abstractions C to entail the specifications S .

We apply the problem frames approach to describe and reason about problems in the control domain. Basically, the required behavior problem frame matches the problems in the control domain where the environment W is to be controlled so that it satisfies certain conditions specified in requirements R . The problem is to write a specification S of a control computer that will impose that control. The requirements are specified in terms of environment phenomena c (Fig. 86.1). In contrast, the

Fig. 86.1 Required behavior problem frame



specifications S only contain information about the shared specification phenomena b at the interface between the computer and the environment. Programs on the other hand, are specified solely in terms of program phenomena a . In our approach we map the shared environment phenomena c to shared specification phenomena b and to program phenomena a in a one to one relationship. That is, each environment phenomena c is modelled in the specification and program either as sensor inputs or actuator setpoints that logically represents a environment.

Based on Jackson's suggested artifacts and entailment relations, we define what is meant by feature and feature interaction in our approach.

Definition 1 (Feature) A feature is a set, $f = (R, S, W, P, C)$, where R represents the requirements in natural language at requirement level. S is the specification at specification level, described in term of domain properties W , that together satisfies the requirements R . P is then the program at implementation level that satisfies the specification S based on the program phenomena C .

Godskesen showed that feature interactions emerging at requirement, specification and implementation levels may be detected and handled at implementation level due to inheritance of feature interactions from one level to subsequent levels. Our approach avoids interactions by searching for a program P at implementation level that satisfy the specifications S at specification level. Additionally, each specification S maps to natural language requirements R in a one to one relationship supporting the traceability of interactions at different levels. We adopt the definition of Feature interactions by Armstrong et. al but applied at the implementation level (Definition 2).

Definition 2 (Feature Interaction) If a program P_1 satisfies a feature specification S_1 assuming program abstraction C_1 (1), and a program P_2 satisfies a feature specification S_2 assuming program abstraction C_2 (2), then it is desirable that their parallel composition satisfy the conjunction of S_1 and S_2 (3). Formally i.e.:

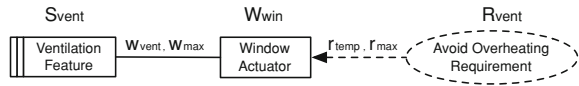
$$P_1, C_1 \vdash S_1 \quad (86.1)$$

$$P_2, C_2 \vdash S_2 \quad (86.2)$$

$$P_1 \parallel P_2, C_s \vdash S_1 \wedge S_2 \quad (86.3)$$

where C_s represents the shared program properties of C_1, C_2 , \parallel is the parallel composition operator and \vdash is the satisfaction relation. Feature interaction occurs when there are shared properties between C_1 and C_2 whose relationship is such that 3 is not true. That is $P_1 \parallel P_2, C_s \not\vdash S_1 \wedge S_2$

Fig. 86.2 Problem structure of ventilation feature F_{vent}



86.4 Example

To explain how our approach works, we base this section on a climate control system from industrial pot plant cultivation in greenhouses [15–17]. Requirements R of the climate control system are represented in natural language and states *what* the system is expected to do in the application domain. For example, consider the following natural language requirements R of a ventilation feature F_{vent} and photosynthesis optimization feature F_{photo} :

- R_{vent} Overheating the greenhouse has to be avoided by using *windows* for ventilation such that the *air temperature* never exceeds a specified *maximum temperature threshold limit*.
- R_{photo} The *leaf temperature* and indoor CO_2 level has to be optimized with respect to the actual *light level* in the greenhouse. The correlation between light level, leaf temperature and CO_2 level is described by a *photosynthesis model* [18].

The problem structure of the ventilation feature can be categorized within Jackson's required behavior problem frame [9]. That is, the ventilation feature can be described as a machine that sends ventilation air temperature setpoints (specification phenomena s_{ventSP}) to a windows actuator to keep the environment in an acceptable state where the greenhouse is not overheated (required behavior).

The requirement R_{vent} is mapped to a specification S_{vent} that is described in terms of the windows domain. The specification S_{vent} describes *how* the ventilation feature has to behave in order to bring about the changes in the environment as described in requirement R_{vent} , assuming the environment to be as specified by the ventilation air temperature setpoint w_{vent} and maximum temperature threshold w_{max} (Fig. 86.2). For example, the specification S_{vent} expresses the ventilation feature should issue ventilation air temperature setpoints w_{vent} that never exceeds a specified maximum air temperature threshold w_{max} .

Similarly, the photosynthesis optimization feature can be expressed as a required behavior problem frame (Fig. 86.3). The photosynthesis optimization feature issues heating and CO_2 setpoints (specification phenomena) to keep the environment in a photosynthesis optimized state (required behavior) given the current indoor light level. How the requirement R_{photo} must be met is specified in the specification S_{photo} . The specification S_{photo} specifies that the requirement R_{photo} is satisfied if the issued heating setpoint w_{temp} and CO_2 setpoint w_{CO_2} are close to the photosynthesis optimal temperature and CO_2 level calculated by a photosynthesis model.

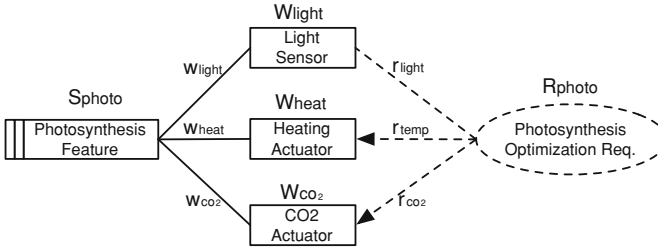
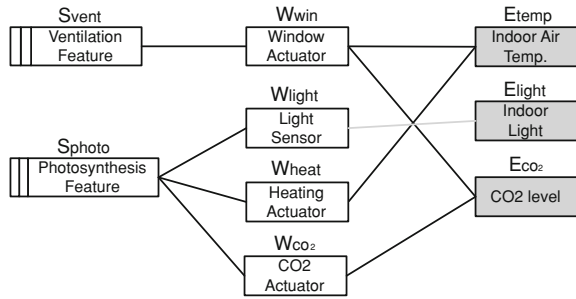


Fig. 86.3 Problem structure of photosynthesis optimization feature F_{photo}

Fig. 86.4 Context diagram illustrating feature interactions as a consequence of features sharing the same controlled environment (grey boxes)



When features are developed independently by third party vendors, they are assumed to work in isolation. That is, the program P_{vent} of feature F_{vent} will satisfy the specification S_{vent} . Additionally, because specification S_{vent} is mapped in a one to one relationship with requirement R_{vent} the satisfaction of specification S_{vent} will result in satisfaction of requirement R_{vent} at requirement level. Expressed by Jackson's entailment relations, that is $P_{vent}, C_{vent} \vdash S_{vent} \rightarrow S_{vent}, W_{win} \vdash R_{vent}$. Similarly, the entailment relations $P_{photo}, C_{photo} \vdash S_{photo} \rightarrow S_{photo}, \{W_{light}, W_{heat}, W_{CO_2}\} \vdash R_{photo}$ are assumed to be satisfied when the photosynthesis optimization feature F_{photo} is deployed in isolation. The program abstractions C_{vent} and C_{photo} represent sensor inputs (w_{max} , w_{light}) and actuator setpoints (w_{vent} , w_{heat} and w_{CO_2}).

A context diagram that includes the environment phenomena can be used to detect whether there is a shared context between features that may lead to interactions. The context diagram of the composed features F_{photo} and F_{vent} illustrates that both the windows and CO₂ actuators indirectly influence the indoor CO₂ level E_{CO_2} (environment phenomena in grey boxes). Furthermore, both the windows and heating actuators influence the indoor air temperature E_{temp} .

Both feature F_{vent} and F_{photo} will work in isolation but when they are composed together they will interact with each other through the same controlled environment as both features influence the indoor air temperature and CO₂ level (Fig. 86.4).

Photosynthesis optimization requires high air temperature and high CO₂ level and for that reason feature F_{photo} will heat the air and dose CO₂ at the same time. The high air temperature caused by feature F_{photo} will trigger the ventilation feature F_{vent} to open the windows to avoid overheating the greenhouse. The result of the interaction is that the dosed CO₂ will be led out into the atmosphere and will be wasted. A dramatic drop of CO₂ level compromises specification S_{photo} that specifies an optimal CO₂ level and as a consequence requirement R_{photo} will no longer be satisfied.

Another interaction occurs when ventilation temperature setpoint w_{vent} exceeds the heating temperature setpoint w_{heat} causing the windows opened and closed repeatedly in short intervals. The windows will open until the ventilation temperature is achieved and the windows will close. Heating will cause the window to open then the windows are closed. This loop where the windows open and close several times in succession will continue infinitely leading to wasteful heating of the greenhouse.

86.5 Approach

In this section, we describe how our approach avoids feature interactions described in Sect. 86.4 by introducing an arbitrator that uses a genetic search algorithm to search for programs that entails separate independently developed specifications.

First, we explain how specifications are expressed in our approach as accept or satisfy methods. The specification S_{vent} is described by an accept method that takes a proposed program as an argument (Listing 86.1). The program argument contains sensor inputs and actuator setpoints proposed by a genetic search algorithm. The body of the accept method is specified by the developer and returns true if the proposed program influences the environment according to the required behaviour described in requirements. That is, the accept method represents hard constraints that has to be satisfied to fulfil the requirements. In case of the feature F_{vent} the accept method returns true if the ventilation air temperature setpoint w_{vent} is below the maximum air temperature threshold w_{max} .

```

1 public boolean accept(Program p) {
2     return p.VentilationAirTempSetpoint <= p.
        IndoorTempMaximumInput;
3 }

```

Listing 86.1 Specification S_{vent} of ventilation feature

The specification S_{photo} is different from S_{vent} as it specifies an optimization of the environment. Optimizations are difficult to specify as constraints but can be expressed in satisfy methods that return a satisfiability value of how well the proposed program satisfied the optimization. A satisfy method represents a

specification of an optimization requirement. Listing 86.2 describe the specification S_{photo} as a satisfy method that returns a satisfiability value calculated as the difference between the heating temperature setpoint and photosynthesis optimal temperature.

```

4 public double satisfy(Program p) {
5     Celsius photoOptTemp =
6         calcPhotoOptimalTemp(p.indoorLightInput, p.
7             photoOptimalPctInput);
8     return absDifference(photoOptTemp, p.
9         heatingSetpoint);
10 }

```

Listing 86.2 Specification S_{photo} of photosynthesis optimization feature

The returned satisfiability value has to be specified within the interval $[0; 1]$ where zero represents the best satisfiability and one the worse satisfiability.

The feature interactions that emerge when the feature F_{photo} and F_{vent} are composed can be avoided by writing two additional specifications. One specification S_{wheat} avoids the wasteful heating interaction by specifying that a program only is acceptable if the heating temperature setpoint is below the ventilation temperature setpoint (Listing 86.3).

```

9 public boolean accept(Program p) {
10     return p.heatingSetpoint < p.
11         ventilationTempSetpoint;
12 }

```

Listing 86.3 Specification S_{wheat} to avoid wasteful heating interaction

A second specification S_{wCO2} prevents the wasteful CO_2 dosing interaction, specifying that ventilation only should be allowed when the CO_2 doser is not dosing CO_2 into the greenhouse (Listing 86.4).

```

12 public boolean accept(Program p) {
13     boolean dosing = p.CO2Input < p.CO2Setpoint;
14     boolean ventilating = p.ventilationTempSetpoint
15         < p.indoorTempInput;
16     return dosing ? !ventilating : true;
17 }

```

Listing 86.4 Specification S_{wCO2} to avoid wasteful CO_2 interaction

It is the task of the arbitrator to find programs that entails the specifications using a genetic search algorithm [19]. That is, programs that satisfies the entailment $P, C \vdash S$. To find such programs, the arbitrator executes a genetic search algorithm for every control cycle of the control system. A short summary of genetic algorithm can be captured by following steps: (1) The genetic search algorithm generates a random population set of candidate programs P , (2) The satisfiability of each candidate program in the population P is evaluated against each specification. The returned results of the method calls are used by the

arbitrator to calculate the fitness of each proposed candidate program as the sum of satisfiability values. An accept count as a satisfiability value of one and a not accept count as a value of one, (3) The genetic algorithm sorts the set of candidate program according to best fitness. The worse half of the population set is then replaced by programs that are either a product of crossover or mutation of randomly selected candidate programs from the best half of the population, (4) Steps 2–3 is repeated over again til the process terminates after a given number of iterations, and (5) The program with the best fitness is selected as the solution program that is effectuated by the climate computer.

If the arbitrator cannot find a program that satisfies all specifications because of conflicting requirements, it provides an explanation of the conflicts as a context sharing problem.

86.6 Discussion

We have chosen to implement the arbitrator using a genetic algorithm, since control programs consist of setpoints that can be manipulated by crossover and mutation operators. In other domains it might be more difficult to apply the same approach without reimplementing the algorithm for that specific domain. The final argument for using a genetic algorithm is the fact that each specification will use models described by non-linear fitness functions, e.g., models for weather forecast ect. Genetic algorithms are known for their ability to handle such complex non-linear fitness landscapes [20].

Needless to say, a genetic algorithm for our problem may be configured in various ways. The current configuration of the algorithm is found by executing a set of test cases confirming that solutions will be found within acceptable time. That is not to say that there is no room for improvement. The performance of the genetic algorithm could definitely be fine-tuned. For example, the parameters could be dynamically adjusted during execution. Furthermore, the starting population could be a set of historical best solutions instead of a random population. Additionally, the algorithm could start out with big population and downsize it dynamically after some generations.

Finally, the approach needs to be more thoroughly tested. It is obvious to compare the approach with other proposed genetic algorithms to evaluate how well the approach performs. An other aspect that should be evaluated, is how well the approach supports explanation compared to other approaches.

86.7 Conclusion

This paper set out to show how requirements in natural language can be described in independently developed feature specifications, how feature interactions can be identified by analysing resources in the system environment and finally how the

control programs can be automatically adapted using a runtime arbitrator to avoid feature interactions when specifications are composed. Through Jackson's framework, we have found that feature interactions can be identified using context diagrams that includes the physical phenomena present in the environment; e.g., air temperature, CO₂ etc. Feature interactions emerge when independent feature shares phenomena indirectly through the same controlled environment. Furthermore, we have showed by example, how requirements in natural language can be mapped, in a one to one relationship, to accept and satisfy method specifications. Finally, we have presented an arbitrator that uses a genetic algorithm to find adapted programs that satisfies the accept and satisfy method specifications and thereby the corresponding requirements. These findings are important contributions to support independent development of control systems as the feature interaction problem is a hindrance to independent extensibility of such systems. While we have specifically focused on the feature interaction problem in control systems, the pervasiveness of the feature interaction problem implies that our findings are likely to be of relevance to other system domains. In terms of future research, we particularly suggest improvement of explanation of feature interactions by analysing the resources in the system environment.

References

1. Szyperski C (1996) Independently extensible systems—software engineering potential and challenges. In: In proceedings of the 19th Australasian computer science conference, 1996
2. Weck W (1997) Independently Extensible Component Frameworks. In: Special issues in object-oriented programming, Workshop reader of the 10th ECOOP' 96, dpunkt, verlag, Heidelberg, pp 177–183
3. Armstrong N, Robin L, Bashar N (2009) Feature interaction as a context sharing problem. In: Feature interactions in software and communication systems X
4. Bisbal J, Cheng BHC (2004) Resource-based approach to feature interaction in adaptive software. In: WOSS '04: proceedings of the 1st ACM SIGSOFT workshop on self-managed systems, New York, NY, USA, pp 23–27
5. Metzger A, (2004) Feature interactions in embedded control systems. *Comput Netw* 45(5):625–644
6. Calder M, Kolberg M, Magill EH, Marganiec SR (2003) Feature interaction: a critical review and considered forecast. *Comput Netw* 41(1):115–141
7. Zambrano A, Vera T, Gordillo SE (2006) Solving aspectual semantic conflicts in resource aware systems. In: RAM-SE, 2006, pp 79–88
8. Liu Y, Meier R (2009) Resource-aware contracts for addressing feature interaction in dynamic adaptive systems. In: 2009 Fifth international conference on autonomic and autonomous systems, vol 0. IEEE, Los Alamitos, pp 346–350
9. Jackson M (1995) Software requirements and specifications: a Lexicon of practice, principles and prejudices, 1st edn. Addison-Wesley Professional, Reading
10. Jackson M (2001) Problem frames: analyzing and structuring software development problems. Addison-Wesley Longman Publishing Co., Inc., Boston
11. Jackson M (2005) Problem frames and software engineering. *Inf Softw Technol* 47(14): 903–912

12. Jackson M (2007) The problem frames approach to software engineering. In: The 14th Asia-Pacific software engineering conference, Dec 2007, p14
13. Gamma E, Helm R, Johnson R, Vlissides J (1994) Design patterns: elements of reusable object-oriented software, 1st edn. Addison-Wesley Professional, Reading
14. Godskesen JC (1995) A formal framework for feature interaction with emphasis on testing. In: Feature interactions in telecommunications systems III. IOS Press, Amsterdam, pp 21–30
15. Aaslyng JM, Ehler N, Karlsen P, Rosenqvist E (1999) IntelliGrow: a component-based climate control system for decreasing the greenhouse energy consumption. *Acta Hort* 507:35–41
16. Aaslyng J, Lund J, Ehler N, Rosenqvist E (2003) IntelliGrow: a greenhouse component-based climate control system. *Environ Model Softw* 18(7):657–666
17. Aaslyng JM, Jakobsen L, Ehler N (2005) Climate control software integration with a greenhouse environmental control computer. *Environ Model Softw* 20:521–527
18. Gijzen H (1992) Simulation of photosynthesis and dry matter production of greenhouse crops. Simulation reports CABO-TT. DLO Centre for Agrobiological Research [and] DLO Winand Staring Centre for Integrated Land, Soil and Water Research, Wageningen Agricultural University, 1992
19. Sørensen JC, Jørgensen BN, Klein M, Demazeau Y (2011) An agent-based extensible climate control system for sustainable greenhouse production. In: The 14th international conference on principles and practice of multi-agent systems, 2011
20. Mitchell M (1998) An introduction to genetic algorithms (complex adaptive systems), 3rd edn. A Bradford Book, Cambridge

Chapter 87

Properties Evaluation of an Approach Based on Probability-Possibility Transformation

M. Pota, M. Esposito and G. De Pietro

Abstract The recent research on classification problems, in fields where vague concepts have to be considered, agree on the utility of fuzzy logic. An important step of inference engines preparation is the definition of fuzzy sets. When probability distributions of concerned variables are known, they can be used to define fuzzy sets, and different methods allow to perform this transformation. A method recently proposed by authors is compared here with other existing methods, in terms of assumptions and properties about the obtained fuzzy set, also considered with respect to the probability distribution it was calculated from. The best existing transformation in terms of compromise between consistency and specificity results to be a particular case of the proposed transformation, which can therefore be considered a more general method. Moreover, it enables, with a small loss of consistency, to find more interpretable fuzzy sets, while the case of less specific fuzzy sets is comprised and justified.

M. Pota (✉) · M. Esposito · G. De Pietro
Institute for High Performance Computing and Networking, ICAR-CNR,
Via P. Castellino 111, Naples 80131, Italy
e-mail: marco.pota@na.icar.cnr.it

M. Esposito
e-mail: massimo.esposito@na.icar.cnr.it

G. De Pietro
e-mail: giuseppe.depietro@na.icar.cnr.it

87.1 Introduction

In many fields, decision-making represents a very challenging issue to be handled. Classifying data in a finite number of conclusions is one of the main objectives of the recent research efforts, especially in medicine, economy and automated processes control. The classification of data is an operation which consists in determining the membership of a new data item to a specific class.

Data acquired from actual cases constitute experience which can be used to classify new incoming cases: namely, domain knowledge has previously been modeled using a set of data items, whose membership to a class is known. Data can be used in knowledge-based Decision Support Systems (DSSs), where they are typically modeled and processed by exploiting a rule representation formalism; they can be also used in data-driven DSSs, where serve as training set for statistical and machine learning models. Therefore, the application of computational intelligence approaches represents a general strategy to reason and learn about unknown relations, in presence of uncertainties and vagueness. A number of applications of both knowledge-based and data-driven DSSs exists, consisting in various modules of computational intelligence and hybrid combination of them. Examples are artificial neural networks [1, 2], fuzzy logic and fuzzy clustering [3, 4], hybrid combinations of artificial neural networks and fuzzy logic [5], and genetic-fuzzy algorithms [6–8], which have been successfully proposed in the literature. Furthermore, some authors have successfully investigated the application of fuzzy clustering by means of statistical analysis [9–11]. All these pieces of research agree on the utility of fuzzy logic [12] in the context of DSSs to manage the uncertainty and vagueness that are typical, for example, in the clinical decision processes.

The main strength point of fuzzy logic relies on the transparency and comprehensibility of its knowledge base. These properties are considered very attractive, in order to allow the resulting rules and membership functions to be studied and interpreted by a medical expert for being further improved or adapted according to experimental data.

Fuzzy DSSs require the definition of fuzzy sets, which could be known a priori, or roughly specified by decision tree algorithms [13]. Other approaches involve statistical analysis and clustering of data [9–11]. Moreover, the membership functions could be optimized in order to maximize the “goodness” of the DSSs [13]. However, the optimization of fuzzy sets undoubtedly brings to various problems, related to the prior definition of the sets’ shape, the risk of overfitting, and/or the definition of sets whose linguistic label is not in agreement with actual ranges. A different approach is then required to face these challenging open issues.

An appealing solution relies on the interpretation of input data by using probability distributions or likelihood functions, in order to successively exploit these kind of information to generate fuzzy sets. As a matter of fact, on the one hand, statistical approaches for data clustering could be used to define probability distributions or likelihood functions; on the other hand, especially in medical field,

knowledge could be promptly acquired in the form of probability distributions or likelihood functions. Moreover, since final users, like physicians in medical settings, are used to think and work according to a statistical interpretation of knowledge, the definition of fuzzy sets starting from statistical data is thought to be able to significantly reduce the existing lack of familiarity, shown by physicians in thinking in a fuzzy fashion, with respect to the classical statistical interpretation.

The problem to solve in this ambit, explained in more detail in the following, is how, given a category ' F ' and a set X of x values for a random variable, a probability distribution of x (which is the function $D(x) = p(x|F)$ defined on X) can be transformed into a fuzzy set F . The problem is well-known in literature [14], and thus different methods could be applied to solve this issue. Usually, the probability distribution is translated into a possibility distribution (the function $\Pi(x|F)$ defined on X), where the difference between the two expressions reproduces the difference between uncertainty and vagueness. The meaning of the membership function of the resulting fuzzy set, which coincides with a possibility distribution, is called "random set view" [15], and can be described as follows: the membership grade $\mu_F(x)$ is the degree of truth associated to the statement.

If the fuzzy set is F , then the variable value is x .

In this framework, possibility is considered to be an upper limit for the probability [14].

The problem of transforming a likelihood function (the function $L(x) = P(F|x)$ defined on X) into a fuzzy set has been also treated in literature. One method allows to obtain a fuzzy set [14]; the associated membership function, which could coincide with the set of possibilities $\Pi(F|x)$, is intended according to the so-called "likelihood view", which is described as follows: the membership grade $\mu_F(x)$ is the degree of truth associated to the statement.

If the variable value is x , then the fuzzy set is F .

This interpretation of a fuzzy set is typically used into fuzzy DSSs, where new data are given in the form of numeric values and have to be translated into fuzzy items.

The two probability-oriented views (upper probability and likelihood) of fuzzy sets and possibility distributions are not antagonistic [14].

If a probability distribution is available, the construction of a fuzzy set has been widely studied, and inherent literature will be detailed in the following. In particular, the method proposed by authors in [11] was considered here, in order to assess if principles of probability-possibility transformations are satisfied. Moreover, a comparative study among existing methods, including the one proposed in [11], is presented, by noticing which of different transformations satisfies different properties, and by applying them to case-study probability distributions associated with different types of probability density function (PDF).

The rest of the paper is organized as follows. Section 87.2 resumes the properties of probability-possibility transformations. In Sect. 87.3 the proposed approach is reminded, while in Sect. 87.4 it is evaluated with reference to its

properties and compared with existing methods, also by the comparative application of methods to synthesized PDFs. Finally, Sect. 87.5 concludes the work.

87.2 Properties of Probability-Possibility Transformations

The transformations between functions describing probability and possibility have been widely studied, starting from the birth of possibility theory developed by Zadeh [16].

Several works start from a probability distribution $D(x) = p(x|F)$, defined on X , where x varies in a subset X of the universe of discourse U and ‘ F ’ is a specified class of data. In the following, $p(x)$ is written instead of $p(x|F)$ since only one class of data is considered at a time.

Most of the methods essentially transforms $D(x)$ into a possibility distribution $\pi : X \rightarrow [0, 1]$, namely the fuzzy set F , which allows the measurement of the possibility $\Pi(A)$ of any finite subset A of X . Distinct solutions have been found, each of them based on a choice of some of the following assumptions:

(A) Normalization. The possibility distribution has to be normalized [17] to ensure $\Pi(X) = 1$. Therefore:

$$\exists x \in X | \pi(x) = 1. \quad (87.1)$$

(B) Consistency. Probability-possibility consistency has to be held [16]. The possibility measures can encode upper probabilities, therefore possibility degrees cannot be less than degrees of probability:

$$\forall A, P(A) \leq \Pi(A). \quad (87.2)$$

A similar concept leads to **consistency index maximization**. Halfway grades of consistency can be encoded by the consistency index:

$$\gamma = \sum_i \pi(x_i) \cdot p(x_i) \quad (87.3)$$

$$\gamma = \int_{x \in X} \pi(x) \cdot PDF(x) dx \quad (87.3')$$

(for discrete and continuous variables, respectively) which has to be maximized [18]. Maximal consistency corresponds to $\gamma = 1$ and means

$$\forall x | p(x) > 0, \pi(x) = 1. \quad (87.4)$$

Consistency furnishes each possibility degree with a lower bound. Moreover, maximizing the consistency index implies to maximize possibility degrees, thus losing information. Conversely, following assumptions C, D and E tend to minimize possibilities.

(C) Specificity. In order to preserve as much information as possible, when the transformation is performed, specificity has to be maximized. In other words, cardinality

$$card = \sum_i \pi(x_i) \quad (87.5)$$

$$card = \int_{x \in X} \pi(x) dx \quad (87.5')$$

(for discrete and continuous systems, respectively) has to be minimized [17].

(D) Uncertainty invariance. Another approach to the same concept of preserving information [19, 20] implies that invariance has to be held between uncertainties encoded by probability and possibility distributions, in other words the entropy of probability distribution should equal the energy of the possibility distribution:

$$H(p) = E(\pi). \quad (87.6)$$

This assumption is debatable, because is based on the prerequisite that possibilistic and probabilistic information measures are commensurate.

(E) Equidistribution. Plausibility, as defined in evidence theory, can be approximated by probability [21]:

$$Pl(A) \cong P(A) \quad (87.7)$$

(F) Order preservation. Since the more probable one event is, the more possible it should be as well, preference has to be preserved [17]:

$$\pi(x_i) > \pi(x_j) \Leftrightarrow p(x_i) > p(x_j). \quad (87.8)$$

One could refer to weak order-preservation if only p order implies π order.

(G) Scaling. A scaling assumption [18, 22, 23] forces each possibility value $\pi(x_i)$ to be a function of only the probability $p(x_i)$ of the same event:

$$\pi(x) = f(p(x)). \quad (87.9)$$

The function f can be ratio-scale, Log-interval scale, etc. However, such assumption can lead to not consistent transformations [24].

87.3 The Proposed Approach

The idea the proposed method is based on, is that the statistical test of hypothesis can be considered as an opinion, just like the answer given by somebody to the question

Is H_0 true?,

where H_0 is called null hypothesis.

Suppose F is a fuzzy set defined on a subset of X ; furthermore, suppose that the probability distribution $D(x)$, associated to x values for which the statement “ x is F ” is true, is known. Now, take a x value and formulate the null hypothesis

$H_0 \equiv “x \text{ is } F”$,

which is equivalent to the hypothesis

$H_0 \equiv “x \text{ is an occurrence of a random variable whose probability distribution is } D”$.

The hypothesis test is thus performed in order to decide if the last statement could be judged true. If the test is applied to a certain probability distribution and repeated a number of times, a polling is simulated. Clearly, the test should not be repeated ever identically, but different perceptions should be randomly simulated. Different perceptions are assumed to be related to different significance levels α used to perform the hypothesis test (significance levels correspond to the probability of rejecting the null hypothesis when it is true). Therefore, to simulate a polling, the test has to be repeated with different α values, to be chosen randomly, according to a certain probability distribution A . The type of distribution for α is unknown and may be considered a degree of freedom associated with the method. The probability distribution A is normalized between two limiting values α_{\min} (≥ 0) and α_{\max} (≤ 1):

$$\begin{cases} \int_{\alpha_{\min}}^{\alpha_{\max}} PDF_A(\alpha) d\alpha = 1 \\ \forall \alpha \notin [\alpha_{\min}, \alpha_{\max}], PDF_A(\alpha) = 0 \end{cases} \quad (87.10)$$

The type of distribution and boundaries may be chosen in order to let the resulting fuzzy set have desired properties, as trapezoidal-like shape, triangular-like shape, little overlapping between different fuzzy sets and so on. The knowledge (or ignorance) about the alternatives to the null hypothesis, induces to perform, in different cases depending on H_I , a one-tailed (on left or right side) or two-tailed test. For each x value, the test is performed by computing the p value associated to x once distribution D is assumed:

$$p - value(\bar{x}) = \sum_{x \geq \bar{x}} p(x) Z \quad (87.11)$$

$$p - value(\bar{x}) = \int_{\bar{x}}^{\sup(X)} PDF_D(x) dx \quad (87.11')$$

on the right side, or

$$p - value(\bar{x}) = \sum_{x \leq \bar{x}} p(x) \quad (87.12)$$

$$p - value(\bar{x}) = \int_{\inf(X)}^{\bar{x}} PDF_D(x) dx \quad (87.12)$$

on the left side, where (87.11) and (87.12) refer to a discrete distribution, whereas in (87.11') and (87.12') PDF_D stands for probability density function of the continuous distribution D .

Then, p value is compared to α value; if

$$p - value(x) > \frac{\alpha}{t}, \quad (87.13)$$

where $t = 1$ for one-tailed tests and $t = 2$ for two-tailed tests, then a “positive answer” is given. Note that “positive answer” is intended to mean “ H_0 is possible”.

This gives an answer to the starting question, and can be encoded as follows:

$$\pi(H_0)(x, \alpha) = \begin{cases} 1 & \text{if } \left[p - value(x) > \frac{\alpha}{t} \right] \\ 0 & \text{if } \left[p - value(x) \leq \frac{\alpha}{t} \right] \end{cases} \quad (87.14)$$

where $\pi(H_0)$ is the perception-based possibility of H_0 , namely $\pi(F|x)$, depending on x and α . Crisp values are produced because only one perception is involved.

The test is repeated for the entire α range, and the probabilities of α values, which give positive answer to the test, are summed up:

$$\mu_F(x) = \int_{\alpha_{\min}}^{\alpha_{\max}} \pi(H_0)(x, \alpha) \cdot PDF_A(\alpha) d\alpha \quad (87.15)$$

The repetition of the test for the same value x is actually a polling, thus, in eq. (87.15), the membership grade $\mu_F(x)$ of the fuzzy set F is computed according to [15].

The membership function of fuzzy set F is thus obtained by considering the set of $\mu_F(x)$ membership grades for all of x values.

87.4 Comparison of Transformations Based on Properties Evaluation

The most notorious transformations could be summarized by respective formulas, given below. Discrete probability distributions are indexed in such a way that

$$p(x_1) \geq p(x_2) \geq \dots \geq p(x_n). \quad (87.16)$$

Modal value of continuous distributions is indicated by x^m , median by x^{me} .

1.

$$\pi(x_i) = \sum_{k=1}^n \min[p(x_i), p(x_k)]. \quad (87.17)$$

$$\pi(\bar{x}) = \int_{x \in X} \min[PDF(x), PDF(\bar{x})] dx \quad (87.17')$$

This was firstly proposed by Dubois and Prade [18], is invertible and is based on normalization A), consistency B) and order preservation F). It does not meet maximum specificity. Equivalent results were achieved by Yager and Kreinovich [19], who found a transformation based on uncertainty invariance D), and by Yamada [25], who based his results on evidence theory assumption E).

2.

$$\pi(x_i) = \sum_{k=i}^n p(x_k), \quad (87.18)$$

$$\pi(\bar{x}) = \int_{x|PDF(x) \leq PDF(\bar{x})} PDF(x) dx \quad (87.18')$$

This was developed by Dubois et al. [17], and is based on normalization A), consistency B), order preservation F), and additionally on maximum specificity, C).

3.

$$\pi(x_i) = \left(\frac{p(x_i)}{p(x_1)} \right)^a, \quad (87.19)$$

$$\pi(\bar{x}) = \left(\frac{PDF(\bar{x})}{PDF(x^m)} \right)^a. \quad (87.19')$$

This was proposed by Klir [22] and is based on uncertainty invariance D) and scaling assumption G). The exponent depends on the distribution and can be computed by imposing condition D).

Of course, (87.17), (87.18) and (87.19) refer to discrete probability distributions, while (87.17'), (87.18') and (87.19') to continuous case.

4. Some works start from a probability distribution to reach a fuzzy set by statistical methods. Some of them use particular assumptions [9, 10], while a method was given in our previous work [11], where a generalization of these

existing methods is formulated, and theoretical support is proposed so as to justify these kinds of transformation. Here, the properties of the proposed method are evaluated.

Since $\alpha_{\max} \leq 1$ and the probability distribution D is normalized (therefore, p -values range from 0 to 1), then there exist at least one x value such that

$$p - \text{value}(x) \geq \alpha_{\max}. \quad (87.20)$$

Therefore, the hypothesis test at that (or those) point(s) gives positive answers for all the α values, whose sum of probabilities equals 1; therefore,

$$\mu_F(x) \in [0, 1], \quad (87.21)$$

namely the resulting fuzzy set is normal (property A) is satisfied). If a continuous PDF is concerned, the limiting value of $\alpha_{\max} = 1$ gives a normal fuzzy set, as well.

Now, the condensed formula is given to calculate $\mu_F(x)$ through the proposed algorithm applied on both sides (by two tails hypothesis test):

$$\mu(\bar{x}) = \int_{\alpha_{\min}}^{\min \left\{ x \left| \int_{x_{me}}^x PDF_D(x) dx \right| \geq \left| \int_{x_{me}}^{\bar{x}} PDF_D(x) dx \right| \right\}} PDF_A(\alpha) d\alpha \quad (87.22)$$

If a uniform distribution is chosen for α , it becomes:

$$\mu(\bar{x}) = \frac{\min \left\{ x \left| \int_{x_{me}}^x PDF_D(x) dx \right| \geq \left| \int_{x_{me}}^{\bar{x}} PDF_D(x) dx \right| \right\} - \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}} \quad (87.23)$$

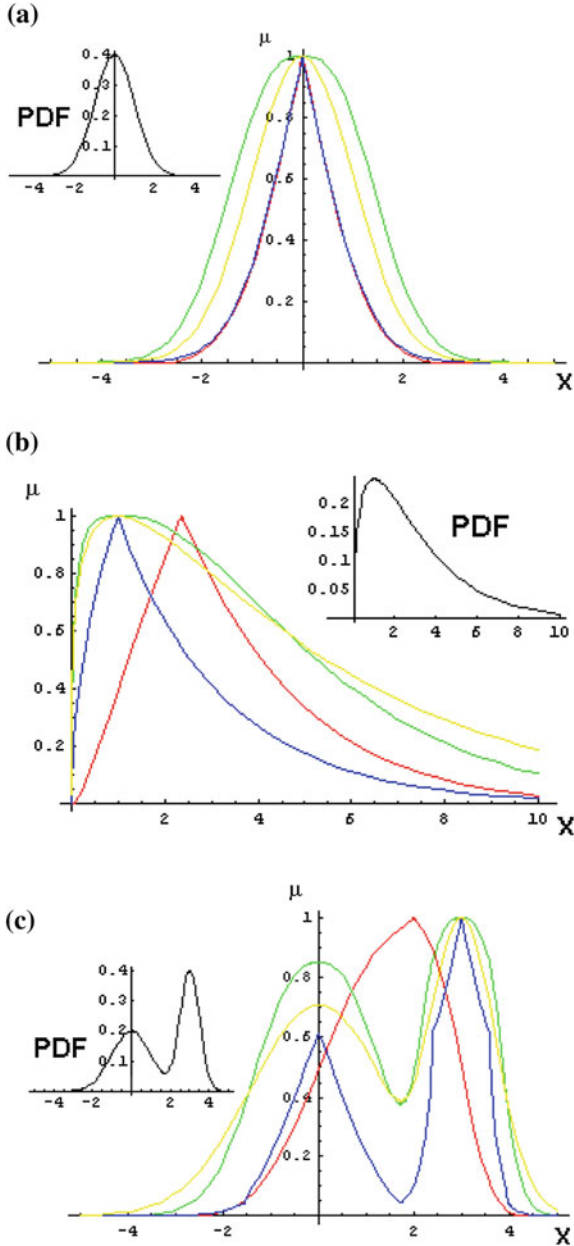
and if $\alpha_{\min} = 0$ and $\alpha_{\max} = 1$:

$$\mu(\bar{x}) = \int_{x \left| \int_{x_{me}}^x PDF_D(x) dx \right| \geq \left| \int_{x_{me}}^{\bar{x}} PDF_D(x) dx \right|} PDF_D(x) dx. \quad (87.24)$$

In this case, if applied to unimodal, symmetric and non-constant probability distributions, this method gives the same result of method 2). Therefore, the transformation satisfies consistency and order preservation, and the resulting fuzzy set is the most specific. Hence, properties A, B, C and F are satisfied.

If α_{\min} is greater than 0, the result is lower than that in transformation 2), which is the maximally specific, thus giving a transformation which does not fully satisfy consistency B. However, if α_{\min} is little, the difference between the two calculated fuzzy sets is very little (less than α_{\min}), while a great improvement is given to the

Fig. 87.1 Three Probability Density Functions:
a Standard Gaussian, **b** χ^2 distribution with 3 degrees of freedom, **c** Convolution of two Gaussians $N[0, 1]$ and $N[3, 0.5]$. Four transformations are applied to construct fuzzy sets: Fuzzy set 1 (in green) uses method proposed by Dubois and Prade [18], Yager and Kreinovich [19] and Yamada [25]; Fuzzy set 2 (in blue) uses method by Dubois et al. [17]; Fuzzy set 3 (in yellow) uses method by Klir [20]; Fuzzy set 4 (in red) uses the method proposed by authors [11], with α uniformly distributed between 0.01 and 1



simplicity of the system, because the new fuzzy set has bounded support. At the same time, conditions of normality A, maximal specificity C and order preservation F are still satisfied.

In Fig. 87.1, the three transformations (87.17'), (87.18') and (87.19'), and the transformation (87.23) were applied to three case-study probability distributions. Resulting fuzzy sets were denoted by numbers corresponding to the numbers associated in this section to the respective methods.

In Fig. 87.1a, the probability distribution is associated with a Standard Gaussian PDF_D ($x^m = 0, \sigma = 1$). The resulting fuzzy sets are shown to be normal, with order ever preserved. Moreover, all fuzzy sets seem to be consistent with probabilities; finally, fuzzy sets 2 and 4 seem absolutely identical, both reaching maximal specificity. Here, for the proposed transformation, a value was chosen of $\alpha_{\min} = 0.01$. Therefore, the difference between the two fuzzy sets is ever smaller than 0.01, which is irrelevant for practical scopes. On the other hand, the support of fuzzy set 4 is bounded, while fuzzy set 2 is unbounded, which let fuzzy set 4 be more suitable to be used for DSSs calculations, because it is more interpretable and does not need to be truncated at arbitrary points.

If α_{\max} is lower than 1, a trapezoidal shape fuzzy set is obtained by this method, which is no more maximally specific.

In case α is supposed to be not uniformly distributed, increasing PDFs bring to not consistent transformations, while decreasing functions for α distribution bring to not maximally specific fuzzy sets.

If a probability distribution (of x) with constant values is concerned, this method furnishes one of the most specific fuzzy sets, which can be chosen by ordering in eq. (87.16) equiprobable events in different manners [17]. In particular, it is the one which assigns the greater possibility to events closer to the median with respect to the other equiprobable events. Also in this case, properties A, B, C and F are satisfied.

If the distribution is not symmetric, the shape of fuzzy set obtained by this method is slightly different from that of method (2). In particular, the maximum possibility is reached in correspondence of median value, while with all other methods the maximum is in correspondence of modal value. Moreover, the presented method gives the same possibility to events which are equally "far from the distribution", not to equiprobable events. Therefore, in some regions the fuzzy set is not maximally specific, while in other regions the transformation results not consistent.

In Fig. 87.1b, all transformations were applied to a χ^2 -distributed (with 3 degrees of freedom) random variable. Here it is shown that if a not symmetric distribution is involved, method (4) gives a fuzzy set quantitatively different from others. Transformation (2) gives (ever) the most specific fuzzy set.

If a distribution with more than one mode is considered, this algorithm gives a fuzzy set qualitatively different from the others. The method assigns a greater possibility to x values which are closer to the median, even if they are less probable. The transformation satisfies this principle rather than preference preservation (F). Also maximal specificity (C) is not reached in this case, while normality (A) remains satisfied, and consistency is the same of other cases.

In Fig. 87.1c, the application of the methods to a PDF obtained by convolution of two Gaussian functions is shown. Again, all fuzzy sets are normal, fuzzy sets

(1, 2 and 3) are consistent and order-preserving, fuzzy set (2) is maximally specific. Fuzzy set 4 does not even show the double modes trend. The membership function is forced to be unimodal, and consequently more decipherable fuzzy sets are constructed.

87.5 Conclusions and Future Work

In this work, different methods used to transform probability distributions into fuzzy sets are evaluated. In particular, a recent method proposed by the authors is compared to the others.

Its application on symmetric unimodal distributions discloses the helpfulness of the proposed method, because, if the right choice of its parameters is done, obtained fuzzy sets are normal, consistent, order preserving and maximally specific. Moreover, little regulation of parameters allows to obtain fuzzy sets with bounded support, even if they are calculated from a random variable with unbounded probability distribution. This peculiarity is not reached by other methods, and is particularly attractive, because bounded fuzzy sets do not need arbitrary cut-offs, and result much more understandable to the final user.

Nevertheless, the application of the method on asymmetric and multimodal distributions reveals its different nature, since the maximum membership grade of the obtained fuzzy sets corresponds to the median, while other methods maximize membership in correspondence of modal values. On the other hand, the membership function is forced to be unimodal, and this characteristic provides the users with much more decipherable fuzzy sets.

Applicability of different methods on real data sets should be assessed in future, because if more than one class is involved, the construction of fuzzy sets should take it into account, and the usefulness of different methods can be measured by considering classification rates of respective DSSs.

References

1. Axer H, Jantzen J, Berks G, Keyserlingk DGV (2000) Aphasia classification using neural networks. In: Proceedings of European Symposium on Intelligent Techniques, Aachen, pp 111–115
2. Axer H, Jantzen J, Keyserlingk DGV (2000) An aphasia database on the internet: a model for computer assisted analysis in aphasiology. *Brain Lang* 75:390–398
3. Berks G, Keyserlingk DGV, Jantzen J, Dotoli M, Axer H (2000) Fuzzy clustering—a versatile mean to explore medical databases. In: Proceedings of European Symposium on Intelligent Techniques
4. Castellano G, Fanelli AM, Mencar C (2003) A fuzzy clustering approach for mining diagnostic rules. *IEEE International Conference on Systems, Man and Cybernetics*
5. Jantzen J, Axer H, Keyserlingk DGV (2000) Diagnosis of aphasia using neural and fuzzy techniques. In: Proceedings Symposium on Computational Intelligence and Learning

6. Tsakonas A et al (2004) Evolving rule-based systems in two medical domains using genetic programming. *Artif Intell Med* 32:195–216
7. Tsakonas A (2006) A comparison of classification accuracy of four genetic programming-evolved intelligent structures. *Inform. Sci.* 176:691–724
8. Esposito M, De Falco I, De Pietro G (2010) An evolutionary-fuzzy approach for supporting diagnosis and monitoring of Multiple Sclerosis. In: 5th Cairo International Biomedical Engineering Conference, Dec 2010, 108–111
9. Akbarzadeh-T M-R, Moshtagh-K M (2007) A hierarchical fuzzy rule-based approach for aphasia diagnosis. *J. Biomedical Inform* 40:465–475
10. Schuerz M, Adlassnig K-P, Lagor C, Scheider B, Grabner G (1999) http://www.erudit.de/erudit/events/esit99/12568_p.pdf
11. Pota M, Esposito M, De Pietro G (2011) Transformation of probability distribution into a fuzzy set interpretable with likelihood view. In: IEEE 11th International Conference on Hybrid Intelligent Systems (HIS 2011), Malacca, Malaysia (in press)
12. Zadeh L (1965) Fuzzy sets. *Inform. Control.* 8:338–353
13. d’Acerno A, De Pietro G, Esposito M (2010) Data driven generation of fuzzy systems: an application to breast cancer detection. In: *Proceedings of CIBB*
14. Dubois D, Prade H (1993) Fuzzy sets and probability: Misunderstandings, bridges and gaps. In: *Second IEEE International Conference on Fuzzy Systems*, San Francisco 2:1059–1068
15. Bilgic T, Türksen IB Measurement of membership functions: Theoretical and empirical work. In: Dubois D, Prade H (eds) *Handbook of fuzzy sets and systems vol 1, Fundamentals of fuzzy sets*, Kluwer, Dordrecht 195–232
16. Zadeh L (1978) Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Syst* 1:3–28
17. Dubois D, Foulloy L, Mauris G, Prade H (2004) Probability-Possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable Comput* 10:273–297
18. Dubois D, Prade H (1986) Fuzzy sets and statistical data. *Eur J Oper Res* 25:345–356
19. Yager R, Kreinovich V (2004) Entropy conserving probability transforms and the entailment principle, Technical Report, MII-2518. Machine Intelligence Institute, Iona College
20. Klir GJ (1990) A principle of uncertainty and information invariance. *Int. Journal of General Systems* 17:249–275
21. Dubois D, Prade H (1982) On several representations of uncertain body of evidence. In: *Fuzzy Informatics and Decision Processes*, Gupta M.M, Sanchez E (eds), North-Holland Pub., Amsterdam 167–181
22. Geer JF, Klir G (1992) A mathematical analysis of information-preserving transformations between probabilistic and possibilistic formulations of uncertainty. *Int J Gen Syst* 20:361–377
23. Shafer G (1976) *A mathematical theory of evidence*. Princeton University Press, Princeton
24. Dubois D, Prade H (1980) *Fuzzy sets and systems: theory and applicatios*. Academic Press, New York
25. Yamada K (2001) Probability-Possibility transformation based on evidence theory. In: *IFSA World Congress and 20th NAIPS International Conference*, 2001, Joint 9th, 70–75

Chapter 88

Functional Verification of Class Invariants in CleanJava

Carmen Avila and Yoonsik Cheon

Abstract In Cleanroom-style functional program verification, a program is viewed as a mathematical function from one program state to another, and the program is verified by comparing two functions, the implemented and the expected behaviors. The technique requires a minimal mathematical background and supports forward reasoning, but it does not support assertions such as class invariants. However, class invariants are not only a practical programming tool but also play a key role in the correctness proof of a program by specifying conditions and constraints that an object has to satisfy and thus defining valid states of the object. We suggest a way to integrate the notion of class invariants in functional program verification by using CleanJava as a specification notation and a verification framework as well; CleanJava is a formal annotation language for Java to support Cleanroom-style functional program verification. We propose a small extension to CleanJava to specify class invariants and to its proof logic to verify the class invariants. Our extension closely reflects the way programmers specify and reason about the correctness of a program informally. It allows one to use class invariants in the framework of Cleanroom-style functional specification and verification.

C. Avila (✉) · Y. Cheon

Department of Computer Science, The University of Texas at El Paso,
El Paso, TX, USA
e-mail: ceavila3@miners.utep.edu

Y. Cheon
e-mail: ycheon@utep.edu

88.1 Introduction

An assertion is a predicate or boolean expression, placed in a program, that should be always true at that place [1]. Assertions such as class invariants and operation pre- and post-conditions became popular as a practical programming tool for verifying, testing and debugging programs [2]. If an assertion evaluates to false at runtime, it indicates that there is an error in the code for that particular execution, thus an assertion can be used for runtime verification of code and for narrowing down a problematic part of the code. Assertions also play a key role in verifying statically the correctness of a program [3]. A class invariant, for example, specifies a condition that all objects of a class must satisfy while they can be observable by clients. It defines valid states of an object and ensures that an object remains in a consistent state. It must be proved that all methods of the class preserve the class invariant.

A functional program verification technique such as Cleanroom [4] views a program as a mathematical function from one program state to another and proves its correctness by essentially comparing two functions, the function computed by the program and its specification [5]. Since the technique uses equational reasoning based on sets and functions, it requires a minimal mathematical background. Unlike Hoare logic [1], it also supports forward reasoning and thus reflects the way programmers reason about the correctness of a program informally. There is a formal notation to support Cleanroom-style functional program verification. CleanJava is such a formal annotation language for the Java programming language [6]. In CleanJava, a specification function is written using a subset of Java expressions enriched with CleanJava-specific extensions, and every section of Java code is annotated with its expected behavior for formal verification of the correctness of the code (see Sect. 88.2).

One problem of a functional program verification technique, however, is that it does not work well with assertions, especially with class invariants. In fact, CleanJava does not provide any built-in language construct to express class invariants. This poses a serious problem both in writing a specification and using it for a correctness proof. In CleanJava, for example, the behavior of a method is specified as a mathematical function, and thus a class invariant must be expressed in a functional form and merged to the specification of each method of the class. The resulting specifications become less readable, reusable, and maintainable, and the correctness verification is not modular in that it cannot be decomposed into those of an invariant property and a method-specific property.

In this paper we propose a way to integrate the notion of class invariants in the functional program verification by using CleanJava as a platform for our study. We suggest two approaches: an *invariant function* and an *invariant clause*. In the first approach, a user-defined function is introduced to test a class invariant. This invariant function is referred to in the specification of each method of a class. The second approach supports an invariant as a built-in language feature by extending CleanJava and its proof logic. It adds a special clause to express an invariant of a

class and extends the proof logic to ensure that the specified invariant be established by constructors and preserved by all methods of the class. Although this approach requires a language extension, it provides a better solution by cleanly separating the specification and verification of an invariant from those of methods.

Since invariants are a well-known concept, it is not surprising to find existing work on using invariants in a Cleanroom-style verification [5]. However, the topic's treatment is shallow in an informal setting without giving a systematic way of translating an invariant or a formal treatment of its proof rules.

The main contribution of our work is that it enables one to use class invariants in the framework of a Cleanroom-style functional specification and verification technique and thus makes the technique more closely resemble the way programmers specify and reason about the correctness of a program informally. We expect this to have a positive effect on teaching and practicing the functional program verification.

The rest of this paper is structured as follows. In Sect. 88.2 we give a quick overview of CleanJava and functional program verification. In Sect. 88.3 we illustrate the problem of the functional verification not supporting class invariants. In Sect. 88.4 we describe our two approaches for integrating the notion of invariants in a functional verification technique, followed by a comparison of these approaches. In Sect. 88.5 we provide a concluding remark along with future work.

88.2 Background: CleanJava

CleanJava is a formal annotation language for the Java programming language to support a Cleanroom-style functional program verification [6]. In the functional program verification, a program is viewed as a mathematical function from one program state to another. In essence, functional verification involves calculating the function computed by code, called a *code function*, and comparing it with the intention of the code written as a function, called an *intended function* [5]. CleanJava provides a notation for writing intended functions. A *concurrent assignment* notation, $[x_1, x_2, \dots, x_n := e_1, e_2, \dots, e_n]$, is used to express these functions by only stating changes that happen. It states that x_i 's new value is e_i , evaluated concurrently in the initial state—the state just before executing the code; the value of a state variable that does not appear in the left-hand side remains the same. For example, $[x, y := y, x]$ is a function that swaps two variables x and y .

Figure 88.1 shows sample Java code annotated with intended functions written in CleanJava. It describes an `AddressBook` class containing a collection of contacts. A CleanJava annotation is written in a special kind of comments either preceded by `//@` or enclosed in `/*@ . . . @*/`, and an intended function is written in the Java expression syntax with a few CleanJava-specific extensions. The first annotation states that the constructor initializes the `db` field to a new empty list. The intended function of the `hasContact` method is interesting. It specifies a partial function defined only when the argument (`n`) is not null; as shown, a

```

class AddressBook {
    private List<Contact> db;

    //@ [db := new ArrayList<Contact>()]
    public AddressBook() {
        db = new ArrayList<Contact>();
    }

    /*@ f0:[n != null ->
        result := db->exists(getName().equals(n))] @*/
    public boolean hasContact(String n) {
        //@ f1:[r, i := false, 0]
        boolean r = false;
        int i = 0;

        //@ f2:[r, i := r || b, anything] where
            boolean b = db.subList(i, db.size())
                ->exists(getName().equals(n)) @*/
        while (i < db.size()) {
            //@ [r, i := r || db.get(i).getName().equals(n), i++]
            if (contacts.get(i).getName().equals(n))
                r = true;
            i++;
        }

        //@ f3:[result := r]
        return r;
    }
}

```

Fig. 88.1 Sample CleanJava code

concurrent assignment may have an optional condition or guard followed by an arrow (\rightarrow) symbol. The function states that, given a non-null name (n), the method tests if there is a contact with the given name in db . The pseudo variable $result$ denotes the return value of a method, and $exists$ is a CleanJava iteration operator that tests if a collection contains at least one element that satisfies a given condition. The body of the method is also interesting. Each section of code is documented with its intended function. In the function f_2 , the keyword **anything** indicates that we do not care about the final value of the loop variable i , and a **where** clause introduces local definitions such as that of b .

It would be instructive to sketch a correctness proof of the `hasContact` method, which involves the following.

- Proof that the composition of functions f_1 , f_2 , and f_3 is correct with respect to (\sqsubseteq) , or a refinement of, f_0 , i.e., $f_1;f_2;f_3 \sqsubseteq f_0$, where $;$ denotes a functional composition.
- Proof that f_1 , f_2 , and f_3 are correctly refined.

In the functional verification, a proof is often trivial or straightforward because a code function can be easily calculated and directly compared with an intended function; e.g., f_1 and f_3 are both code and intended functions. However, one also need to use different techniques such as a case analysis and an induction based on

the structure of the code as in the proof of f_2 [6]. Below we discharge the first proof obligation, where b_i is $db.subList(i, db.size()) \rightarrow exists(getName().equals(n))$ and $?$ is short for **anything**.

$$\begin{aligned}
 f_1; f_2; f_3 &\equiv [r, i := false, 0]; [r, i := r || b_i, ?]; [result := r]; \\
 &\equiv [r, i := b_0, ?]; [result := r]; \\
 &\equiv [r, i, result := b_0, ?, b_0] \\
 &\sqsubseteq [result := b_0] \\
 &\equiv f_0
 \end{aligned}$$

88.3 The Problem

A functional program verification technique is fundamentally different from an assertion-based technique such as Hoare logic [1]. It is direct and constructive in that for each state variable such as a program variable one must state its final value explicitly. On the other hand, an assertion-based technique is indirect and constraint-based in that one specifies the condition that the final state has to satisfy by stating a relationship among state variables. The final value of a state variable is not defined directly but instead is constrained and given indirectly by the specified condition.

Because of this fundamental difference, a functional verification technique does not work very well with assertions such as class invariants. In fact, CleanJava does not provide a built-in language construct for specifying class invariants. This is a serious concern in practice because class invariants are a popular programming idiom and cannot be directly expressed in CleanJava. To illustrate this problem, let's consider the `AddressBook` class from the previous section. One possible class invariant for this class would be $db \neq null \ \&\& \ db \rightarrow isUnique(getName())$, stating the non-nullness of the `db` field and the uniqueness of contact names; the `isUnique` operator is a CleanJava iterator asserting the uniqueness of given values. How to express this invariant in CleanJava? An invariant must be merged to, and expressed in, the intended function of each operation of the class to ensure its establishment by a constructor and its preservation by each method, as shown below.

```

class AddressBook {
    private List<Contact> db;

    /*@ [db := new ArrayList<Contact>] @*/
    public AddressBook() { ... }

    /*@ [n != null && db != null && db->isUnique(getName()) ->
        result := db->exists(getName().equals(n))] @*/
    public boolean hasContact(String n) { ... }
}

```

Note that the constructor's intended function remains the same. This is because the new value for `db`, an empty list, obviously implies the invariant. For the `hasContact` method, the invariant becomes the optional condition part of the concurrent assignment and the rest are unchanged. This is because a method assumes an invariant, and this particular method does not change any state variable, meaning that the invariant is trivially preserved. For a mutation method, say `addContact`, the invariant must become the condition of its intended function and be implied by new values of state variables.

There are several shortcomings in the above approach of not explicitly stating a class invariant and scattering it all over method specifications. There are problems of specification readability, reusability, and maintainability. The specifications of an invariant property and the behavior of a method are tangled, and an invariant specification is duplicated in almost every method specification. The approach also makes a correctness verification hard and non-modular in that the verification of an invariant property and that of a method-specific property cannot be performed separately, as the specifications of these two properties are tangled and are not distinguished.

88.4 Our Approach

In this section we describe our approaches for supporting invariants. We propose two approaches: an invariant function and an invariant clause. The first approach allows one to systematically translate an invariant to CleanJava annotations without requiring an extension to CleanJava or its proof logic. On the other hand, the second approach does require an extension to both the notation and the proof logic of CleanJava, but it cleanly separate the specification and verification of class invariants from those of methods.

88.4.1 An Invariant Function

This approach is to express an invariant in the intended function of each method. An invariant becomes part of the intended function of a method and is verified along with the intended function. This approach is similar to the view that an invariant is conjoined to pre- and post-conditions of an operation. To eliminate a duplication of an invariant expression in multiple intended functions, we introduce a user-defined function that tests an invariant. This function is called an *invariant function* and is responsible for testing all the invariants of a class.

Suppose we have an invariant I written in terms of a state variable, say x , and a method with an intended function $[P \rightarrow y := E]$, where the type of y is T . Then, our approach produces the following user-defined function and intended functions.

```
fun inv(x) = I
f1: [inv(x) && P -> y := E]
f2: [inv(x) && P ->
    y := findAny(T z | z == E && inv(z))]
```

The first annotation introduces a user-defined function named `inv` that tests the invariant of a class. The state variables appearing in the invariant become the arguments of the invariant function so that the invariant can be tested in both the initial and the final states. The next two annotations show translated intended functions. Depending on whether a state variable appearing in the invariant is changed or not, either intended functions f_1 or f_2 is used. If x and y are different state variables—i.e., the state variable appearing in the invariant is not changed, the first one (f_1) is used; otherwise, the second one (f_2) is used. As expected, the invariant constrains the condition (P) and the final values of state variables (E). The CleanJava operator `findAny` denotes an arbitrary value that satisfies a given condition. In f_2 , the argument to the second `inv` call is z —the final value of y —because the expressions in concurrent assignments are evaluated in the initial state and the `inv` call is to check the invariant in the final state.

Let's apply this approach to our `AddressBook` class. The revised intended functions are shown below.

```
class AddressBook {
  private List<Contact> db;
  // @ fun inv(db) = db != null && db->isUnique(getName())

  /* @ [db := findAny(List<Contact> l | inv(l) &&
    l.equals(new ArrayList<Contact>()))] @ */
  public AddressBook() { ... }

  /* @ [n != null && inv(db) ->
    result := db->exists(getName().equals(n))] @ */
  public boolean hasContact(String n) { ... }
}
```

An invariant function is defined in the first annotation. In CleanJava, one does not have to declare the signature of a user defined function; it is inferred [6]. The constructor's intended function was translated using the f_2 pattern. However, since a constructor does not assume an invariant in the initial state, the invariant function does not appear in the optional condition part of the concurrent assignment. The intended function of the `hasContact` method is translated using the f_1 pattern.

88.4.2 An Invariant Clause

This approach is to support the notion of invariants as a built-in language feature of CleanJava. For this, we propose to introduce a new CleanJava language construct called an *invariant clause*. An invariant clause can appear only in the member declaration level and specifies the invariant of a class. It must be established by all constructors and preserved by all methods of the class. For example, shown below is the `AddressBook` class annotated using an invariant clause. Note that the intended functions of the constructor and the method are unchanged.

```
class AddressBook {
    private List<Contact> db;
    //@ inv: [db != null && db->isUnique(getName())]

    //@ [db := new ArrayList<Contact>()]
    public AddressBook() { ... }

    /*@ [n != null ->
        result := db->exists(getName().equals(n))] @*/
    public boolean hasContact(String n) { ... }
}
```

A natural next question is how to verify a class invariant specified using an invariant clause. We extend the proof rules of CleanJava to support the invariant clause. Consider a class with an invariant I specified using an invariant clause. For a constructor C with an intended function f in the form of $[P \rightarrow x := E]$, we have the following extended proof obligations.

- (1) C is correct with respect to f , i.e., $C \sqsubseteq f$.
- (2) C establishes I . For this, one needs to prove:
 - (a) $P \Rightarrow I$ if I is not written in terms of x , or
 - (b) $P \Rightarrow I[E/x]$ otherwise, where $I[E/x]$ means I with every free occurrence of x replaced with E .

For a method M with an intended function f in the form of $[P \rightarrow x := E]$, we have the following extended proof obligations.

- (1) M is correct with respect to f provided that I holds in the initial state, i.e., $M \sqsubseteq [P \ \&\& \ I \rightarrow x := E]$.
- (2) M preserves I . For this, one needs to prove:
 - (a) $I \wedge P \Rightarrow I$ if I is not written in terms of x , or
 - (b) $I \wedge P \Rightarrow I[E/x]$ otherwise, where $I[E/x]$ means I with every free occurrence of x replaced with E .

As an example, let's prove the invariant of the `AddressBook` class. For the constructor, we need to discharge the proof obligation 2.b: $P \Rightarrow I[E/x]$ because the constructor changes the state variable `db` appearing in the invariant. Note that the constructor does not have an optional condition (P), leaving as a proof obligation $I[E/x], db! = null \ \&\& \ db \rightarrow isUnique(getName())$ where `db` is `new ArrayList<Contact> .` The proof is straightforward because a new empty list is not null and contains no contact. For the `hasContact` method, we have to discharge the proof obligation 2.a: $I \wedge P \Rightarrow I$, as it does not change any state variable. However, there is nothing to prove; it's a tautology.

88.4.3 Comparison

The invariant function approach allows one to systematically translate class invariants to intended functions. Since invariants are factored out to user-defined functions, they are not duplicated in intended functions. The strength of this approach is that it does not require a language extension or the proof rules. However, it does not completely address the original problems of readability, reusability, maintainability, and verifiability. For example, specifications are still tangled and scattered, and the use of `findAny` operator in an intended function makes a specification complicate and hard to read and understand.

An invariant clause addresses all the aforementioned problems by cleanly separating an invariant specification from method specifications. It supports a separation of concerns in a verification; an invariant verification and a method verification can now be performed separately and in a modular way. Another strength of this approach is that it can also support the inheritance of an invariant by making a subclass to inherit the invariants of its superclasses. However, the approach requires an extension to both the language and its proof rules.

88.5 Conclusion

We suggested two approaches for supporting class invariants in Cleanroom-style functional program verification. The first approach systematically translates class invariants to intended functions by factoring them out. It does not require a notational or proof logic extension but is subject to the problems of readability,

reusability, maintainability, and verifiability. The second approach supports class invariants as a built-in language concept. For this, we introduced a new language construct, called an *invariant clause*, and defined its meanings in terms of proof rules. This approach addresses all the aforementioned problems associated with the first approach and closely reflects the way programmers specify and reason about the correctness of a program informally. In our study, we assumed that state variables are independent without aliasing and one state variable is not contained or owned by another. A related question is the granularity of frame axioms that assert which objects—the whole or part?—are allowed to be changed. These are future research problems.

Acknowledgments This work was supported in part by NSF grants CNS-0707874 and DUE-0837567.

References

1. Hoare CAR (1969) An axiomatic basis for computer programming. *Commun ACM* 12(10):576–580, 583
2. Rosenblum DS (1995) A practical approach to programming with assertions. *IEEE Trans Softw Eng* 21(1):19–31
3. Cheon Y, Leavens GT (2002) A simple and practical approach to unit testing: the JML and JUnit way. In: ECOOP 2002, Málaga, Spain. LNCS, vol 2374. Springer, Berlin, pp 231–255
4. Mills HD, Dyer M, Linger R (1987) Cleanroom software engineering. *IEEE Softw* 4(5):19–25
5. Staveland A (1999) Toward zero defect programming. Addison-Wesley, Reading
6. Cheon Y, Yeep C, Vela M (2011) Cleanjava: a formal notation for functional program verification. In: ITNG 2011: 8th international conference on information technology: new generations. IEEE Computer Society, Las Vegas, pp 221–226, 11–13 April 2011

Chapter 89

Normalization Rules of the Object-Oriented Data Model

Vojtěch Merunka and Jakub Tůma

Abstract There are only very few approaches to normalizing object-oriented data. Approach to object-oriented database is called class normalization. In this paper we present an approach to normalization of the object-oriented conceptual model based on UML class diagrams. First part of the paper describes the current status in the area of formal methods used for object-oriented data modeling. Second part presents four normalization rules, which are based on own experience and modified Ambler-Beck approach. These normalization rules are introduced on an example. Our method has been used in education at several universities. It has been and is also used for database design in software development projects, which we carried out. Recently, development of the CASE tool based on this approach has been started.

89.1 Introduction

Nowadays many various kinds of object-oriented software applications are used practically. We have the long-term experience with Gemstone database and Smalltalk programming, for example. Although there already are many theoretical works individually demonstrating suitability of non-relational object-oriented data model, only the procedures based on experience with imperative object-oriented programming languages are used in the area of analysis and design of data structures.

V. Merunka (✉) · J. Tůma
Department of Information Engineering, Faculty of Economics and Management,
Czech University of Life Sciences Prague, Kamýcká 129, 165 21,
Praha 6-Suchbát, Czech Republic
e-mail: merunka@pef.czu.cz

J. Tůma
e-mail: jtuma@pef.czu.cz

However some techniques like behavioral design patterns or object library components, which are optimal for algorithms in software application, can fundamentally complicate their effective database processing. Design patterns are used by programmers to improve quality of applications. For example behavioral patterns like decorator are not properly to be used with object-oriented database. As a consequence of this situation, we can see wrong usage of relationships and hierarchies among objects, breakneck tricks in code, etc. The problem of these applications is not that they do not work. Unfortunately really monstrous constructions work thanks to modern components and development systems and this is why the discussion with designers about the need to rebuild their software is very hard.

Moreover, relational design techniques as normalization, decomposition and synthesis cannot be easily used in object-oriented data structures. This is why various proposals of object-specific normalization techniques appeared in the world of software developer's community. Unfortunately no generally accepted and widely used technique or method for object-oriented data design has been introduced so far. Our solution to this problem is adapting the Ambler-Beck approach. It has been developed as a part of agile programming techniques. Our contribution is in modification of this approach towards specific data structures in object-oriented models.

Therefore we decided to discuss the formal techniques of object-oriented design. A data structure is the fundament of almost all software applications and object technology becomes the mainstream. In addition, many myths exist in the community of object-oriented software vendors and developers. For example very popular is the myth about no need for any normalization, about easiness of programming etc.

89.2 The Issue of Different Software Technologies

89.2.1 MDA

MDA is an abbreviation for Model Driven Architecture. MDA defines an approach that separates a specification of business system description (CIM—Computation Independent Model) from its computer implementation specification (PIM—Platform Independent Model); and this computer specification from the final solution on a concrete technological platform (PSM—Platform Specific Model). Each specification represents an individual viewpoint of the same problem. According to MDA, there is a mutual relationship between these three views, and the models should transform from one to another when a system is created. MDA is created and maintained by the Object Management Consortium [1].

89.2.2 Object-Oriented Programming

The object-oriented approach has its origins in the researching of operating systems, graphic user interfaces, and particularly in programming languages, that took place

in the 1970s. It differs from other software engineering approaches by incorporating non-traditional ways of thinking into the field of informatics. We look at systems by abstracting the real world in the same way as in ontological, philosophical streams. The basic element is an object that describes data structures and their behavior. In most other modeling approaches, data and behavior are described separately, and, to a certain extent, independently. OOP has been and still is explained in many books, but we think that this one [2] written by OOP pioneers, belongs to the best. The OOP can be regarded as one implementation option of PSM of more possible implementation ways. The interesting question is the position of PIM. In ideal case, this model should be independent on the following PSM. However, this does not happen in practice. Either the object-oriented data model derived from the UML or older Entity-Relation data model, which is closer to the relational database technology, is usually used for conceptual modeling on the level of abstraction correspondent to PIM. If we try to figure out how the independent conceptual data model for PIM should really look like, we will find out, that we need to use following modeling concepts:

1. Entity
2. Link between entities—however we need to distinguish between:
 - (a) IS-A relationship, i.e. taxonomy or inheritance,
 - (b) ASSOCIATED-TO relationship, i.e. link creating tuples of entities, and
 - (c) HAS-A relationship, i.e. link describing hierarchic structures or data compositions.

Detailed overview of various approaches can be found in the Table 89.1. It is obvious that on the conceptual modeling level, the relational data model and existing object-oriented data model are incompatible. That is why we presume that formal techniques known from the relational database field are not suitable for object-oriented data modeling and vice versa.

A concept of object identity is the next problem of simple adoption of relational technique for the object-oriented data model. In RDM, the identity of record is created by a value of chosen attributes (primary keys). In object data model identity of object is based on addresses into virtual memory and is independent on any value changes.

89.2.3 Object-Oriented Databases

Database systems are based on various data models, e.g. network (and its sub-species hierarchical data model), relational object-relational and object-oriented data model.

Nowadays relational database model dominates. But recent practice shows that object databases are able to compete with relational databases. Object databases are based on two substantially different data models:

Table 89.1 Possible approaches of the conceptual data modeling

| Feature | Model | Comment |
|---------|--------------------------------|---|
| R | Entity-relational | This is the traditional RDBMS model based on Chen |
| C | Network | This is the model of the network databases (IDMS) |
| I | No name | This conceptual model does not exist in the Software Engineering Or does anywhere? |
| RC | Hybrid network-relational | This is the RDBMS model combined with data containers (e.g. NF ² = non-first normal form databases) |
| RI | Extended entity-relational | This is the RDBMS model extended by the inheritance (e.g. IDEF1X) |
| CI | OOP model | This is the data model of the recent object-oriented programming languages (e.g. Java, Smalltalk, C#,...) and many programming-language-based OODBMS |
| RCI | The universal conceptual model | This data model includes all conceptual features and reflects the proposed ODMG 2.0 and 3.0 standard, but is not directly implemented in recent object-oriented programming languages |

R presence of the association relationship (i.e. RELATED-TO), *C* presence of the composition relationship (i.e. HAS-A), *I* presence of the inheritance relationship (i.e. IS-A)

1. Object-relational (or hybrid) data model (ORDM) introduces an evolutionary trend of design. It concerns on addition of the original relational data model with the support of some structures and operations known from programming languages. Most of the big producers of relational database systems (e.g. Oracle) chose this alternative. Object relational data model stays principally the same relational data model, but with extended functionality.
2. Object-oriented data model (ODM) introduces a new revolutionary trend of development. It concerns new data model, which is not built as an extension of relational data model at all. The impedance problem with storing and retrieving of object-oriented data in relational and also in object-relational databases was the main reason for creating the ODM. This is the reason, why the construction of new database models, which would be able to work with objects better, has risen. ODM and RDM differ distinctively from each other. In RDM, tables are the only possible form of logical data representation and their physical storage as well. On the other hand, ODM is similar to network databases, as we knew them in IDMS systems. The ODM can be interpreted as the renaissance of network data model. In a very simple way, it can be described by the following equation:

$$\text{network data model} + \text{objects} + \text{methods} + \text{polymorphism} = \text{ODM}$$

It is reasonable to assume that the importance of object databases will grow in the near future, because there are now many applications, where object-oriented database shows its advantages. Common attribute of these applications is large amount of complex data structures and their variability during their lifetime.

Those systems can work with more than hundred or thousand various mutually composed and changing data types. Moreover, the queries over these structures require common polymorphism and abstraction. In those systems, for example, we need to write down the queries over sets containing elements of various types. At the same time we expect that while adding or updating data types it will not be required to change already written queries and related data structures. Good example of those systems are data-warehouses. Those systems are characteristic not only for company management systems, but also for various governance evidence systems, hospital information systems and information systems containing ecological information, agricultural information, historiographical information as well, decision support in marketing and finance [3–5].

On the other hand it is necessary to note that relational database works very well in area, where database structure is constant. This means that new data types not are added during lifetime of such system. Moreover, relational databases traditionally achieve very good performance if the database consists of large amount but simply structured records.

89.2.4 Miscellaneous Approaches to Object-Oriented Normalization

Some various papers aroused since 1980s (for example [6]). First papers applied to the enlargement of relational techniques, but we can meet the papers specialized to object-oriented data structures in recent last years. Object-oriented database normalization is called class normalization which is introduced [7]. There are several research groups in the world interested in object databases. The results of their studies are used in object databases construction. The international organization ODMG—Object Database Management Group—supports publications and conferences on this topic.

89.2.5 Nootenboom's OONF

According to Hank Nootenboom the first three relational normal forms are universally valid for the object-oriented data model as well as for other possible data models [8]. He introduces the concept of only one additional normal form for objects as a substitute for fourth and fifth relational normal forms, having the following definition:

A collection of objects is in OONF if it is in 3NF and contains meaningful data elements only.

89.2.6 Khodorkovsky's ONF, 4ONF, 5ONF and 6ONF

The paper [9] proposes object normal forms, which concerns the right relation among objects and methods. The rules of the defined object normal forms are based on modification of relational definitions of 4NF, 5NF (and 6NF, which is author's original refinement of 5NF). The author calls these modifications of classical definitions as 4ONF, 5ONF and 6ONF.

The paper is considered to be more elaborated formulation of almost similar ideas as the example above. The author says, that 1NF, 2NF and 3NF are common for relational and object databases.

89.2.7 Australian-Swiss ONF

The authors [10] present only one ONF on various types of functional dependencies among objects. Concretely, path dependency concerns a composition of objects and navigability among objects, local dependency concerns relations of internal object and global dependency concerns behavioral requirements on application. Object-oriented structure is in ONF, if user requirements on applications are covered by a set of functional dependencies. This method relates to the behavioral requirement of object databases, but it is not specifically focused on the conceptual modeling of data [13–17].

89.2.8 Three Ambler-Beck's Object Normal Forms

Ambler and Beck are pioneers of the agile approach in programming. They introduced three object-oriented normal forms for object-oriented applications [11, 12]. These normal forms are analogous with first, second and third relational normal form. The authors talk about these object normal forms as a tool for objects classes' normalization complementary with technique of design patterns. Let's look at their proposals in detail:

A class is in 1ONF when specific behavior required by an attribute that is actually a collection of similar attributes is encapsulated within its own class. An object schema is in 1ONF when all of its classes are in 1ONF.

It is evident from the definition and the example that authors wanted to build the first normal form analogically to the first relational normal form. From experience, it is little confusing that object can be non-normalized even if it already has associated collection of encapsulated objects. See attribute seminars of class Student in the Fig. 89.1. In this example the class Student contains the collection seminars, but the class Student is still in 0ONF. The collection seminars from 0ONF does not differ much from the relation takes in 1ONF in the Fig. 89.2. The difference between 0ONF and 1ONF is only in presence of specific methods of class Student (Figs. 89.3, 89.4).

Fig. 89.1 0ONF

| Student |
|---|
| studentNumber name address phoneNumber seminars |
| addSeminar() dropSeminar() printSchedule() setProfessor() setCourseName() getSeminarLength() |

A class is in second object normal form (2ONF) when it is in 1ONF and when share behavior that is needed by more than one instance of the class is encapsulated within its own class (es). An object schema is in 2ONF when all of its classes are in 2ONF.

As the definition and the example show, the 2ONF requires to detach attributes, which are shared by more objects, into separate objects. In our experience, this definition is well accepted. Also, this definition offers analogous result, as the second relational object form in relational databases.

A class is in third object normal form (3ONF) when it is in 2ONF and when it encapsulates only one set of cohesive behaviors. An object schema is in 3ONF when all of its classes are in 3ONF.

It is possible to recognize, that the third and the last object normal form by Ambler gives analogous results as the third relational normal form. This is our experience as well. It concerns the characteristics within some objects, which might be interpreted and behave as an independent object. In this case we need to exclude them into new separate object.

89.3 Our Experience

We have good results with Ambler-Beck approach. But we have found that object-oriented community expects bit different technique:

1. It has to be very simple, precise, and understandable and should work with minimum of abstract concepts, similarly as the classical relational normalization. We suppose that introduction of difficult definitions distinctively exceeding over the range of classical normal forms by having a lot of types of concepts and relations, is not the right way.
2. It should be focused concretely on object-oriented modeling of data structures. We need to model structures of objects used for data storage and data manipulation. We do not need to model objects responsible for functional

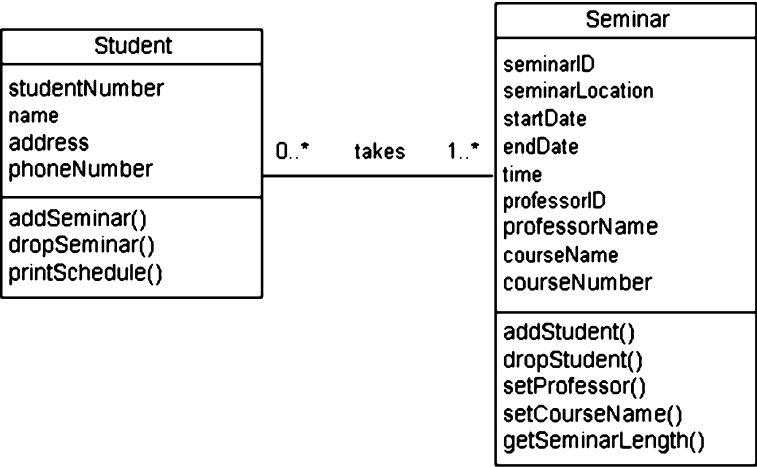


Fig. 89.2 Ambler’s and Beck’s 1ONF

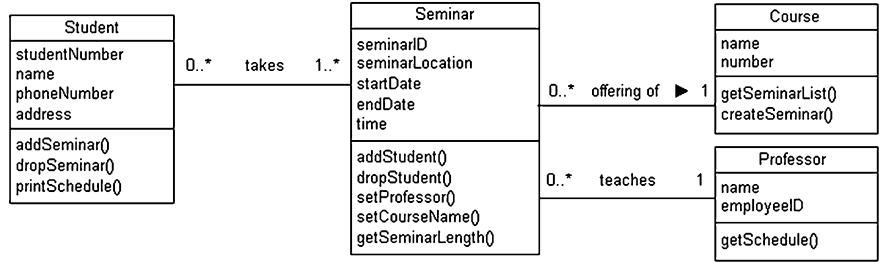


Fig. 89.3 Ambler’s and Beck’s 2ONF

behavior of applications. For these another behavioral objects we already have design patterns and other programming techniques. We do not need to duplicate these proved techniques. We think that original Ambler’s approach needlessly tries to solve everything in one.

We have to define, what do we exactly understand by the concept of data object. Data objects serve only for data storing and manipulation. We will not work with data elements and with methods separately. This is proposed by [4]. We will define only one common concept of an attribute. By an attribute, we will understand the data property of an object, regardless if the data property is coming from a data element or if this data property is a result of a method.

Of course, there is a question, if such simplification is not too much. Ambler-Beck’s original approach works separately with data and methods and uses both of them separately. But we think that we can allow this simplification for the data objects, because our approach is not aimed for behavioral design of application structure.

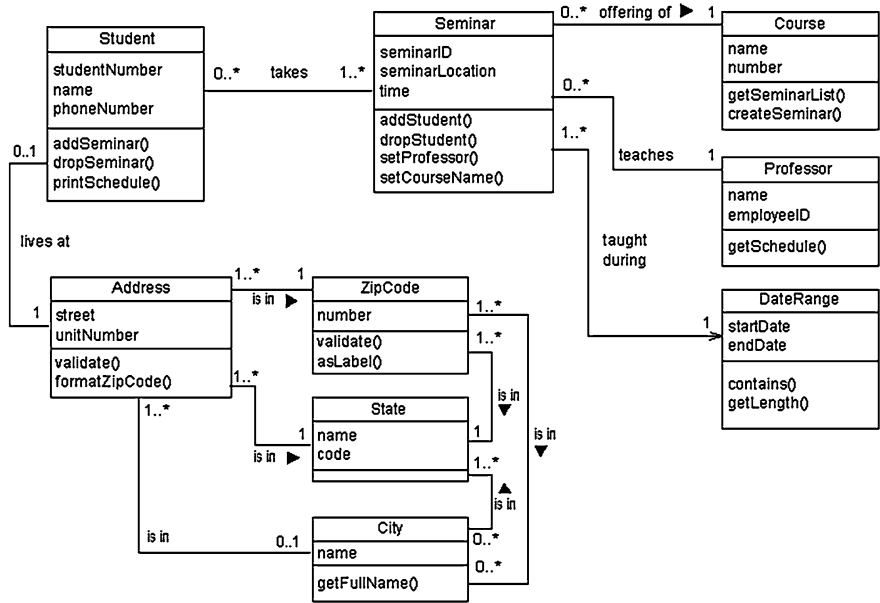


Fig. 89.4 Ambler’s and Beck’s 3ONF

89.3.1 A First Normal Form Rule

Definition 1 A class is in the first object normal form (1ONF) when its objects do not contain group of repetitive attributes. Repetitive attributes must be extracted into objects of a new class. The group of repetitive attributes is then replaced by the link at the collection of the new objects. An object schema is in 1ONF when all of its classes are in 1ONF.

More formally; Let us have an object a in the object system Ω as $\alpha \in \Omega$, where for $k > 1$ (length of collections of similar attributes) and $n > 1$ (number of repetition of these collections) is $data(a) = [\dots, x_1^1, \dots, x_k^1, \dots, x_1^n, \dots, x_k^n]$ having $\forall i \in (1, \dots, k): class(x_i^1) = class(x_i^2) = \dots = class(x_i^n)$.

Then it is required to modify object a and create new objects $b_j \in \Omega$ for $j \in (1, \dots, n)$ as $data(a) = [\dots, \{b_j\}, \dots]$ and $data(b_j) = [x_1^j, \dots, x_k^j]$.

In the Fig. 89.5 there is the example of data structure in non-normalized form and in the Fig. 89.6 there is the same example in 1ONF. On the contrary with the original Ambler-Beck’s approach, we do not assume designers recognize groups of repetitive attributes automatically and extract them out into independent classes. The problem is not always trivial as in presented example. Repetitive attributes can exist under various names, which are not easy visible on the first sight (Figs. 89.7, 89.8).

Fig. 89.5 Unnormalized model

| Order | Supply |
|----------------------|----------------------|
| supplier firstname | supplier firstname |
| supplier surname | supplier surname |
| supplier address | supplier address |
| client firstname | client firstname |
| client surname | client surname |
| client address | client address |
| order date | supply date |
| payment mode | payment mode |
| first product name | first product name |
| first product price | first product price |
| second product name | second product name |
| second product price | second product price |
| third product name | third product name |
| third product price | third product price |
| ... | ... |

89.3.2 *Second Normal form Rule*

Definition 2 A class is in the second object normal form (2ONF) when it is in 1ONF and when its objects do not contain attribute or group of attributes, which are shared with another object. Shared attributes must be extracted into new objects of a new class, and in all objects, where they appeared, must be replaced by the link to the object of the new class. An object schema is in 2ONF when all of its classes are in 2ONF.

More formally; Let us have two objects $a, b \in \Omega$ for $k > 1$ (length of a collection of shared attributes) as $data(a) = [\dots, x_1, \dots, x_k, \dots]$ and $data(b) = [\dots, y_1, \dots, y_k, \dots]$ having $A_i \in (1, \dots, k) : x_i = y_i$.

Then it is required to modify objects a and b and create new object $c \in \Omega$ as $data(a) = [\dots, c, \dots]$ and $data(b) = [\dots, c, \dots]$ and $data(c) = [x_1, \dots, x_k] = [y_1, \dots, y_k]$.

It concerns the attributes supplier’s first name, supplier’s surname and his address and client’s first name, client’s surname, his address and method of payment in our example. Because these attributes are common for both concrete order and supply, it was necessary to create the new object class Contract.

89.3.3 *Third Normal form Rule*

Definition 3 A class is in the third object normal form (3ONF) when it is in 2ONF and when its objects do not contain attribute or group of attributes, which have the independent interpretation in the modeled system. These attributes must be extracted into objects of a new class and in objects, where they appeared, must be replaced by the link to this new object. An object schema is in 3ONF when all of its classes are in 3ONF.

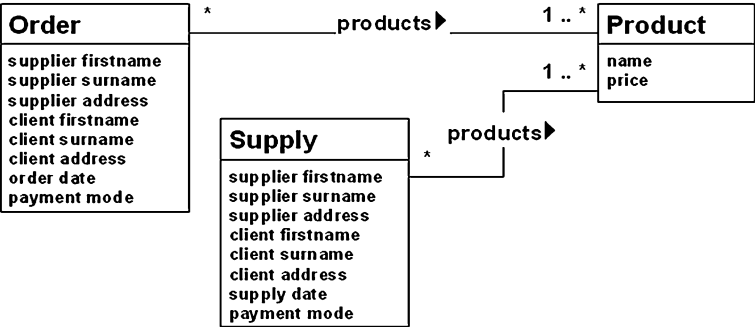


Fig. 89.6 Model in 1ONF

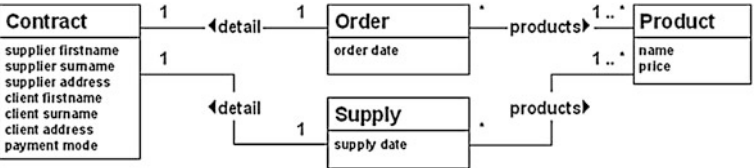


Fig. 89.7 Model in 2ONF

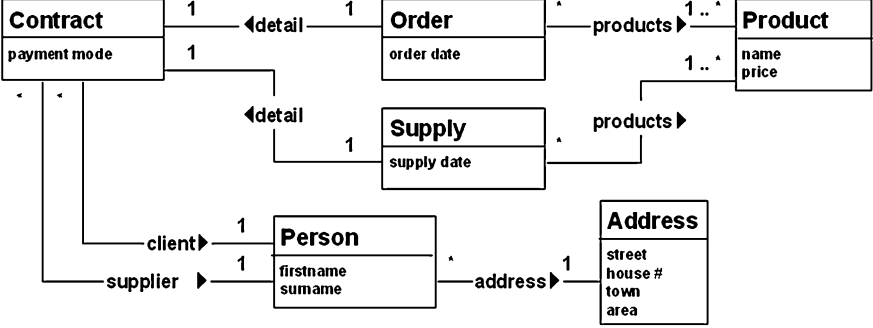


Fig. 89.8 Model in 3ONF

More formally; Let us have two objects $\alpha \in \Omega$ for $k > 1$ (length of a collection of independent attributes) as $data(\alpha) = [..., x_1, ..., x_k, ...]$, where $[x_1, ..., x_k]$ is collection of independent attributes.

Then it is required to create object $b \in \Omega$ and modify object α as $data(\alpha) = [..., b, ...]$ and $data(b) = [x_1, ..., x_k]$

It concerns the data about suppliers and clients in the objects of the class **Contract**. These attributes represent some persons having independent interpretation on contracts. The same applies to addresses.

89.3.4 Fourth Normal form Rule

Definition 4 A class is in the fourth object normal form (4ONF) when it is in 3ONF and when there is no other class in the system, which defines the same attributes. These attributes must be extracted from classes, where they are duplicated, and affected classes must be connected using class inheritance in order to exclude data definition duplicates. If there is no existing class to be reused as a inheritance superclass, a new superclass must be added into the system. An object schema is in 4ONF when all of its classes are in 4ONF.

More formally; For each two objects a, b in the object system Ω as $a, b \in \Omega$ having $data(a) = [x_1, \dots, x_k]$ and $data(b) = [\dots, y_1, \dots, y_k, \dots]$ where $\forall i \in (1, \dots, k): class(x_i) = class(y_i)$, classes of these objects a, b must have inheritance relationship as $class(a)$ predecessor of- $class(b)$ in order to avoid duplicates.

89.4 Conclusions

The object-oriented approach used practically is ahead of theoretical foundation and formal techniques. Perspective and practically used technology—the object-oriented approach—still does not have comprehensible and universally accepted theoretical foundation and formal techniques. It is known, that several research centers are interested in this theme, but any coherent and widely accepted results were not yet published in recent years. Absence of reputable formal tools and techniques is the big problem of this promising technology.

Theoretical conclusion is behavioral design patterns are properly for inner component object-oriented applications and class normalization is for object-oriented database. Fundamentally class normalization is a technique for improving the quality of your object schemas [12].

Therefore we suppose that near future may bring maybe some alternative approaches, more or less similar to our approach we presented in this paper. Our method has been used in education at University of Thessaly in Volos, Alexandrian Technological Institute in Thessaloniki, Lehigh University in Pennsylvania, Czech Technical University and Czech University of Life Sciences. It was also used for database design in software development projects, which we carried out for a large international consulting company Deloitte. Recently, the project on object-oriented CASE tool supporting this approach sponsored by a consortium of software companies has been started. Our future research will focus on describing the rules of our object-oriented normal forms as a sequence of refactoring steps.

Acknowledgments The authors would like to acknowledge the support of the Czech Ministry of Education, Youth and Sports by the grant project MSM 6046070904.

References

1. David SF (2003) Model driven architecture: applying MDA to enterprise computing. Wiley, New York. ISBN 0-471-31920-1
2. Goldberg A, Rubin SK (1995) Succeeding with objects—decision frameworks for project management. Addison Wesley, Reading. ISBN 0-201-62878-3
3. Kroha P (1995) Objects and databases. McGraw Hill, London. ISBN 0-07-707790-3
4. Loomis M, Chaundri A Object databases in practice. ISBN 013899725X
5. Vanicek J (2004) Data gathering for science and research. *Agric Econ* 50(1):29–34
6. Mok WY, Ng Y-K, Embley DW (1992) An improved nested normal form for use in object-oriented software systems. In: Proceedings of the 2nd international computer science conference: data and knowledge engineering: theory and applications, December 1992. Hong Kong, pp 446–452
7. Beck K (2003) Agile database techniques—effective strategies for the agile software developer. Wiley, New York. ISBN 0471202835
8. Nootenboom Henk Jan: Nuts—a online column about software design. <http://www.sum-it.nl/en200239.html>
9. Khodorkovsky VV (2002) On normalization of relations in databases. *Progr Comput Softw* 28 (1):4–52
10. Tari Z, Stokes J, Spaccapietra S (1997) Object normal forms and dependency constraints for object-oriented schemata. *ACM Trans Database Syst* 22(4):513–569
11. Ambler S (1997) Building object applications that work, your step-by-step hand-book for developing robust systems using object technology. Cambridge University Press/SIGS Books, New York. ISBN 0521-64826-2
12. Ambler S (2009) Object orientation § bringing Data professionals and application developers together. <http://www.agiledata.org/essays/>. Accessed Jan 2009
13. Barry D (1996) The object database handbook: how to select, implement, and use object-oriented databases. Wiley, New York. ISBN 0471147184
14. Blaha M, Premerlani M (1998) Object-oriented modeling and design for database applications. Prentice Hall, Upper Saddle River. ISBN 0-13-123829-9
15. Catell RG (2000) The object data normal: ODMG 3.0. Morgan Kaufmann, San Mateo. ISBN 1558606475
16. Gemstone Object Server § documentation & non-commercial version download. <http://www.gemstone.com>. Accessed Jan 2009
17. Yonghui W, Zhou A (2001) Research on normalization design for complex object schemes, info-tech and info-net. In: Proceedings of ICII, vol 5. Beijing, pp 101–106

Chapter 90

Location Based Overlapping Mobility Aware Network Model

Abdul Razaque, Aziz Alotaibi and Khaled Elliethy

Abstract The mobile devices provide dynamic behavior in different geographical locations. The location based information plays integral part for enhancing the performance of mobile devices. The right selection of location improves the smooth transition process of context aware services to mobile devices. For example location based information services cover the shopping, travel information, entertainment, different navigation, event information and tracking services. All of these services depend on the actual position of the users. The efficiency and smooth transfer of data contexts also relate to the type of services available to mobile devices. Therefore, the proper selection of technology and positioning methods need to be chosen carefully to obtain the desired accuracy. Several models have been proposed to improve the performance of mobile devices. This paper proposes new approach of location based overlapping that augments the performance of mobile devices and also provides better quality of service (QoS). The proposal is supported by an algorithm that helps to choose the best location for mobile device where signals are stronger. The mobile devices select strong signal-providing base station and each mobile device is free to select any overlapped area. To prove the claim, the proposed model is simulated in ns2 and findings are compared with existing models.

A. Razaque (✉) · A. Alotaibi · K. Elliethy
Wireless and Mobile Communication laboratory (WMC), Department of Computer Science
and Engineering, University of Bridgeport, Bridgeport, USA
e-mail: arazaque@bridgeport.edu

A. Alotaibi
e-mail: aziza@bridgeport.edu

K. Elliethy
e-mail: elleithy@bridgeport.edu

90.1 Introduction

The primary positioning methods utilized for serving of the cells are considered as cell based triangulation and satellite based methods [1]. Cell based positioning methods solely evaluate pre-defined radio parameters in the air interface. There is a variety of parameters that can be used in GSM such as cell identifier, timing advance, signal levels of the service cell and the neighboring cells. These parameters have limited accuracy despite they can operate with minimal impact on the network architecture without any modifications to the handsets. Therefore, cell based positioning algorithm is primarily used by location based services application. However, this technique is not free from imperfections as the overlapping of cells may lead to obtaining cell numbers that are different from the desired mobile station.

The triangulation method [2] traces mobile location by observing the propagation time measurement for the same signal from various points within the network. While the accuracy levels are higher than the cell based method; separate measuring equipment is required for conducting the observations. The best positioning accuracy can be acquired by using the satellite based method. Data collection is dependent on the mobile terminals that are capable of sending, receiving and processing signals received via satellite. A major drawback of this method is the reliance on GPS which gives the advance positioning and tracking methods.

Radio parameters such as cell identifier, cell identifier + timing advance, cell identifier + timing advance of several adjacent cells can be adopted for creating network measurement reports. These reports in GSM contain the above mentioned parameters on the service cell and the adjacent cell. Network Measurement Reports employ one of the two evaluation methods; The Distance Calculation Method or the Data Base Correlation. The former makes several assumptions on this propagation model. In particular, the HATA model [3] allows us to know the distances from the signal strengths. This notion also applies to fading; therefore certain filtering techniques may be put in place in order to get the exact positioning. The Data base correlation method uses tools to grid several points within the network. The desired position can be estimated by calculating the measured NMR and comparing it with the desired value.

This paper proposes an improved version of cell identifier advance wherein the TA values are deployed to measure the distance between the base station and the mobile station. Within this parameter, each base station is equipped with the sector antennas and the MS regularly sends the signal strength of the serving and the neighboring base stations. Neighbor sector included in the same base station are also incorporated and the direction of the MS is calculated in this manner. "The ratio of the signal levels received at the MS from any pair of sectors of the same BTS is independent of the distance between BTS and MS. Also, since the signals take the same path, this ratio is not subject to shadowing effects" [4]. This is not the case for the overlapped area.

Potential solutions to this barrier include division of the area into locations with Omni directional antennas equipped base stations. In case the MS enters the overlapped area, the location with the higher signal strength is served. Although the Omni directional antennas radiate power in all directions, power radiation decreases with the elevation angle. The radiation pattern is spherical in shape and there is uniformity in all directions as far as the gains from this antenna are concerned. In [Sect. 90.2](#), discuss the salient features of related approaches. [Section 90.3](#) describes the proposed overlapping mobility aware network model. [Section 90.4](#) gives the simulation and the analysis of result. Finally, conclusions are given in [Sect. 90.5](#).

90.2 Related Work

This part explains the salient features of location based mobile network stations. Barkhuus et al., have the location tracking services based on either the user's location or the self-tracking potential of the device, for instance, position awareness service [\[5\]](#). The authors compare between privacy concerns and the use of location based services. The results indicate that in spite of the similarity in their usage, there is greater privacy concern generated from the use of location tracking services in comparison to position aware services. The study concludes that the position awareness services have the potential of further development.

Location Based Services (LBS) is applied in M-commerce using the actual position of the terminal in the service provision is a special case of M-Commerce. The author of [\[4\]](#) focuses on the requirements for the LBS. Since technology serves as an important facilitator and a limiting factor, the paper examines technology aspects pertaining to LBS. Furthermore it discusses the requirements for the user, system and the infrastructure. The paper concludes with a design and implementation presentation of an LBS application that runs on Java based handsets.

In [\[6\]](#), the authors cover the technological aspects and market opportunities for location based service. The authors in [\[7\]](#) propose a real life scenario about the location based service (LBS) and discuss about the system that may help and detect the actual place of accident precisely and quickly.

An open mobile alliance location based service (LBS) standard is discussed in [\[8\]](#). The standard is used for detecting positions of mobile devices and covers the wireless access protocol (WAP) in the old forum and Location Interoperability Forum (LIF).

All the approaches presented in this section help to detect the position of mobile devices but our approach provides the best network signals strength to mobile devices to achieve better Quality of service (QoS) and maintain a high data delivery rate in dynamic and scalable environments [\[10–13\]](#).

90.3 Proposed Overlapping Mobility Aware Network Model

The model presented in this section divides the whole regions into small regions. The divided regions are given the labels loc1, loc2, loc3 and loc-n as shown in Fig. 90.1.

All of these locations cover the base stations. Each of the base stations is equipped with an Omni directional antenna. The radiations that are used for spreading the signals to all directions are shown in Fig. 90.2. The main components of the model are Omni directional antenna, base station and the divided regions like loc 1, loc2, loc3 and loc-n.

Let us consider an example as given in Fig. 90.3. Assume that the mobile station enters location 3, service is now provided by many base stations such as BS5, BS6 (loc1), BS2, BS3. BS4 (loc2), BS7, BS8 (loc-n). Not all base stations serves simultaneously but only a single base station serves which provides the stronger signal. The stronger signals are calculated on base station 2 and location 3. Signal strength is considered as magnitude of electric field and calculated significant distance from the transmitting antenna at the reference point. The signal strength (SS) can be calculated with following formulas:

$$V = \frac{e * \lambda}{\pi} \quad (90.1)$$

Here, v is voltage; λ is the wavelength and value of π is 3.14. At the receiver side, the terminated voltage is made $\frac{1}{2}$ and half is lost between matched termination and source impedance. If terminated voltage is not set half then system will not be boot again because full voltage does not support.

We know that wave length relates to frequency; thus

$$\text{Frequency} * \text{wavelength} = \text{speed of light } 3 \times 10^8 \text{ m/s} \quad (90.2)$$

The frequency is derived from number of channels.

We know that the channel bandwidth is 8 MHz at phase alternating line in Data over cable service interface specification (DOCSIS) standard.

Channel 21 communicates with 470 MHz approximately.

$$\text{Therefore, Channel } 44 = (44 - 21) * 8 \quad (90.3)$$

$$f = 8 * (\text{channel} - 21) + 470 \text{ MHz} \quad (90.4)$$

Substitute the values:

$$= 8 * (44 - 21) + 470 = 654 \text{ MHz}$$

$$\lambda = 3 * 10^8 / 654 * 10^6$$

$$= 0.46 \text{ m}$$

To find the received signal voltage of dipole with by applying Eq. (90.1).

Fig. 90.1 Showing the division of region into locations

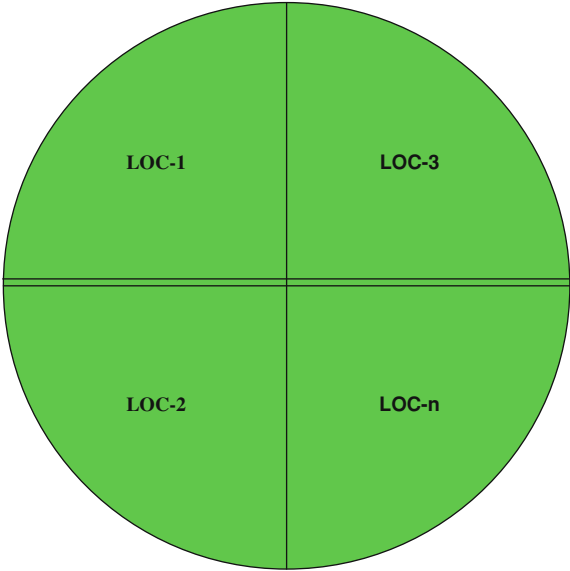
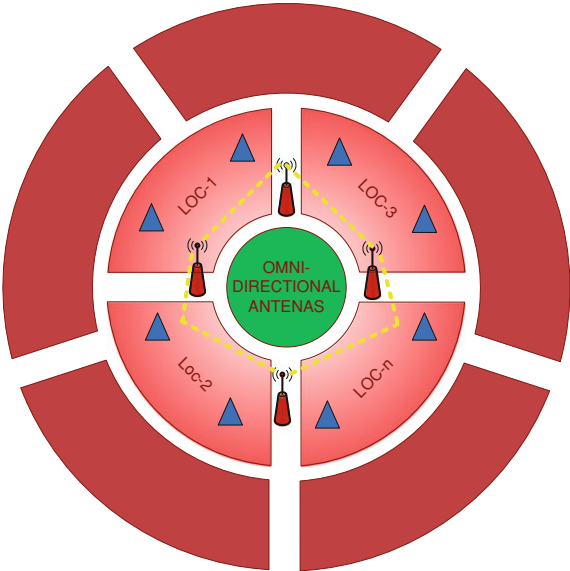


Fig. 90.2 Using Omni directional antenna in each base station



$$V = \frac{e * \lambda}{\pi}$$

Substitute the values

$$V = 0.5 * 0.095 * 0.46 / 3.14159$$

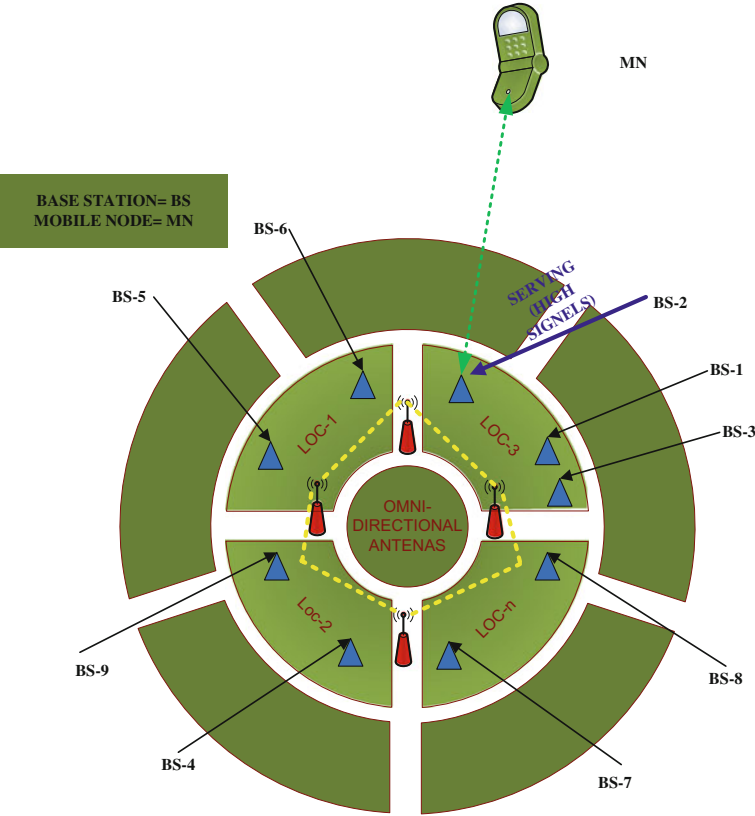


Fig. 90.3 Mobile station in the loc2

$V = 0.0069$; this is actual voltage needed to detect high signal cell.
Now, converting into dBuV

$$V = 20 \log_{10}(0.0069/10^{-6})$$
$$V = 77 \text{ dBuV}$$

By adding the value of aerial (VA) that is 7.3 dB

$$SS = V + VA$$

Substitute the values:

$$SS = 77 + 7.3 = 84 \text{ dBuV}$$

The received signal strength = 84 dBuV
The probability and position is calculated as follows:
The coordinates ($a1, b1$) and ($a2, b2$) for each base station is given. In addition the average distances are $s1$ and $s2$.

$$sbs = \sqrt{(a_2 - a_1)^2} + \sqrt{(b_2 - b_1)^2}$$

Where Sbs = Distance of base station;

$$l_1 = s_1 + s_2 - S_{bs}$$

$$l_2 = \sqrt{s_2^2 - l_1^2}$$

$$\sin(i) = b_2 - b_1 / s_{bs}$$

$$\cos(j) = a_2 - a_1 / s_{bs}$$

8 & 9 meets the point P, then we can obtain:

$$a_p = 7a_2 - l_1 \cdot \cos(i)$$

$$b_p = b_2 - l_1 \cdot \sin(i)$$

The most probable position of mobile station (MS) depends on two solutions:

$$a = a_p - l_2 \cdot \sin(i) = a_2 - l_1 \cdot \cos(i) - l_2 \cdot \sin(i)$$

$$b = b_p - l_2 \cdot \cos(i) = b_2 - l_1 \cdot \sin(i) - l_2 \cdot \cos(i)$$

Therefore,

$$a = a_p + l_2 \cdot \sin(i) = a_2 - l_1 \cdot \cos(i) + l_2 \cdot \sin(i)$$

$$b = b_p - l_2 \cdot \cos(i) = b_2 - l_1 \cdot \sin(i) - l_2 \cdot \cos(i)$$

This is different from the other models because irrespective of the overlapped region, it serves the MS. Generally in all other techniques it may go wrong in getting the correct cell number, using of the GPS, other additional measurement equipment. The proposal is easy to implement and algorithm 1 explains the process of penetrating the location.

Algorithm:

1. initialization weak=a1; Strong=a2; neutral=a3; no range=an; BS=N
2. if MN=N || MN ≠ N
3. MN Is detected
4. if MN==a1,a2,a3,an
- Start searching for the stronger signal
5. If (loc1 > loc2 && loc1 > loc3 && loc1 > locn)
6. Mn=loc1
7. else if (loc2 > loc3&& loc2 > locn)
8. Mn=loc2
9. else if (loc3> locn)
10. Mn=loc3
11. else Mn=locn
- //Mobile node chooses the appropriate location
12. MN→LOCn→N

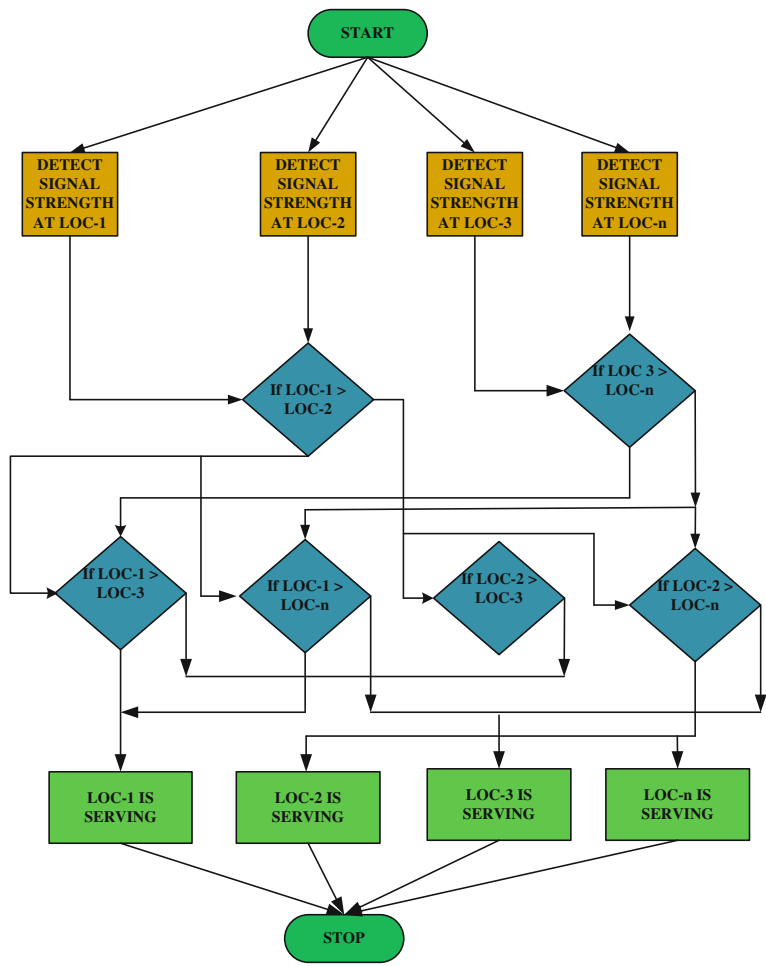


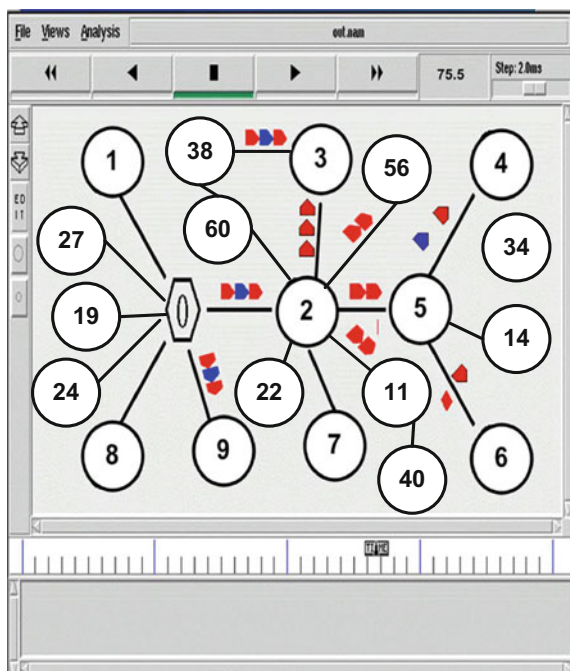
Fig. 90.4 Showing the process of detecting the location

The process of searching high signal location, the flowchart explained in Fig. 90.4 helps to detect the correct location with appropriate base station to be connected.

90.4 Simulation and Analysis of Result

Ns2.28 on Red Hat 8 is used for simulation. The Random Waypoint mobile scenario is generated. The simulator gives a proper model for signal propagation and transmission range is 250 m. The sensing and interference range is 550 m.

Fig. 90.5 NAM screenshot of simulated scenario in ns2



Overlapped, cell based and triangular approaches have been simulated by using TCP.

Hence, same scenario is simulated to ensure performance of proposed model and comparing with already existing models. The length of packet is 1,040 bytes including 40 bytes overhead. In this simulation, 9 mobile nodes are participating and trying to get the stronger signal based location. NAM screenshot for overlapping mobility aware network model is shown in Fig. 90.5.

Scenario is mobility aware, and nodes move within rectangular field of 600 * 1,200 m. RW generates mobile scenario and starts searching the location for nodes. Constant values for pause time have been set, which are 3 s. Total simulation time is 75.5 s. The minimum speed of the node (V_{\min}) is 0 m/s and maximum speed (V_{\max}) are 10 m/s respectively. The moving speed of node is randomly obtained through uniform division [V_{\min} , V_{\max}]. We run simulations, which cover combination of the pause time and moving speed of nodes [9].

On the basis of simulation, goodput performance for each model and hit rate for proposed approach is obtained and shown in Figs. 90.6 and 90.7.

Goodput performance is obtained by following formula:

$$\text{Goodput \%} = \sum_1^n \text{Ftp received} * 100 / \sum_1^n \text{Ftp total transmitted}$$

Fig. 90.6 Goodput of overlapping model with respect to existing schemes

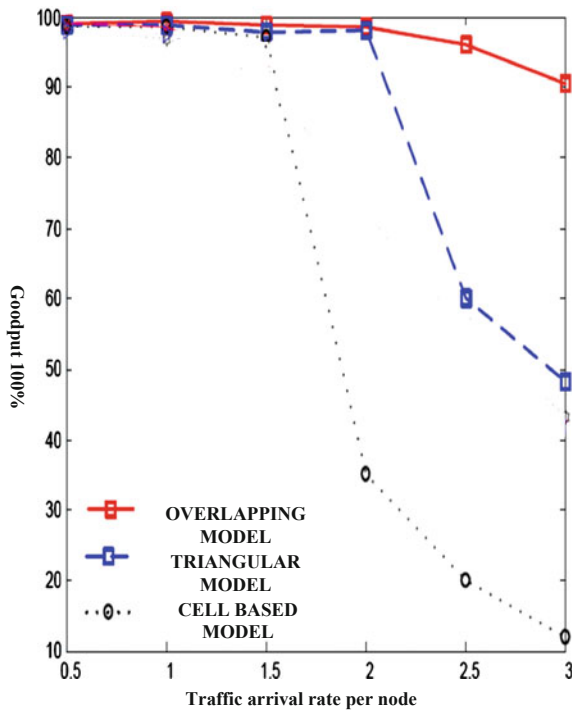
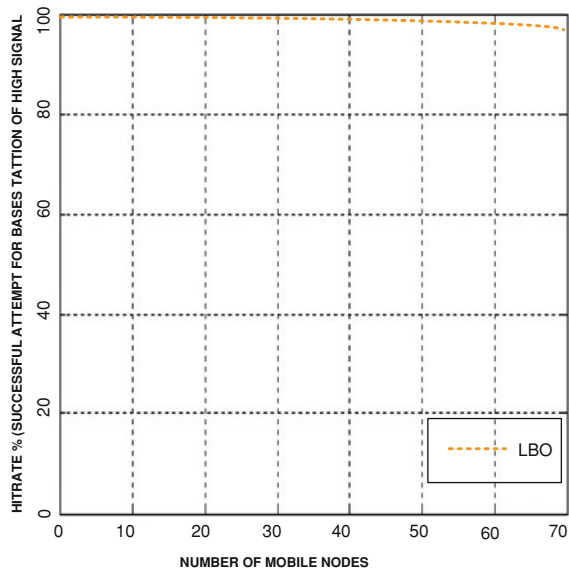


Fig. 90.7 Hit rate ratio versus number of mobile nodes at overlapping mobility aware network model



On the basis of findings, it is clear that each scheme decreases the performance but regardless our scheme gives better goodput. Other schemes incur several problems relating to cache usage.

Hit ratio shows the number of attempts made by different number of mobile devices. The purpose of calculating hit rate is to measure the scalability of this approach. In this simulation, hit rate is calculated when all mobile devices try to achieve highest signal producing base station. The hit rate is 100 % obtained up to 50 mobile nodes. If number of mobile nodes increases then hit rate decreases. The reason of decreasing the hit rate is not scalability issue because underlying routing protocols do not have sound capability to maintain the connectivity. At the same time, some mobile nodes join and some leave that cause the failure of connection. The mobility is also major concern for reducing the hit rate. If number of mobile devices increase then used random way point mobility model does not have convergence capacity to control the moment of mobile devices. From other side, decrease hit rate is less than 2 % that is also negligible because all mobile devices make attempt concurrently for getting highest signal at base station. In real environment, all mobile nodes do not make attempt at the same time. Furthermore simulated area is very small that also incur more chances of discontinuation.

$$\text{Hit rate} = \frac{\text{number of successful attempts} \times \text{number of mobile devices} \times 100}{\text{total number of attempts} \times \text{number of mobile devices}}$$

Multiple routes obviously give benefits but creates disadvantage due to high mobility. In larger networks, the source-routing principle can also generate a trouble. Our scheme has best option for selecting any area where network connectivity is better than rest of areas. Node joins to any location that is the reason, overlapped scheme give healthy performance.

90.5 Conclusion

In this paper, a unified approach for detecting high signal strength in various locations is proposed. This approach utilizes overlapping area to improve the communication performance. We have adopted Omni antenna to provide the signals to all directions for satisfying QoS. The mathematical model aids in providing strong signal in several locations. With this model, it is possible to increase the data delivery rate and employed in various fields. It is also better choice for extremely mobile and dynamic applications, which are not substantiated by centralized administration. This model can be implemented in disaster situations, scattered educational institutions and defense department. These environments need such networks to route data packets through dynamically mobile nodes with high data delivery rate. Although in this paper we have only used normal number of mobile nodes, additional mobile nodes can be deployed without any issues of complexity. The model sustains low network overhead and minimum congestion on the transport level because each node chooses high signal location to communicate with other nodes. The analysis of simulation results show better findings

achieved when compared with other popular existing models. Hence, significant amount of network resources are saved by choosing correct base station in any location. Therefore, we conclude that this unified model provides better QoS, scalable and dynamic access for communication. Further enhancement to the model such as accurate radio interference will be augmented which will foster realistic communication environment.

References

1. G3GPP TS 43.059 Functional stage 2 description of location services (LCS) in CERA. Technical report, (Release 4)
2. Cafbay JJ, Suber GL (1998) Overview of radiolocation in CDMA cellular systems. *IEEE Commun Mag* 36(4):38–45
3. Hata M (1980) Empirical formula for propagation loss in Land mobile radio services. *IEEE Trans Veh Technol* 29(3):317–325
4. Tangemann M, Nikolai D, Alcarel SEL AG (2003) A new network-based Positioning Method for Location Services in 2G and 3G mobile communications matthias schreiner, Research & Innovation, P70430 Srunger, German)
5. Barkhuus L, Dey A (2003) Location-based services for mobile technology: a study of users' privacy concerns. In: *Proceedings of the interact 2003, 9th IFIP TC13 international conference on human-computer interaction*, July 2003
6. Priessl B, Bouwman H, Steinfield C (eds) (2004) *Elife, the development of location based services in mobile commerce*. Springer, Berlin
7. Alkhateeb F, Al Maghayreh E, Tubishat M, Aljawarneh S (2010) The use of location based services for very fast and precise accidents' reporting and locating. In: *Proceeding of IEEE international conference on intelligent systems, modeling and simulation (ISMS)*
8. Adams PM, Wayne G, Ashwell B, Baxter R (2003) Location-based services—an overview of the standards. *BT Technol J* 21(1):34–43
9. Camp T, Boleng J, Williams B, Wilcox L, Navidi W (2003) Performance comparison of two location based routing protocols for ad hoc networks, colorado school of mines, NSF Grants ANI-9996156 and ANI- 0073699
10. Drane C, Macmughan M, Scott C (1998) Positioning GSM telephone?. *IEEE Communns* 36(4):44–59
11. Wang SS, Green M, Malkawi M (2002) Mobile positioning technologies and location services. In: *Proceedings of the 2002 IEEE WOan d wireless conference*, Boston, MA, USA, Aug 2002, pp 9–12
12. Hellebrandt M, Mathar R (1999) Location tracking of mobiles in cellular radio networks. *IEEE Trans Veh Technol* 48(5):1558–1562
13. Tsalgatidou A, veijalainen J, Jouni M, A Katasonov, Hadjiefthymiades S (2000) Mobile E-commerce and location based services technology and requirements. Technical report, financially supported by Finnish National Technology Agency and the Special Account for Research Grants of the National and Kapodistrian University of Athens (ELKE)

Chapter 91

Expert System for Evaluating Learning Management Systems Based on Traceability

E. Valdez-Silva, P. Y. Reyes, M. A. Alvarez, J. Rojas
and V. Menendez-Dominguez

Abstract Nowadays, the quality of learning management systems (LMS) is an important feature that helps ensure a good service. Many schools use them in their academic activities; however, the quality of such systems has not been analyzed in detail, therefore it is necessary to create rules to govern their operation and development. This paper proposes an expert system thought to assist the user in the evaluation of learning management systems. The system proposed in this paper considers quality standards in software engineering and distance education to establish a traceability model that combines techniques of data analysis, based on the evaluations of the characteristics of a learning management systems, along with the perception of users, and provides information that is necessary to enhance the quality of the system.

91.1 Introduction

In recent years, Internet, along with different information and communication technologies have been incorporated in a relevant way to people's social life, especially to the academic life, helping develop new training proposals online with this scenario [1].

E. Valdez-Silva (✉) · P. Y. Reyes · M. A. Alvarez
Department of Postgraduate and Research, Universidad Politécnica de Aguascalientes,
Aguascalientes, Mexico
e-mail: emilia.valdez@upa.edu.mx

J. Rojas
Cloud Technologies Consulting, Mexico City, Mexico

V. Menendez-Dominguez
Autonomous University of Yucatán, Merida, Mexico

The progress of distance education (DL) technology has showed a significant growth due to its independence from educational paradigms based on space and time; Therefore, the efficiency of learning management systems and content is critical in the success of distance teaching and learning.

Learning management systems (LMS), are increasingly demanded by society as an alternative paradigm in learning systems, so their evaluation is really important to ensure the quality of the service [2]. In this context, it is necessary for the LMS to be created through the use of software engineering (SE) fundamentals and DL.

Given the existing diversity of solutions, it is necessary to carry out investigation in which the quality of learning management systems is the main actor. An automated approach to this analysis may be relevant to ease the evaluation of this process. Assessment can be assisted through the use of an expert system to suggest and highlight desired or identified features of a learning management system in particular.

Currently, however, no expert system has been reported that has been designed taking international standards DL and SE into account, which may help institutions and organizations make more accurate assessments of the use of LMS, thus no results have been produced that may help improve these services.

By using an IS approach and based on SE and DL models and standards, the expert system guarantees the quality of the assessments, since it follows a formal process and methodology in its implementation and operation.

Therefore, this research emphasizes the importance of evaluating the distance learning technological tools that facilitate course administration, the creating of materials by students and the control and monitoring of teachers' tools. All this in order that the LMS cater for current needs of the institutions, based on platform assessments and user perception.

Consequently, this paper addresses this issue by: (i) offering a literature review of the LMS, (ii) mentioning both SE and DL standards and norms on the LMSs, (iii) offering the expert systems descriptions, (iv) proposing the structure of the LMS expert system evaluator, especially its (v) traceability model, (vi) presenting the derived outcomes and finally the conclusions.

91.2 Learning Management Systems

In distance education, learning management systems play an important role. For this reason, the determination of their quality is relevant. The objective of LMS's is to seek the best technological solution for managing online courses through the web [3].

IEEE defines an LMS as an information system that offers the possibility of registering students, establishing the schedule of access of learning resources, controlling and guiding the learning process and analyzing the performance of while monitoring the same students [4].

An LMS is conceived as a system that is generally characterized by its ability to integrate the tools and resources that help manage, administer, organize, coordinate, design and deliver training programs through the Internet, activities it is necessary to carry out in order to achieve significant learning from the students [5].

The tendency of LMS's shows a service-oriented architecture [6], which facilitates their integration with other systems such as content management systems, for storing and exchanging resources.

Many institutions use LMS's to offer online training; mainly because they provide an environment that enables updating, maintenance and expansion of educational spaces, while favoring the collaboration of their many users. These tools also allow for the management of academic content [7].

In order to establish the quality of LMS's, it is important for them to show an efficient performance in the today use. In this sense, different institutions or organizations are now working in the development of standards and specifications for LMS in various aspects of their architecture, interface, functionality and development.

91.3 Software Engineering Standards and Distance Education

For the purposes of this proposal a review of the different standards related to the SE and the DL is presented. The following standards describe the specifications which have been considered for the expert system model.

91.3.1 Software Engineering Standards

ISO 15504 SPICE: is an open, international standard for assessing and improving the capability and process maturity together with ISO 12207. The standard is aimed at evaluating and improving the quality of the process of software development and maintenance [8].

CMM (capability maturity model): is a software quality model that classifies companies according to their maturity levels. These levels are used to identify the maturity of the processes followed to produce software [9].

MOPROSOFT (process model for software industry): is a model created to promote the improvement and evaluation of software processes used within the program for the development of software industry in Mexico [10].

91.3.2 Distance Education Standards

SCORM (sharable content object reference model): this specification is considered to be the most comprehensive instructional content sharing tool, and it's widely

used in the field of e-Learning. It is the result of work by the ADL (advanced distributed learning) initiative created to help standardize and streamline the distribution of educational resources and training [11].

The AICC-AGR 012: consists of a series of specifications developed by the AICC (aviation industry CBT committee) and it sets standards and technological recommendations to achieve better practices of LMS's as user-designed systems, hardware and software, icons etc. CBT [12].

IMS common cartridge, learning tools interoperability and learning information services: is a collection of specifications proposed by the IMS global learning. These specifications help develop and promote open access to facilitate online learning activities. Their aim is to create a format that will take the recommendations of IEEE and AICC into account [2].

IEEE 1484.11.1-4: This standard evaluates the architecture and code of the LMS's [13], emphasizing the use of its components and computer systems in education as well as in interoperability of the platforms. It was developed by IEEE LTSC (learning technology standards committee).

ISO/IEC TR 29163-4:2009: This standard covers the essentials of the code and architecture of the LMS. It gathers all the different standards of IMS, AICC and ADL in order to make specifications for the LMS's [14].

PROJECT OKI (open knowledge initiative): this project defines an architecture where components of the software environment communicate with each other and with other systems. The specifications enable sustainable interoperability and integration by defining standards based on a service oriented architecture (SOA) and the definition of interfaces, called open service (OSIDs) [15].

W3C standards: World Wide Web Consortium is an international community where (W3C) experts work together to develop Web standards and accessibility [16]. They ensure the availability of published resources on the Web, regardless of the characteristics of the access devices.

91.3.3 Expert Systems

Expert systems are computer systems that simulate the process of learning, memorization, reasoning, communication and action of a human expert in any area of science. Their aim is to emulate the behavior of an expert in a particular domain and to provide high performance expertise applied to particular situations [17].

Expert systems store the knowledge of specialists in a given field, and solve problems through the logical deduction of conclusions.

Expert systems specialist knowledge stored for a given field and solves problems through logical deduction of conclusions.

Although some investigators usually find a connection between expert systems and quality assessment, little is mentioned on this topic in most of the articles written in this area of expert system investigation; for example, Declan Dagger et al. [6] analyze the evolution and current challenges that LMS platforms must

address to achieve interoperability of information. In working on the iClass project, they present a system that sees a connection with the next generation of service-oriented LMS's; Expertus Training Industry, Inc. [18] conducted an investigation on the state of the LMS's. The survey was designed to help improve the understanding between teaching professionals and the degree of their LMS; also LMS main challenges, features and functions that are most important in order, to include the information obtained in future development of LMS's; Georgiakakis [19] proposed a system based on surveys. Since a learning management system based on criteria of usefulness and quality of use is being evaluated.

91.4 Expert System for the Assessment LMS (SEAMY)

The objective of this research project is the creation of an expert system which development was based on a traceability model to assist in the assessment of the quality of an LMS called SEAMY. System development involves the following of a research model that in turn involves the application of the following phases: (See Fig. 91.1):

- (1) Analysis of DL projects in order to choose the indicators or services to be used to evaluate the LMS.
- (2) Search and analysis of SE and DL standards to identify information related to the LMS; then it is classified and utilized in expert system design.
- (3) Analysis of the features and aspects of the indicators defined in the previous phase. This is to determine which indicators are to be included in the conceptual framework. The conceptual framework is also constituted by a group of indicators defined trying to obtain a better systematization; this in order to have the basis for the traceability model expert system.
- (4) Definition of a traceability model to calculate the average and instant frequencies of evaluation, taking as a base the assessment criteria of the different standards of SE and DL (described in detail in the next section).
- (5) Development of the expert system. At this stage, an expert system is developed, based on the conceptual framework defined above.
- (6) Validation of the proposal. At this stage reports of results must be written, and conclusions of the expert system model are presented to a panel of experts for their validation.

After seeking information of LMS features and functionality in software engineering standards and distance education, 290 indicators were selected that form the conceptual framework of all base theories.

Classification consists of the major categories as shown in Table 91.1.

Figure 91.2 shows the conceptual framework for expert system, which includes the following elements:

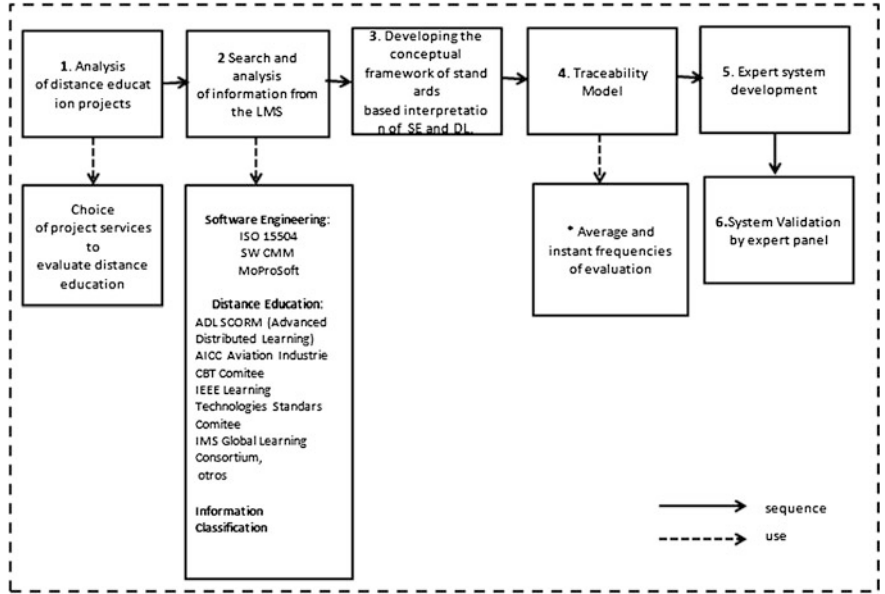


Fig. 91.1 Research model

Table 91.1 Classification of Main Categories

| Aspects | Indicators group |
|---------------------------|---|
| 1. Technical requirements | A. Hardware/software B. Price/license |
| 2. Management tools | A. Configuration tools B. Services tools C. Accessibility and compatibility tools D. Control tools |
| 3. Student activity tools | A. Communication tools B. Support tools C. Assessment tools |

- (i) The user interface expert system will be a SEAMY Web application that invokes the inference engine.
- (ii) The inference engine contains control rules of the SEAMY expert system, along with the traceability model. The traceability model provides a measure of evaluation.
- (iii) The knowledge base contains the basis of the standard theories of distance education. The base theories of the expert system are the questions of the group of indicators and of the indicators of the standards to be analyzed.

The user interface exchanges inference rules with the inference engine, depending on the traceability model templates. This model uses four Likert scale

Fig. 91.2 Expert system framework

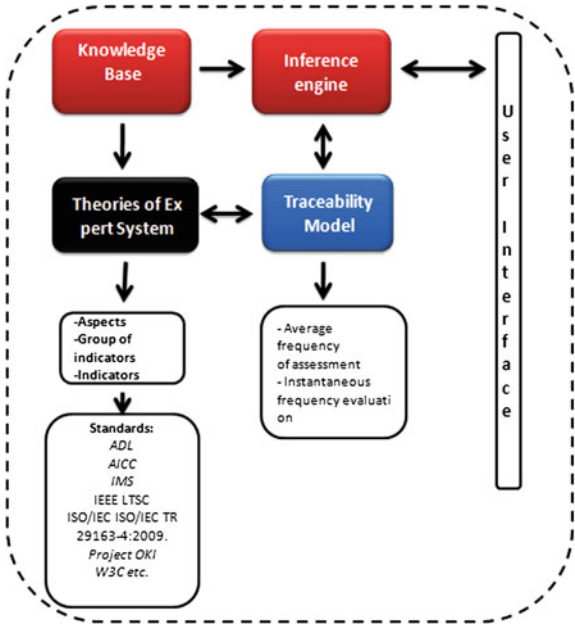


Fig. 91.3 Likert scale

| | | |
|--------------|---------------------------------|---|
| | <input type="text" value="25"/> | <input type="checkbox"/> Agree |
| Positive(+) | <input type="text" value="15"/> | <input type="checkbox"/> Strongly Agree |
| | <input type="text" value="25"/> | <input type="checkbox"/> Neither agree nor disagree |
| <hr/> | | |
| Negative (-) | <input type="text" value="18"/> | <input type="checkbox"/> Strongly disagree |
| | <input type="text" value="17"/> | <input type="checkbox"/> Disagree |

values (Fig. 91.3). Thus, the inference engine can exchange weights for measuring the indicators and templates based on theories of the expert system, in addition to a more accurate assessment given by the average frequency of assessment and the instant evaluation frequency.

The knowledge base is compose of the theories of the expert system, which highlights areas, groups of indicators and particular indicators of the standards used of distance education in the LMS. A very important element in the framework of the expert system is the traceability model.

91.5 Traceability Model

The formulation of questions is carried out taking into account the classification of information indicators that was obtained. The traceability model plays an important role because it helps obtain more accurate evaluations.

Table 91.2 LIKERT scale values in the traceability model

| | |
|-------------------|-------------|
| Agree | P4 = 1.0000 |
| Strongly agree | P3 = 0.6065 |
| Strongly disagree | P2 = 0.1273 |
| Disagree | P1 = 0.0101 |

The traceability model involves the Likert rating scale. This scale is the most frequently utilized. This scale combines sentences that express a favorable or unfavorable attitude toward the object of interest, making this a procedure for evaluating questions (items) [20].

A situation that occurs with the Likert scale has to do with the neutral answer “neither agree nor disagree”, and is associated with “moderate agreement” (See Fig. 91.3), which turns out to be an “undecided” value. The neutral response is not indicated, because users tend to move toward positive answers. In view of this, it was decided to exclude the value of undecided, and only to consider four Likert scale values (See Table 91.2). Thus, it is intended for the assessments to be more accurate.

The equation for the four values of the Likert scale is as follows:

$$\text{Gauss} = \frac{\gamma^2}{e^2 * \sigma^2} \tag{1}$$

Where: γ = Magnitude of the distribution σ = Is the question

The distribution of weights for each of the questions is listed on Table 91.2. The values of P1, P2, P3 and P4 were obtained from the Eq. (1) of the Gaussian function, taking as variables the distribution and magnitude of the question.

With the bell curve showing the distribution per share. Thus, depending on the values for the weights of the Likert scale for each of the questions, the average frequency is calculated, making an average of the different answers given by different users, and considering the instant evaluation frequency, the final addition of the current question is made in order to determine how the evaluation pattern advances.

Traceability model was necessary to perform a measurement LMS assessment, establishing a pattern of evaluation of the different questions of each user. The average of these values corresponds to the perception of the question of different users.

The embodiment of the questions on traceability model is asking questions in a positive and negative, this resolves the tendency of users to answer repeating a Likert scale value (“Response Set”).

The main idea of implementing a traceability model is to indicate a pattern of user evaluation. Also present was a perception of the LMS user, displaying the results graphically where the user can view their assessment and perception of other users for each of the questions.

Thus we can measure and quantify the value of each of the questions of the expert system and produce results more accurate assessments.

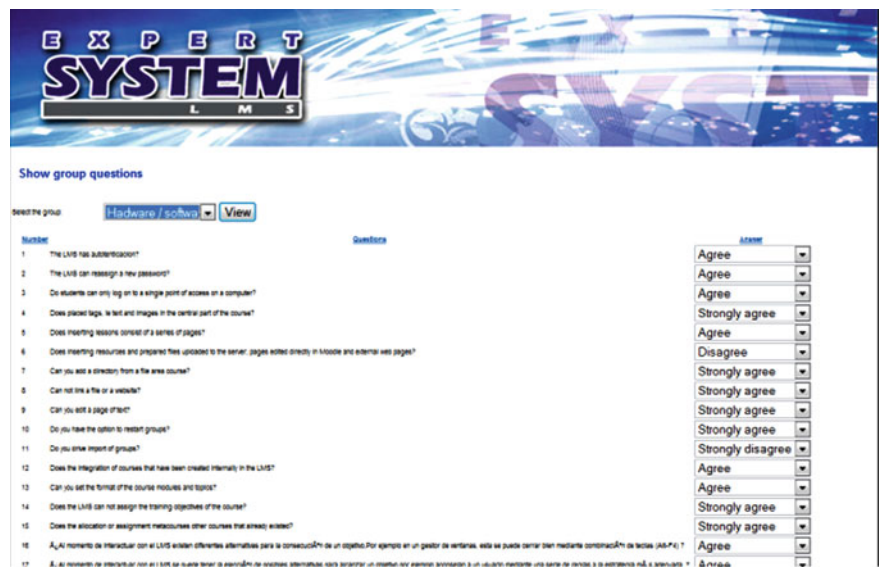


Fig. 91.4 Expert system web interface

91.6 Implementation

A prototype expert system LMS evaluator on web technologies has been developed (Fig. 91.4). The user uses a web interface to evaluate an LMS and get graphic feedback on the quality of the learning management system.

The following chart uses the allocation of weights for each of the questions, depending on the Likert scale. Its responses are classified according to their profile within the LMS: student, teacher, administrator or manager of LMS.

When a user answers a question, the SEAMY makes an average estimate of his/her response compared to the answers of different users. After, the user can view a graph showing the perceived response of others to the same question (Fig. 91.5). It also shows another graph with the current user's evaluation and the ideal pattern of each of the question or prompts.

The user accesses the web interface through a session where his/her profile has previously been uploaded. Next, the same user needs to indicate the educational institution he belongs to and his evaluation profile in order to start answering SEAMY.

The way the SEAMY questions are organized is according to the conceptual framework (see Table 91.1); so, all of the questions are organized by groups.

Once the questionnaire has been completed, the user can view a graph of his/her evaluation, the average of previous users' evaluations and the perception of other users.

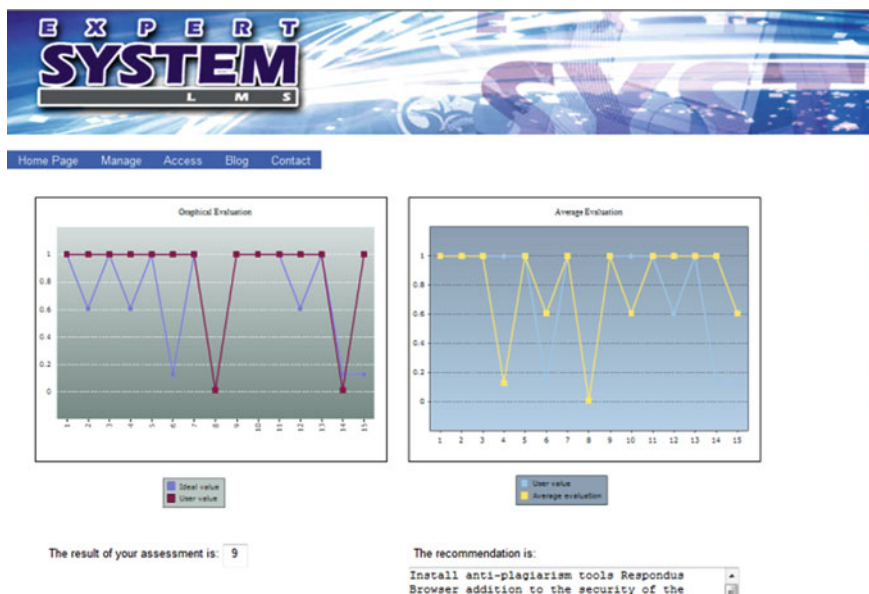


Fig. 91.5 Valuation graphics

The user can view an overall rating of his/her evaluation and a list of recommendations, depending on the questions that are critical to assist in making management decisions LMS.

An important aspect of the graphics is that they allow the user to visually identify how close his/her perception is regarding the views of other users. Even if quality assessments are subjective, the perception of the majority can be a relevant factor in decision making.

91.7 Conclusions

Quality in learning management systems is currently a recurring problem when it comes to choosing a good LMS, since there are many LMS available. Choosing the right LMS is a problem faced by managers every day DL.

In this paper, we have presented an expert system that has to be based on a traceability model, SE quality standards and distance education, to ensure that results will be more in line with what is being handled recently.

The proposal considers the perceptions of different users in the LMS in addition to an automatic component that optimizes, analyzes and proposes criteria for evaluating the system. It is noteworthy that this system allows researchers to contrast the different perceptions of LMS users and, thereby, establish how close or how dissonant a standard tool can be, and the opinion that users have of its use.

A prototype expert system has been developed that easily incorporates LMS evaluations and user perceptions. By contrasting the specifications or the LMS

evaluation question and what users perceive from other users' opinions of the specifications. This will combine the two factors of technology assessment.

As future work it is necessary for us to develop a module that is integrated with an LMS (Moodle for example), in order to get a concrete validation of the proposal; thus the information here generated can be integrated into the management system, contributing to the decision-making of the administration of an LMS.

References

1. Buontepo MP, Ortiz MC, Maricel Kler N (2010) La inclusion de la Educacion a Distancia en la Universidad Nacional de Nordeste: Cambio e innovacion en sus prácticas institucionales, Educacion a distancia actores y experiencias. CREAD 1:131
2. Naveh G, Tubin D, Pliskin N (2010) Student LMS use and satisfaction academic intuitions: the organizational perspective. Elsevier, Israel
3. Epic (2010) Open source learning management systems, by Mark Aberdour, technical producer, pp 30
4. IEEE learning technology standards committee (2011) Accredited learning technology standards, USA. <http://www.ieeeltsc.org/>
5. Diaz Antón G, Pérez MA (2011) Hacia una ontología sobre LMS la Universidad Simón Bolívar, Venezuela, pp 1
6. Dagger et al (2010) Service-oriented e-learning platforms: from monolithic systems to flexible services. Distance Learn IEEE Internet Comput 11:28–35
7. Boneu, Joseph M (2010) Plataformas abiertas de e-learning para el soporte de contenidos educativos abiertos, Revista de Universidad y Sociedad del Conocimiento (RUSC), Universidad Oberta de Catalunya vol 4. España, pp 36–47
8. Steinmann C, HM and S, ISO 15504 (Spice), Employee Motivation and information using Spice The Road to Software Process Improvement, 2010
9. Phillips Mike (2002) CMMI: improving processes for better products. Lecture notes in computer science, SpringerLink 2559:1
10. Hanna Oktaba et al (2005) Modelo de Procesos del Software, Moprosoft, Version 1.3, pp 123
11. ADL (2011) Advanced distributed learning, SCORM 2004, USA. <http://www.adlnet.gov/Technologies/scorm/default.aspx>
12. AICC (2011) Aviation industry CBT comitee, About the AICC, USA. <http://www.aicc.org/>
13. IEEE (2011) IEEE learning technology standards committee, accredited learning technology standards, USA. <http://www.ieeeltsc.org/>
14. ISO (2011) ISO/IEC TR 29163-4:2009, USA. http://www.iso.org/iso/catalogue_detail.htm?csnumber=53537
15. Lopez Lopez JF (2009) Iniciativa Del Conocimiento Abierto y su Aporte Al E-Learning Tanto En Objetivo Como En Su Proceso. <http://telecsys.unad.edu.co/documentos/revista%20No.2/articuloFernando.pdf>
16. W3C (2011) Consorcio W3C. <http://www.w3c.es/Consortio/>
17. Riley G (2011) Sistemas Expertos: Principios y Programación, 3rd Edn. p 315
18. The current and future state of learning management systems. Expertus and Training Industry, Inc. pp 1–11, 2010
19. Georgiakakis P, Papasalouros A, Retalis S, Siassiakos K, Papaspyrou N (2005) Evaluating the usability of Web-based learning management systems. THEMES Educ 1(6):45–59
20. Cooper DR, Schindler PS (2001) Business research methods, 7th Edn. McGraw-Hill, New York

Chapter 92

How Variability Helps to Make Components More Flexible and Reusable

Yusuf Altunel and Abdül Halim Zaim

Abstract Cross-project and cross-market reusability is the basis for the existence of a component-based development paradigm for the establishment of a software industry based on reusability and conservation of expertise. In this paradigm producer is in need to enhance components' functional capabilities and increase the audiences for successful marketing, so they tend to add extra features to widen the reuse spectrum. However, this in return is recompensed by the component consumers in means of higher prices for unneeded features, overflowing functionality and superfluous code. The component producers need more flexibility to generate specific patterns of components to deal with the changing requests of consumers. In this paper we provide a basis for more flexible components.

92.1 Introduction

Object-oriented technologies have very-well understood advantages but they are weak in reuse development, runtime performance, adaptability, management of complexity, and performance [1]. Therefore a more reuse oriented approach is

The study is supported by NodeSer, Istanbul.

Y. Altunel (✉)

Department of Computer Engineering, Istanbul Kultur University, 34156 Istanbul, Turkey
e-mail: y.altunel@iku.edu.tr

A. H. Zaim

The Institute Of Graduate Studies in Science and Engineering, Istanbul Commerce University, Ragıp Gümüşpala Cad. No: 84 Eminönü, 34378 Istanbul, Turkey
e-mail: azaim@iticu.edu.tr

needed and component-based approaches can provide such a shift to realize the common and systematically enforced way of reusability both for in-house and cross-market development. Reusable component design requires abstraction, hiding, functional independence, and refinement activities [2]. Designing components with variation points and variants can increase the component's intended reuse [3].

The main concentration of component-based development is to make reuse of the software instead of developing every piece of code from scratch [1] and establish a software production industry with maximized reusability instead of building everything from scratch. Common principles to reach to this goal are standardization, easy integration, conservation of expertise, reduction of development complexity, and enhancement of maintainability. Accordingly, component-based development can help the organizations to save development time, reduce both time to market and development costs, simplify the system production, and provide the ability to outsource the technical expertise.

Two parties appear in new division of labor in industrial structure: component producers and consumers. Producers can be professionals, internal units or external organizations, specialized in certain technical abilities and practicing technical details of the domain. They will gain the experiences, knowledge and best-practices to generate the components with smart solutions otherwise consumers should make investment, lose time, and deal with complexities.

In such paradigm, the consumers require the confidence, reliability, efficiency, and compliance to standards additional to guaranteed support; otherwise the risks are not worth of the potential gains. This sensitive balance can be established, if the producers are able to develop components to satisfy both the common and specific expectations of the consumers at the same time. The producer should be able to develop products with various capabilities to solve as many problems as possible so that its profitability is maximized. The strategy seems to work only if components are filled with extra features so that even very unique desires of consumer can be satisfied. However, this makes the components full of features that are not required in many of reuse options.

Producer should be able to define and measure the reuse potentials of a component to make better analysis and decisions at different levels from production to the marketing. Actually the consumer side is also in need to explore capabilities of components to make purchase or not purchase decisions. One way to accomplish this mission is to create formal specifications of component's structural characteristics so both parties can get some insights about its abilities and potentials. Such a formalism should be capable to categorize component features according to the reuse options to identify the features always required (mandatory), sometimes needed (optional), only be used interchangeably (exclusive), multiply occurring (repetitive), and depending on the others (dependencies).

In this paper we provide a basis for component variability, which is a critical point in understanding and measuring the component reuse potentials. Variability is studied in product-line approaches but we try to look to the issue in a broader sense but in limits of software components.

The paper is organized as follows: We start with the explanation of formal approaches and the importance of formal specification of components. After that, we provide the definitions and discussion on the variability as a quality attribute. In the following section, we explore the attributes of variability. Next, we provide a literature study on the variability issue. Final section is about conclusive remarks and the further studies.

92.2 Formal Specifications and Components

Formal systems are widely used by traditional engineering industries and accepted as the core behind the whole engineering processes. They are also accepted as useful in software development as they have the potential to enhance the quality and simplify the software verification against its specification [4, 5] but not ‘practical’ and ‘non-cost effective’ [5].

Mathematical foundations can help the component-based software to enhance the functionality, prevent ambiguity, and simplify the process of integration. Additionally, formal systems are useful in search and identification of correct components as well as improvement of component evaluation. The formalisms are also helpful to generate code out of specifications, assess the satisfaction level of non-functional characteristics, produce dynamically changing software, define and check the conformance to the architectural reuse configurations [6].

Components, as they construct hierarchically structured composition structures, can become quite complex. So, production of a component requires the act of exploration of such a composition hierarchy which can be done by applying formally defined set of activities recurrently [7]. Such a work should deal with revealing the width, depth, and length of decomposition. Depth is the number of levels in decomposition process, width is the (average) number of sub-components for each parent, and the length is the number of variants produced for each component [8].

Formal approaches help in specification of components and analysis of their reuse potentials. This gives the consumer the opportunity to understand how easy to add a component to an application. Formal approaches such as algebraic descriptions can be useful to create component specifications for multi-reuse and properly define individual variants, so that the consumer can produce specific components optimized for special usages, additional to generic ones.

92.3 Variability as a Quality Attribute

‘Variability’ which is the noun form of ‘variable’ is defined as a quality or state of varying, changing; or alternation [9]. According to Bachmann & Clements, variability is defined as the ability about an ‘asset’, system, or development

environment to produce artifacts which are different from each other. This definition is further adapted to the scope of product-lines as the ability of a ‘core asset’ to adapt to usages in different product contexts [10]. Another definition is given by Svahnberg et al. as the ability to change, customize and configure software in a particular context [11].

The variability can be accepted as a quality attribute and qualitative measure to clarify the capabilities of a component in satisfaction of varying needs. Functional variability is relatively easier to realize since no architectural modification will be required comparing with other quality attributes like performance, security or maintainability [12].

The component variability study has practical results to identify the configuration options of a component. Additionally, the components can be analyzed to understand whether specific needs can be fulfilled or not. The study helps to comprehend how to increase the reuse options of a component to cover more functional and non-functional requirements. Finally, the variability study shows the potential of reducing overall costs, complexities and time of development providing the information about the quantitative properties of reused assets in component production as they are or with minor modifications.

92.3.1 Variability by Product

Single entity-based variation is usually maintained by means of configuration options. A single entity rarely provides ‘rich’ functional and non-functional variation potentials. Variability is ordinarily maintained by almost all software intensive systems and software is rarely designed rigid, i.e. providing at least some configurability. So product variability or the variability of a single product is the natural way of software implementation, although it is rarely identified and measured.

When the components are the case, inheritance is a questionable approach and should be used with care. Instead, composition seems to be a better mechanism so that components can be deployed individually. With this approach, component variants can be generated out of sub-components. The variation capacity can be measured as the capability of component abilities in production of various variants.

92.3.2 Variability by Product Set (Families)

If the expectations are high and various applications are aimed, a set of components sharing the commonalities can provide a high capacity for variation. In object-oriented approach, a base class with derivation options can provide such a potential.

Diversity of customer needs direct many software product companies to develop software products with different capabilities instead of single products with maximum capabilities [13]. The ‘family of products’ can be defined as the set of products or applications having commonalities or interrelated with each other. According to this definition the product family is much more than “the assets that are produced in the same product-line” as defined by product-line approaches. So, in reality the challenge can be stated as how to define, implement, maintain and measure the commonalities and differences between the software systems.

92.3.3 Functional Variability

Functional variability can be defined as the diversion in the functionality or variation in functional characteristics of a component. For example a Math component can provide the calculation of pi value, square root of a number, and power of functionalities to satisfy various mathematical computations needed by other parts of the software.

At the syntactic level, functional diversity is usually about the functional abilities that a component can serve. In this manner, number of methods, attributes, and sub-components are direct counts. Additionally, the provided interfaces of a component can be accepted as the functional variability level of a component for integration.

92.3.4 Non-Functional Variability

Component’s potential can vary for non-functional properties such as performance, reliability, portability, security, etc. A component with adjustable level of abilities can be useful to satisfy various non-functional conditions. So, non-functional variability can be defined as the ability of a component to provide different levels of non-functional abilities when the reuse conditions are changed. Hence, it will represent the varying capabilities and usefulness of the component that can be satisfied under different conditions and environments.

92.4 Time of Variability and its Attributes

Certain quality attributes are enhanced by means of the variability level of the component. Variability helps to enhance the flexibility and reusability of a component. Flexibility as a quality attribute can be maintained at the architecture level to make components handle various application differences. Flexibility is much more about the usage of the component, so component should be designed and implemented so to handle the flexibility.

Table 92.1 Attributes and time of variability

| Variability by | Party | Time of Variability |
|----------------|----------------------|--|
| Adaptation | Producer | Design Implementation Maintenance |
| Configuration | Consumer | Integration |
| Customization | Producer | Implementation |
| Evolution | Producer | Design Implementation Maintenance |
| Extension | Producer Consumer | Design Implementation Pre-integration |
| Modification | Producer | Implementation |
| Portability | Producer | Design Implementation Maintenance |
| Scalability | Producer | Design Implementation |
| Suitability | Producer Consumer | Design Implementation Pre-integration |
| Tailoring | Producer Consumer | Design Implementation Pre-integration Integration |
| Upgradability | Producer Consumer | Implementation Maintenance |

Component variability is affected by certain other quality attributes such as adaptability, configurability, customizability, suitability, etc. We collected a set of such attributes that seem directly related in realizing the variability. The variability attributes, responsibilities, and the time of variability are presented in Table 92.1.

92.4.1 Adaptability

A component is adaptable if it is designed to easily change its internal properties and external capabilities under unanticipated situations. The configuration/reconfiguration ability is one of the easiest ways to make a component adaptable. However, component internals can be restructured to make adaptation and such a work might require reworking the component implementation life cycle. This process can be optimized by means of automated generation of component variants. So, whenever adaptation requirement is required a proper variant of component can be generated according to the description of the component internals. Consecutively, variability by adaptation can be realized by the producer side and accessing component internals at implementation level.

92.4.2 Configurability

Configuration is the easiest way to adapt the component and make it useful for different types of applications. Any reusable entity is expected to be deployed with some certain levels of configurability so that it will fit the application's specific conditions. Variability by configuration therefore is a common and straightforward operation normally done by the consumer side.

92.4.3 Customizability

Customization is defined as the process of changing the internal properties according to the changing needs of customer and can be done both programmatically, and through customization interfaces [14]. Without customization a component is rarely reusable, so it is one of the basic ingredients of reusability.

Components are black-box units and internal ingredients are not accessible to modification of third parties. However, customization requires accessing and making changes on component internals. Therefore customization is done by the producer side and it requires certain component production life cycle activities. So variability by customization is normally maintained at implementation level and realized by the producer side.

92.4.4 Evolvability

Evolution is not an ordinary case to come across, so components are rarely designed and implemented as evolvable. Evolution requires the self-regulated adaptation of the component according to the changing conditions over time. Such a component should be designed flexible to add, delete and change component features even after deployment. So, variability by evolution is tricky and requires advanced techniques. Even if it can be implemented, it should be done by the producer at the time of implementation.

92.4.5 Extensibility

The components can be extended by means of object oriented derivation, aspects, and automated code generation techniques. Component extension requires knowledge about component internals but no modification of internals is needed. Normally, extension is realized by the consumer at pre-integration time. However, the component should be designed properly for extension and it is a design issue of

the producer. So variability by extension is the issue of both producer and consumer at the same time.

92.4.6 Modifiability

This shows how easy to modify the component for various reasons such as maintenance, bug fixes, enhancements, etc.

Modification on component internals can result in side effects that should be carefully managed. Proper ways of modification is addition, deletion, and modification on component features. After the modification, the system using the component should be analyzed for its stability. So compatibility with previous versions should be managed. However, variability by modification is the issue of the consumer at the implementation time of component.

92.4.7 Portability

Portability is the ability to operate in the same manner but in different conditions. Normally, components are created for pre-specified environment which can be an operating system, development environment, framework, middleware, etc. Therefore, if the environment will be changed major modifications are expected over the component internals additional to the style of development.

92.4.8 Scalability

A scalable component can provide services to accelerating requests with the ability to be enlarged to accommodate such growth. Scalability requires certain design specialties at the architecture level and it falls within the producer's responsibilities. So, variability by scalability is maintained by the producer at the implementation time of the component.

92.4.9 Suitability

Scalability represents how a component can meet stakeholder's needs. Component's unsuitable parts can be enhanced by means of other variability mechanisms such as configuration, customization, extension, etc. Accordingly, suitability level shows whether the component is in need of such mechanisms at the time of component selection. A more suitable component can be selected according the

results of such study. Similarly, suitability can be enhanced by the producer by making adaptable, configurable, customizable, etc. components. Therefore, suitability is also related with the implementation time at the producer cite.

92.4.10 Tailorability

Tailorability of a software component specifies how a component can be modified when its capabilities are insufficient, inappropriate, or not covering the needs in acceptable fashion. Tailoring might require some combinations of adaptation, configuration, customization, and extension at the same time so that an unsuitable part is replaced or a gap in component properties is properly filled. As easily can be seen, tailorability requires both producer and consumer side management covering different levels development activities.

92.4.11 Upgradability

A component can be upgradable if it is replaceable and backward/forward compatible. An upgrade is done to make an enhancement, add some features or remove some parts within the component. This is performed by the producer by implementation. However, consumer makes the replacement at the maintenance time to make use of the enhancement comes with the component.

92.5 Variation Capacity as Component Variability Measure

We provide a measure for producer's side called *Variation Capacity* (VC) to understand component abilities in variant production. It is based on counting the number of mandatory, optional, exclusive, repetitive, and dependent features. Mandatory and repetitive features have no effect on the calculation since mandatory features are always included and repetitive attributes can be implemented as single entry by the help of data structures such as array, linked list, queue, or collections. The actual variation capacity depends on the optional, exclusive and dependent features.

The variation capacity of a component $VC(C)$ is the multiplicative order of the unique optional and exclusive feature combinations. The number of optional feature combinations (op) is of 2's exponent of the number of optional features. The dependencies between optional features restrict the possible number of unique combinations since dependent features should stay together.

Dependencies, reduces the possible independent combination options of features. If dependency is between the optional features, they should be combined as a single combination.

The number of exclusive feature combinations is a double of exclusive features (*ex*). Dependency between the exclusive features is a contradiction, since features cannot be both exclusive and dependent at the same time.

After these discussions we can provide a formula to evaluate a component's variation capacity $VC(C)$ as in 92.1.

$$VC(C) = 2 \text{ ex } 2^{(op-dp)} \quad (92.1)$$

92.6 Literature Study

We provide the approach to the variability in other approaches including the product-line architectures, feature oriented programming, aspect oriented approaches, and generative programming.

One way of handling the variation is to introduce 'variation points' to provide alternative functional and non-functional features such as introducing different formulas of tax calculations, different persistency mechanisms and so on. There are some risks in variation points missing or unnecessary variation points with superfluous complexity because of over-generalization. Variation points are typically implemented using design patterns and reuse in class libraries is done by introducing or refactoring classes. In return this increases the complexity of design process, and results in performance degradation due to dynamic binding of additional levels for indirection which remains at runtime [1].

Software product lines or software product line development in software world is the collection of software engineering methods, tools and techniques to develop software products out of common set of software assets by means of common product-line. The main difference from other opportunistic reuse is that the product-lines enables systematic reuse [15] of software artifacts in one or more products rather than putting the components into libraries and leaving them to the arbitrary reuse by chance [10]. The approach has the potential to make a jump of competitive advantages as happened in adaption of mass production and mass customization paradigms in manufacturing.

In product-line, a single and coherent development activity is undertaken to build a set of products from core asset-base which is a collection of artifacts specifically designed for use across the "portfolio". Portfolio might be the architecture and its documentation, specifications, software components, tools such as component or application generators, performance models, schedules, budgets, test plans, test cases, work plans, and process descriptions. Product-lines require describing and using the variability at a level, as the products are generated out of reusable assets, which should have a certain level of variability. Features are the key points in implementation of variability in product-line approaches.

Feature Oriented Programming (FOP) or Feature Oriented Software Development (FOSD) is defined as a general paradigm for program synthesis in software

product lines. ‘Feature’ here is defined as a reflection of a stakeholder’s requirement incrementing in functionality, which is distinguishing elements for variants of program or software system [16] cited in [17]. Features are the modular domain concepts that can be translated into codes by means of ‘feature modules’ to enable code composition in incremental generation of programs. Features are implemented by a whole set of classes to cooperate in completion of task rather than single classes. The features are used to develop different compositions to get different programs [9].

AOP is for localizing, separating, and modularizing crosscutting concerns. Aspects as the main abstraction mechanisms in Aspect Oriented Programming (AOP), encapsulates code scattered across the implementation of concern and help to achieve the separation of crosscutting concerns in building complex software. A concern is an issue or problem which is interesting to stakeholder and rather related to the features. Crosscutting concerns are special since they do not reflect hierarchical or block structures of other concerns in program [9].

Aspects and features are two different decomposition types of abstractions. Aspects are small units used for few concerns otherwise their implementation lead to scattered code, whereas feature is an increment in program functionality. Feature and aspect-based decompositions have different intention and language mechanisms. Aspects can affect code with several features and feature modules can contain code from several aspects. Aspect-based and feature-based decomposition can be integrated to lead to decomposition of software in three dimensions: classes, aspects, and features [9].

The Generative Programming (GP) is the discipline in the field to fulfill specific requirements by generating specialized and highly optimized systems. The motivation behind the GP is to decrease the gap between the code and domain concepts, achieve high reusability and adaptability, simplify management of component variants, and increase both space and execution time efficiencies. Generative Programming is identified as a software development paradigm to achieve intentionality, reusability, and adaptability by maintaining runtime performance and computing resources [1].

The GP provides a diagrammatic representation of the feature relationships, which is called as “feature diagram”. In this means, mandatory, optional, alternative (exclusive), and or-features can be represented using these diagrams. Simple edges with filled ending with filled circles represent the “mandatory features”, simple edges ending with empty circles are for “optional features”, edges connected by an arc are used for “alternative features”, and finally edges with filled arcs are used for “or-features” [16].

The aim of variability in product-line is to build and maintain products over a specified period of time to maximize the return on investment (ROI) and it is supported with the “variation mechanism” which helps to control the adaptation and support of product-line development for benefiting from the similarities that exist between similar applications. Consequently it becomes a switch decision to make control of the inclusion of one or the other component, and this is quite

acceptable rather than comparing with recoding or copying thousands lines of code to make the reuse of the component.

Maintaining a single component with the adaptability for a range of variation can be obtained by owning versions of the component, and this is how the variation mechanism is implemented. A core asset can be developed applying the “major” activities such as identification of cores assets that will remain the same for all products, choosing the variation mechanism to support the variation requirements, and providing instructions to describe how to use the variation mechanisms of core assets [10].

Component generator produces a component as a product asset out of its specifications. A software component is created as a single asset for particular product by using the user’s manual and tool to analyze the generated code, additional to the generator [10]. Aspect-oriented programming (AOP) on the other hand makes it possible to create variants of methods by automatically injecting code to enable satisfaction of non-functional characteristics.

Our approach has similarities with the GP with some minor differences in operator definitions. Additionally, we use the “feature” as a concept to describe component’s ingredients rather than “an increment in program functionality” to be implemented by a set of classes cooperating to complete a class. Usually, features extend a program by adding several new classes and by applying several new roles to existing classes simultaneously. Hence, the implementation of a feature cuts across several places in the base program [9]. However, a feature in our algebraic approach represents a single attribute, method, or sub-component which is a sub-atomic unit of the component or its variants.

Finally, component variability measures seem not so popular estimates in the literature [18]. In this paper we presented such a metric called *variation capacity* based on the previous studies of component variants.

92.7 Conclusions

In this paper we explored the variability concept to increase the flexibility of components and increase the reusability potential of a component. Variability requires producer–consumer type of division of labor between two different parties, even if both parties are the member of the same organization and this happens when components are produced for internal usage. Each parties has own responsibilities and expectations to gain the intended benefits. The consumers in this structure require the confidence, reliability, efficiency, and compliance to standards additional to guaranteed support; whereas the producers will gain conservation of expertise, experiences, knowledge and best-practices to generate the components with smart solutions.

The variability can be maintained by various non-functional attributes. In this paper we provided a bucket of such attributes and defined the responsibilities of producer and consumers plus the time of realization. Additionally we proposed

variation capacity as a measure for producer's side to understand the component capability of in production of variants. This measure can easily be calculated if the optional, exclusive and dependent features are explored. As a feature study, we work on other estimation techniques of the component variability taking into the consideration of other attributes.

In this paper, we also presented other approaches to the concept of the variability to illuminate the similarities and differences with our approach.

The variability level provides a very strong impression about the reusability options of a component, which can be used to estimate the potential market of the component. The consumer can use the measure to understand whether the component can properly be integrated with or without tailoring or customizations. Accordingly, the results help the consumer better understand the real value of component.

References

1. Czarnecki K (1998) Generative programming principles and techniques of software engineering based on automated configuration and fragment-based component models. A dissertation submitted in partial fulfillment of the requirements for the degree of Doktor-Ingenieur, Department of Computer Science and Automation, Technical University of Ilmenau
2. Muller PA.; Instant UML; Wrox Press, Canada 1997
3. Kazman R (2001) Software architecture. In: J.T. Yao (ed) Handbook of software engineering and knowledge engineering, Vol 1. World Scientific Publication, Singapore
4. Pressman RS (2001) Software engineering a practitioner's approach, 5th edn. McGraw Hill, New York
5. Sommerville I (2006) Software engineering, 8th edn. Addison-Wesley, Harlow
6. Dimov A, Ilieva S (2004) System level modeling of component based software systems. International conference on computer systems and technologies—CompSysTech '2004, pp II 7–1
7. Altunel Y, Tolun MR (2007) Component-based software development with component variants. The IASTED international conference on software engineering as part of the 25th IASTED international multi-conference on applied informatics, Innsbruck, Austria, 13–15 Feb 2007
8. Altunel Y, Tolun MR (2007) Component-based project estimation issues for recursive development. In: Sobh T (ed) Proceedings of CISSE 2007 international conference on systems, computing sciences and software engineering (SCSS 07), Later published in advances in computer and information sciences and engineering, Springer, 2008, ISBN: 978-1-4020-8740-0, pp 577–581, 3–12 Dec 2007
9. 14.09.2011,14:37:<http://www.webster-dictionary.net/>, <http://www.webster-dictionary.net/definition/variable>
10. Bachmann F, Clements PC (2005) Variability in software product lines. September 2005, Technical report, CMU/SEI-2005-TR-012, ESC-TR-2005-012
11. Svahnberg M, van Gurp J, Bosch J (2005) A taxonomy of variability realization techniques. In: Centre IBM (ed) Software—practice and experience, vol 35, no 8
12. Myllärniemi V, Männistö T, Raatikainen M (2006) Quality attribute variability within a software product family architecture. In: Conference on the quality of software architectures (QoSA) 2006

13. Myllärniemi V, Raatikainen M, Männistö T (2007) KumbangSec: an approach for modelling functional and security variability in software architectures. In: Proceedings of first international workshop on variability modelling of software-intensive systems, VaMoS 2007, Limerick, Ireland, 16–18 Jan 2007, pp 61–70
14. Sharma A, Kumar R, Grover PS (2005) Classification of metrics for component-based systems. In: Proceedings of the national conference on frontiers in applied and computational mathematics (FACM-2005), Allied Publishers Private Limited, Punjab
15. Course web page: CS 415—software product line engineering, Instructor: Dr. Bedir Tekinerdoğan, <<http://www.cs.bilkent.edu.tr/~bedir/CS415-SPLE/>>. Accessed 17 May 2011
16. Kang K, Cohen S, Hess J, Novak W, Peterson A (1990) Feature-oriented domain analysis (FODA) feasibility study. Technical report CMU/SEI-90-TR-21, SEI, CMU
17. Apel S, Lengauer C, Möller B, Kästner C (2008) An algebra for features and feature composition. In: Meseguer J, Roùsu G (eds.) AMAST 2008, LNCS 5140, pp 36–50, Springer-Verlag, Berlin
18. Sharma A (2009) Design and analysis of metrics for component-based software systems. Ph.D Thesis, Thapar University School of Mathematics and Computer Applications, Punjab

Chapter 93

Computing and Automation in the AEC Industry: Early Steps Towards a Mass Customized Architecture

Neander Silva, Diogo Santos and Ecilamar Lima

Abstract In this paper we demonstrate through examples and comparisons that digital fabrication is starting to produce impact in the Brazilian architecture, towards mass customization, not only through some exceptional buildings, but also through small experiences involving ordinary design needs.

93.1 Introduction

Some of the most important applications of computer science in the architecture, engineering and construction (AEC) industry have been in the fields of representation and fabrication.

According to Mitchell [1], representation is the “creation and manipulation of signs—things that ‘stand for’ or ‘take place of’ something else”. We represent, for example, spoken language through writing, calculations with numbers and artifacts through scaled models and drawings.

Representation plays a double folded role in the AEC industry: firstly, it is an essential part of the design thinking process. It is not possible to fully design without resorting to a concurrent representation system [2]. Secondly, representation is a means of design communication to clients and builders [3].

Computer systems initially represented low level entities such as lines, arcs, polygons and planes. As hardware became more powerful and capable of processing

N. Silva (✉) · D. Santos · E. Lima

Laboratório de Fabricação Digital e Customização em Massa, (Laboratory of Digital Fabrication & Mass Customization), Faculdade de Arquitetura e Urbanismo, ICC Norte, Ala A, Subsolo, Sala ASS589/9, Universidade de Brasília, Distrito Federal, 70910-900 Brasília, Brazil
<http://lecomp.fau.unb.br/>

more complex algorithms, new applications were designed for representing higher level entities such as three-dimensional generic solids. As processing power continued to grow, it is now possible to represent specific object-oriented construction components and subassemblies. These computer representations encapsulate not only geometry, but also properties, behavior and inter-relationships in what came to be called Building Information Modeling (BIM).

However, these applications have been often misunderstood by some architects. For instance, the issue of representing architecture through computer systems is frequently regarded as a peripheral practice, with no implications, either in the design process or in its product. This point of view is generally based on the assumption that design thinking and design representation are two distinct and sequential processes. However, as we already pointed out, design cannot take place without a representation system.

In a contradictory statement, those who hold to this view, very often also hold to a kind of 'drawing worship', in which hand drawing is considered an eternal and irreplaceable representation system [4].

However, the profession of architect, as we know it today, the one who solely designs for others to build, is relatively new in the course of human history [5]. Not much more than 500 years have elapsed since it came into being replacing the master builder by the end of High Middle Ages. This represents roughly less than 10 % of the time elapsed since the invention of phonetic writing, around 4000 BC, which marks the beginning of human history.

In the same way it is with designing by drawing and by drafting, a system which arose together with the profession of architect as we know it today. It was this system that allowed master builders to progressively distance themselves from construction sites and to become solely designers [5].

Nevertheless, many seem to consider and to act as the profession and its closest counterpart, the drawing/drafting system, had both always existed. Consequently, many professionals and teachers think that drawing and drafting have always had a central role to play in designing and building [6].

However, the historic evidence does not support these views. Before High Middle Ages, master builders used many different ways to represent their ideas and to have them materialized. Drawing was just one of them and it was not the most important [5–8]. If drawing and drafting have not always been the prevailing way of designing, there is no need to believe that they should play this role forever. Also, the advent of interactive three-dimensional computer modeling seems to be challenging those views. New representational systems can and have been brought to play innovative roles in the design [9] and in the construction processes [10].

The construction industry has been based to this point in time in mass standardization. The produced components are generic elements that will be customized later in the life cycle of the product. The mass produced components are classified into specific categories and produced in a limited array of forms and sizes. They are then stored, indexed and catalogued until they eventually, if sold, end up in a combination of elements in a factory or as a part of a building in the construction site [11].

Computer resources that allow the computerized manufacturing of artifacts directly from three-dimensional virtual models came to be called digital fabrication [10]. This process allows the production of construction components through computer controlled machines. These components may be produced by order without the need for an indexing system and shipped straight to the construction site. Therefore, substantial savings are made with labor, transportation, storage and cataloging systems.

As a new paradigm, the mass customization provided by digital fabrication allows that construction components may be produced for specific purposes, to become singular elements in unique contexts of specific buildings. The savings obtained in the automation of this process mean that the costs of unique components are hardly above those of the old standardized ones [12–16].

93.2 Research Problem

Digital fabrication technology is already available in emergent economies such as Brazil as it was demonstrated in earlier works [12]. However, has it had any impact towards mass customization in the Brazilian architecture? If so, did it find application just in exceptional buildings or has it found use in more day-to-day design needs?

93.3 Hypothesis

We believe that digital fabrication is starting to produce impact in the Brazilian architecture towards mass customization, not only through some exceptional buildings, but also through small experiences involving ordinary design needs.

93.4 Research Method and Results

We demonstrate the above hypothesis through surveying and describing a number of examples of architectural and interior design works from Brazil which have made use of digital fabrication.

We also explore some design works from one of the authors of this paper in which this technology was and is being used in the production of architectural and interior design components. At least one of these experiences reveals that digitally fabricated components can even result in a lower cost than those mass standardized ones.

Several examples of digital fabrication were found in the central region of Brazil, the area in which our research has been taking place. Some of them have already been mentioned in previous works [12], particularly in the area of arts such



Fig. 93.1 “Sphere”, by artist Darlan Rosa, DF, Brazil. (Source: authors’ photography)

as that shown in Fig. 93.1. The so called “spheroids” by artist Darlan Rosa (<http://darlanrosa.com/>) were one of the earliest examples of digitally fabricated artifacts in our region.

However, the application of digital fabrication is no longer limited to the area of arts, but has already found broader use in the construction industry. Figure 93.2 shows the entrance of a pedestrian overpass of the Metro System of Brasília, structural design by Márcio Buzar, which was digitally fabricated and assembled by CPC Estruturas, a company based in the Federal District of Brazil (<http://www.cpcestruturas.com.br/>).

Figure 93.3 shows an external view of the same pedestrian overpass which connects a Metro station to a major shopping mall.

Other developments were found in the area of interior design. Figure 93.4 shows the front façade board of a shopping in Brasília, designed by Diogo Santos, which was cut with a CNC plasma cutter by the local company Ferro e Aço Badaruco (<http://www.badaruco.com.br/>).

Figure 93.5 shows a stand for the telephone company Oi with curved wooden benches and counter, designed by Diogo Santos, which were also CNC cut.

However, the most important example of digital fabrication are shown in Figs. 93.6 and 93.7 because it illustrates the viability a potential impact of mass customization. Figure 93.6 shows a staircase mass produced by Ferro e Aço Badaruco (<http://www.badaruco.com.br/>) for which the cost estimated by the manufactures themselves was US\$3,806.00.

Figure 93.7 shows a custom staircase, designed by Diogo Santos, for the same purpose as the previous one, for which the estimated cost, to be digitally fabricated



Fig. 93.2 Pedestrian overpass—metro of Brasília, structural design of Márcio Buzar (Source: <http://www.cpcestruturas.com.br/>)



Fig. 93.3 Pedestrian overpass—metro of Brasília, structural design of Márcio Buzar (Source: <http://www.cpcestruturas.com.br/>)



Fig. 93.4 Front façade board of shopping in Brasília, designed by Diogo Santos, CNC cut steel (Source: authors' photography)

Fig. 93.5 Stand of Oi telephone company, with curved wooden benches, CNC cut, designed by Diogo Santos (Source: authors' photography)



Fig. 93.6 Comparative study: mass standardized staircase, by Ferro e Aço Badaruco, Brasília, Brazil (Source: authors' image)



Fig. 93.7 Comparative study: customized staircase, designed by Diogo Santos, CNC cut and assembled by Ferro e Aço Badaruco, Brasília, Brazil (Source: authors' image)



by the aforementioned company, was US\$2,417.00. This result seems to be consistent with other examples found in earlier works such as that of Bernhard Franken's comparison between a mass standardized pavilion and a customized one in which he reports that the last one was 1/3 less expensive than the first one [16].

93.5 Conclusion

We believe we have demonstrated that digital fabrication is starting to produce impact in the Brazilian architecture. This process is not limited to exceptional buildings, but is also finding way in small experiences involving ordinary needs, from interior design and construction components to simple buildings.

The last example, the comparison between staircases, shows that digital fabrication technology is also having an impact towards mass customization in the Brazilian architecture. Cultural obstacles, such as the ingrained belief that a singular artifact is necessarily more expensive than a mass standardized one, are being progressively removed.

References

1. Mitchell W (1995) Representation. In: Lentricchia F, McLaughlin T (eds) Critical terms for literary study, 2nd edn. University of Chicago Press, Chicago, p 11
2. Merleau-Ponty Maurice (1945) Cezanne's doubt. Essay, Paris
3. Zevi Bruno (1957) Architecture as space—how to look at architecture. Horizon Press, New York, pp 22–23
4. Lyn F, Dulaney R (2009) A Case for Drawing. ARCC J 6(1):23–30
5. Robbins E (1997) Why architects draw. MIT Press, Cambridge, pp 10–11
6. Lyn F, Dulaney R (2009) A case for drawing. ARCC J 6(1):24
7. Coulton JJ (1977) Ancient Greek architects at work—problems of structure and design. Cornell University Press, New York
8. Kostof S (1977) The architect—chapters in the history of the profession. University of California Press, Berkeley
9. Kolarevic B (2003) Architecture in the digital age—design and manufacturing. Taylor & Francis, New York, p 9
10. Kolarevic B (2003) Digital production. In: Kolarevic B (ed) Architecture in the digital age—design and manufacturing. Taylor & Francis, New York, pp 31–51
11. Polette Doug, Landers Jack M (1995) Construction systems. Goodheart-Willcox Company, Inc., Publishers, pp 45–49
12. Silva N, Bridges A, Lima E (2009) Computer numerical control, mass customization and architectural education in an emergent economy. In: EIAE 09, international conference on engineering education, instructional technology, assessment, and e-learning, Bridgeport
13. Kieran S, Timberlake J (2004) Refabricating architecture. McGraw-Hill, New York, pp 131–153
14. Kolarevic B (2005) Digital production. In: Branko kolarevic, architecture in the digital age—design and manufacturing, Taylor & Francis, New York, pp 52–53

15. Schodek D, Bechthold M, Griggs K, Kao KM, Steinberg M (2005) Digital design and manufacturing—CAD/CAM applications in architecture and design. John Wiley & Sons, New Jersey, pp 339–344
16. Franken Bernhard (2005) Real as data. In: In branko kolarevic, architecture in the digital age—design and manufacturing. Taylor & Francis, New York, p 138

Chapter 94

Three Dimensional SPMD Matrix–Matrix Multiplication Algorithm and a Stacked Many-Core Processor Architecture

Ahmed S. Zekri

Abstract Current applications in image and media processing, scientific and engineering computing require a tremendous processing and higher memory bandwidth to gain high performance. Three dimensional multi/manycore processors stacked with memory layer(s) may provide good processing facilities to enhance the performance of these applications. In this paper, we introduce a proposal of a 3-D stacked many-core processor architecture composing of a number of processing elements (PEs) layers stacked with one or more memory layer shared among all PEs. Unlike many 3-D machine architectures, the proposed model uses local communications between PEs in both horizontal and vertical links avoiding the cost of building specialized interconnection networks. We present a novel memory efficient SPMD blocked algorithm for performing the kernel matrix–matrix multiply operation (MMM), on the 3D processor architecture. Our analytical evaluation of the 3-D stacked architecture showed a near linear speedup as the number of PE layers increases while data communication and redistribution is overlapped with computing.

94.1 Introduction

The emerging Through Silicon Via (TSV) technology is a promising design approach to hide the processor-memory speed gap by allowing processor and memory layers to be stacked together forming 3-D architectures delivering high

A. S. Zekri (✉)

Department of Mathematics and Computer Science, Faculty of Science,
Alexandria University, El-Shatbi, Alexandria 21526, Egypt
e-mail: ahmed.zekri@alex-sci.edu.eg

performance for many scientific, media and engineering applications [1, 2]. The resulting 3-D architectures require efficient 3-D algorithms to exploit the tremendous processing power and the huge memory bandwidth of these architectures.

2-D algorithms for performing matrix–matrix operations have been devised since the roots of parallel computer systems. For 2-D mesh architectures, the algorithms can be differentiated according to the amount of data input and output at each step of computing into two groups. The first group is due to using two vectors and one matrix at each iteration of the algorithm [3–5]. These algorithms highlight the Level-2 BLAS to compute the MMM operation. The algorithms in the second group relies on using three matrices (or sub-matrices) at each step of computing as well as avoiding data broadcasting. The well-known Cannon algorithm [6] and Systolic algorithm [7] belong to this group.

3-D algorithms have been studied and implemented on many shared and distributed memory machines [8–10]. A primary gain of 3-D algorithms is they reduce communication to a factor of $P^{1/6}$ over 2-D counterparts [8, 11], where P is the total number of processors.

In this paper, we present a 3-D processor architecture that is stacked with memory to provide large bandwidth to the processing elements (PEs). The problem data are initially stored in the memory layer shared by all PE layers. Data movement is overlapped with computing in the PEs since we used the PE layer next to the memory chips for data load/store and redistribution to meet the data alignment required in PEs for correct computing. To evaluate the performance of the 3-D processor, we devised a 3-D blocked matrix–matrix multiplication algorithm that utilize the huge bandwidth of the stacked memory and minimize communication time by copying parts of input matrix before computing and streaming other parts during processing since the local memory inside PEs is not enough to hold all problem data. Our 3-D algorithm achieved a high performance of the 3-D processor by overlapping computations and communications and using more PE layers. Unlike other 3-D architectures, our proposed architecture does not have any dedicated special networks for global communications and data alignment operations.

This paper is organized as follows. In Sect. 94.2, we present the 3-D processor architecture. In Sect. 94.3, we outline the 3-D algorithm. In Sect. 94.4, we analyze the 3-D algorithm and show its performance on the proposed architecture. Section 94.5 concludes the paper.

94.2 3-D Many-Core Processor Architecture

A primary goal of building 3D processor chips is to overcome communication latency across 2D chip boundaries. Our proposed 3-D architecture composed of L processor layers stacked with one or more memory layers (see Figs. 94.1, 94.2). Each processing element (PE) layer is an $N \times N$ PEs connected by a 2-D torus communication network. Each processing element $PE^\ell(i, j)$, $0 \leq i, j \leq N - 1$, is identified by its 2-D Cartesian coordinate (i, j) inside the 2-D layer, and the layer

Fig. 94.1 Schematic diagram of the 3-D processor consisting of L PEs layers stacked with DRAM memory layer(s)

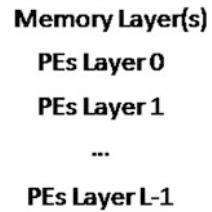
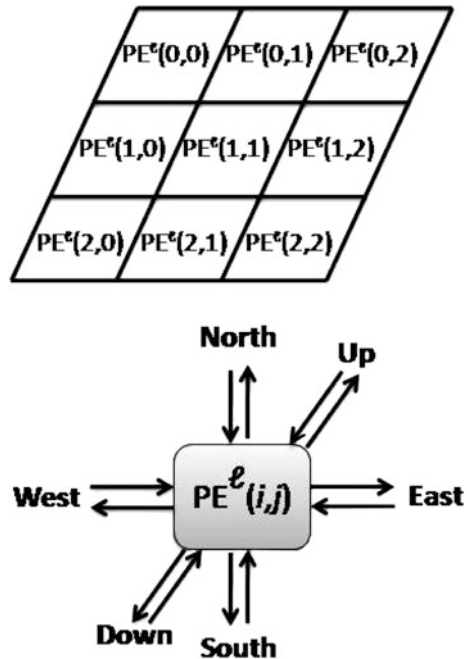


Fig. 94.2 **a** One PEs layer of the 3-D processor showing the Cartesian coordinates (i,j) of PEs; **b** a processing element $PE^\ell(i,j)$ showing the communication links with horizontal (East, West, North, South) and vertical (Up, Down) PEs, ℓ is the current layer number



number, $0 \leq \ell \leq L - 1$. For example, $PE^2(1,0)$ is the PE at location $(1,0)$ inside layer 2. Every PE has a full-fledged RISC processor with a data and instruction caches beside a DRAM local memory for storing data and programs during processing. There are six communication channels for each PE: East, West, North and South for intra-layer communication between neighbor PEs, while Up and Down connections for inter-layer communication to/from vertical PEs in the above and below adjacent PE layers.

The stacked memory layer is an $N \times N$ separate SDRAM memory chips (MCs) that are vertically connected to the PE layer below it immediately, which we will call Layer 0. Each MC is directly connected to one PE. The other PE layers are numbered from 1 to $L-1$ as we move down from Layer 0. Every access from PE layers, $1, 2, \dots, L - 1$, to the stacked memory layer is done through Layer 0 which loads/stores data and sends it down using message-passing communication to

Layer 1 which sends it to Layer 2, and so on until it reaches the desired layer. Each memory chip $MC(i,j)$, $0 \leq i, j \leq N - 1$, has a unique ID represented as a Cartesian coordinate (i,j) .

94.3 3D Matrix–Matrix Product Algorithm

The 3-D processor architecture can be employed to perform both 3-D and 2-D matrix–matrix algorithms. For 2-D algorithms, the architecture is used as a stack of L 2-D tori of multi-cores where each layer is dedicated to solve a different matrix–matrix operation. For 3-D algorithms, we consider distributing the computing among all layers where they cooperate to solve the whole problem. In fact, since the 3-D architecture is a natural mapping of the index space of the matrix–matrix multiply-add operation (MMM), several 3-D matrix–matrix algorithms for mesh processor arrays can be implemented in our 3-D processor by mapping the computation at the index points of the 3-D index space to corresponding PEs.

In this section, we used the large memory bandwidth coming from the stacked memory to devise a novel 3-D algorithm to compute the MMM operation $C \leftarrow C + A \times B$ where \leftarrow denotes assignment, $A = [a(i,k)]$, $B = [b(k,j)]$, and $C = [c(i,j)]$, where $0 \leq i \leq m_1 - 1$, $0 \leq j \leq m_2 - 1$, and $0 \leq k \leq m_3 - 1$. The main idea behind the algorithm is maximizing data reuse and overlapping computations with data communications.

94.3.1 The Algorithm

We assume the three matrices A , B , and C with dimensions $m_1 \times m_3$, $m_3 \times m_2$, and $m_1 \times m_2$, respectively, are divided into blocks of size $\beta \times \beta$ where for simplicity the dimensions of the matrices are multiples of β . Now A , B , and C are regarded as $N_1 \times N_3$, $N_3 \times N_2$, and $N_1 \times N_2$, respectively.

In order to distribute the three matrices into the $N \times N$ memory chips of the stacked memory layer, we further divide the matrices A , B , and C into $N \times N$ larger blocks each of size $n_1 \times n_3$, $n_3 \times n_2$, and $n_1 \times n_2$, respectively, where $n_1 = N_1/N$, $n_2 = N_2/N$, $n_3 = N_3/N$ and assuming N_1, N_2 , and N_3 are divisible by N . So, each memory chip will have one larger block of each of A , B , and C , which we will call a super-block, see Fig. 94.3. We denote a super-block X^{ij} , $0 \leq i, j \leq N - 1$, of matrix X where the superscript indicates the memory chip coordinate which stores the super-block. That is, the super-block X^{ij} will reside in memory chip $MC(i,j)$.

To compute the blocks of matrix C simultaneously using all PE layers, each super-block C^{ij} is divided into $L - 1$ horizontal block-panels where each block-panel is computed in processor $PE^l(i,j)$ of each layer. This block-panel has

Fig. 94.3 The distribution of matrices A , B , and C among the memory chips $MC(i,j)$, $0 \leq i, j \leq N-1$, $N=2$ and $L=2$

| | | | |
|----------------|---------------|----------------|---------------|
| $A^{00}(0,0)$ | $A^{00}(0,1)$ | $A^{01}(0,2)$ | $A^{01}(0,3)$ |
| $B^{00}(0,0)$ | $B^{00}(0,1)$ | $B^{01}(0,2)$ | $B^{01}(0,3)$ |
| $C^{00}(0,0)$ | $C^{00}(0,1)$ | $C^{01}(0,2)$ | $C^{01}(0,3)$ |
| MC[0,0] | | MC[0,1] | |
| $A^{00}(1,0)$ | $A^{00}(1,1)$ | $A^{01}(1,2)$ | $A^{01}(1,3)$ |
| $B^{00}(1,0)$ | $B^{00}(1,1)$ | $B^{01}(1,2)$ | $B^{01}(1,3)$ |
| $C^{00}(1,0)$ | $C^{00}(1,1)$ | $C^{01}(1,2)$ | $C^{01}(1,3)$ |
| MC[1,0] | | MC[1,1] | |
| $A^{10}(2,0)$ | $A^{10}(2,1)$ | $A^{00}(2,2)$ | $A^{00}(2,3)$ |
| $B^{10}(2,0)$ | $B^{10}(2,1)$ | $B^{00}(2,2)$ | $B^{00}(2,3)$ |
| $C^{10}(2,0)$ | $C^{10}(2,1)$ | $C^{00}(2,2)$ | $C^{00}(2,3)$ |
| MC[1,0] | | MC[1,1] | |
| $A^{10}(3,0)$ | $A^{10}(3,1)$ | $A^{11}(3,2)$ | $A^{11}(3,3)$ |
| $B^{10}(3,0)$ | $B^{10}(3,1)$ | $B^{11}(3,2)$ | $B^{11}(3,3)$ |
| $C^{10}(3,0)$ | $C^{10}(3,1)$ | $C^{11}(3,2)$ | $C^{11}(3,3)$ |

$n_1/(L-1) \times n_2$ small blocks of size $\beta \times \beta$. To overlap computing in PE layers with loading/storing operations, we used the PEs of Layer 0 for concurrent loading/storing of data from the memory layer to the other PE layers. Also, we assume one super-block of each of A , B , and C reside in the corresponding memory chip.

Our algorithm is described using MATLAB colon notation in the pseudo-code program in Listing 1 below ($B^{uv}(:,0)$ means all $\beta \times \beta$ blocks of super-block B^{uv} located at column 0). This program will be executed by all PEs concurrently in SPMD computing mode. For making the description more clear, we separated the details of operations Multiply(block,block,block), Align(block,direction), Load(block,layer#), and Store(block) into Listing 2, 3, 4, and 5, respectively. Note that, the pseudo-codes in these listings are for execution for one PE, and all PEs of all layers execute concurrently.

- Initial stage:** This is a communication step where the $N \times N$ super-blocks B^{ij} , $0 \leq i, j \leq N-1$ are loaded simultaneously by all PEs of Layer 0 so that $PE^0(i,j)$ loads from its corresponding memory chip $MC(i,j)$ vertically located above it. These super-blocks are simultaneously copied to other corresponding PEs in layers $1 \dots L-1$. For example, the super-block B^{12} is copied to $PE^\ell(1,2)$ in layers $1 \leq \ell \leq L-1$.
- Main stage:** This is a computation step of the algorithm which goes as follows: for each layer $1 \leq \ell \leq L-1$, each PE^ℓ loads one block row of the corresponding super-block of A and use it to compute one block-row of the corresponding super-block of matrix C . This is done concurrently in all PE layers. To compute one $\beta \times \beta$ block of C inside each PE, the PEs in the same layer perform local block matrix–matrix multiplication and inter-communicate together using intra-layer horizontal local connections so that the correct blocks of $A^{ij}(I,K)$ and $B^{jk}(K,J)$ meet in correct PEs of the same layer to compute blocks $C^{ij}(I,J)$. We used Cannon's 2-D algorithm in each layer to compute the blocks $C^{ij}(I,J)$ where the Align() operation is applied to align the $\beta \times \beta$ blocks of super-blocks of

matrices A and B . After computing one block row of super-block C^{ij} inside processors $PE^\ell(i,j)$, a Store() operation is applied to store results back to the corresponding memory chip $MC(i,j)$ and another block row of C^{ij} is computed till all super-blocks C^{ij} of matrix C are computed in all PE layers.

Listing 1:

Description of the algorithm executed in $PE^\ell(u,v)$,
where $0 \leq u, v \leq N-1$.

```

01. %INITIAL STAGE:
02. If  $l = 0$  Then
03.   Load( $B^{uv}(:, 0), \ell$ )
04.   For  $q = 1$  To  $n_2-1$ 
05.     Align( $B^{uv}(:, q-1), \text{north}$ )
06.     Load( $B^{uv}(:, q), \ell$ )
07.     Send( $B^{uv}(:, q-1), \text{down}$ )
08.   EndFor
09.   Align( $B^{uv}(:, n_2-1), \text{north}$ )
10.   Send( $B^{uv}(:, n_2-1), \text{down}$ )
11. El self  $l = L - 1$  Then
12.   For  $q = 0$  To  $n_2-1$ 
13.     Receive( $B^{uv}(:, q), \text{up}$ )
14.   end for
15. Else
16.   For  $q = 0$  To  $n_2-1$ 
17.     Receive( $B^{uv}(:, q), \text{up}$ )
18.     Send( $B^{uv}(:, q), \text{down}$ )
19.   EndFor
20. EndIf
21. % MAIN STAGE:
22. If  $l \neq 0$  Then
23.   For  $I = (l-1)n_1/(L-1)$  To  $n_1l/(L-1)-1$ 
24.     Load( $A^{uv}(I, :), \ell$ )
25.     Align( $A^{uv}(I, :), \text{west}$ )
26.     For  $J = 0$  To  $n_2-1$ 
27.       For  $K = 0$  To  $n_3-1$ 
28.         Multiply( $C^{ij}(I, J),$ 
29.            $A^{ij}(I, K), B^{ij}(K, J))$ 
30.       EndFor
31.     EndFor
32.   Store( $C^{uv}(I, :)$ )
33. EndFor
34. EndIf

```

Listing 2:

Multiplying blocks $A^{uv}(_, _)$ and $B^{uv}(_, _)$ inside $PE^\ell(u, v)$.

```

01. Multiply( $C^{uv}(\_, \_)$ ,  $A^{uv}(\_, \_)$ ,  $B^{uv}(\_, \_)$ )
02. {
03. For  $step = 0$  To  $N - 1$ 
04.    $C^{uv}(\_, \_) += A^{uv}(\_, \_) \times B^{uv}(\_, \_)$ 
05.   Send( $A^{uv}(\_, \_)$ , west) to  $PE^\ell(u, [v-1] \bmod N)$ 
06.   Send( $A^{uv}(\_, \_)$ , north) to  $PE^\ell([u-1] \bmod N, v)$ 
07.   Receive( $A^{uv}(\_, \_)$ , east) from  $PE^\ell(u, [v+1] \bmod N)$ 
08.   Receive( $A^{uv}(\_, \_)$ , south) from  $PE^\ell([u+1] \bmod N, v)$ 
09. EndFor
10. }
```

Listing 3:

Alignment of block $X^{uv}(_, _)$ in current PE,
but relative to PEs of current layer ℓ .

```

01. Align( $X^{uv}(\_, \_)$ , direction)
02. {
03. If direction = west Then
04.   Send( $X^{uv}(\_, \_)$ , west) to  $PE^\ell(u, [v-u] \bmod N)$ 
05.   Receive( $X^{uv}(\_, \_)$ , east) from  $PE^\ell(u, [v+u] \bmod N)$ 
06. ElseIf direction = north Then
07.   Send( $X^{uv}(\_, \_)$ , north) to  $PE^\ell([u-v] \bmod N, v)$ 
08.   Receive( $X^{uv}(\_, \_)$ , south) from  $PE^\ell([u+v] \bmod N, v)$ 
09. EndIf
10. }
```

Listing 4:

Loading block $X^{uv}(_, _)$ of super-block X^{uv}
from $MC(u, v)$ into current $PE^\ell(u, v)$.

```

01. Load( $X^{uv}(\_, \_)$ ,  $\ell$ )
02. {
03. % send signal to  $PE^0(u, v)$  at Layer 0 to do:
04. Load( $X^{uv}(\_, \_)$ , 0)
05. Send( $X^{uv}(\_, \_)$ , down)
06. % send signals to  $PE^w(u, v)$ ,  $1 \leq w \leq \ell-1$  to do:
07. Receive( $X^{uv}(\_, \_)$ , up)
08. Send( $X^{uv}(\_, \_)$ , down)
09. % in current  $PE^\ell(u, v)$  do:
10. Receive( $X^{uv}(\_, \_)$ , up)
11. }
```


Listing 5:

Storing block $X^{uv}(_, _)$ into $MC(u, v)$.

```

01. Store( $X^{uv}(\_, \_)$ )
02. {
03. % in current PE do:
04. Send( $X^{uv}(\_, \_)$ , up)
05. % send signal to  $PE^w(u, v)$ ,  $1 \leq w \leq l - 1$  to do:
06. Receive( $X^{uv}(\_, \_)$ , down)
07. Send( $X^{uv}(\_, \_)$ , up)
08. % send signal to  $PE^0(u, v)$  at Layer 0 to do:
09. Store( $X^{uv}(\_, \_)$ )
10. }

```

94.4 Performance Evaluation

In this section, we analyze and evaluate the performance of the proposed 3-D processor in executing our blocked MMM algorithm. Before starting the execution, the three matrices are input and distributed among the memory chips as we showed in the previous section. Also, we assume all PEs in the 3-D processor are loaded with one program (Listing 1), and they execute asynchronously in SPMD mode. The computing inside all layers will be independent from one another since different blocks of matrix C are computed in different PE layers. However, all PE layers will share in communication especially passing data/results between PEs layers and memory layer. In addition, PEs of the same layer collaborate and use intra-layer local communications only during computing the corresponding blocks of C .

94.4.1 Model Parameters

1. Memory operations: We assume the super-blocks of matrices A , B , and C are layed out in the memory chips in block data layout format [12] so that any $\beta \times \beta$ block is stored continuously in memory. We choose β to be the optimal size for the blocks, i.e., the length of the PE cache line in words. For processors at Layer 0, $PE^0(i, j)$, the latency of SDRAM chip for read/write one word is t_{start}^m , and using enough memory banks inside each memory chip, a pipelined load/store of data can be achieved. So, the time of moving one $\beta \times \beta$ block of matrix elements from a memory chip to one PE in Layer 0 is modeled as:

$$T_{load}^0 = t_{start}^m + t_{word}^m \times \beta^2,$$

where, t_{word}^m is the memory per word transfer time that is equal to one PE clock cycle, after some startup latency, in the pipelined model.

2. Message-passing operations: Sending/receiving one $\beta \times \beta$ block between adjacent vertical PE neighbors is modeled as:

$$T_{send} = t_{start}^v + t_{word}^v \times \beta^2,$$

where, t_{start}^v is the startup time for sending a message between vertically adjacent PEs, and t_{word}^v is the vertical per word transfer time. Similarly, sending/receiving between horizontally adjacent PEs in the same layer can be modeled as the last equation, where t_{start}^h and t_{word}^h are the horizontal startup and per-word times, respectively. As Listing 4 shows the implementation of loading a block in PE ^{ℓ} at Layer ℓ , the time to load one $\beta \times \beta$ block into a PE in Layer $1 \leq \ell \leq L - 1$ is estimated as:

$$T_{load}^\ell = T_{load}^0 + \ell \times T_{send},$$

because a PE can send and receive simultaneously either vertically or horizontally. Also, the message passing Send() operation is non-blocking to allow the processor to proceed execution whenever appropriate and do not wait till the receiver responds. But, the Receive() operation is a blocking one. The alignment required in each PEs layer is implemented in concurrent message-passing communication as in Listing 3. Therefore, the concurrent alignment time of block rows of super-blocks A^{ij} inside the PEs can be estimated using pipelined cut through horizontal routing as:

$$T_{align} = t_{start}^h + (N - 1) \times t_h + t_{word}^h \times n_3 \times \beta^2,$$

where, t_h is the per-hop time to transfer the header of the message between neighbor layer PEs.

3. Parallel overhead: The computations inside each PE layer of the 3-D processor are independent from the computations in other PE layers because different blocks of matrix C are computed in each layer. Therefore, the computing in all layers can proceed in parallel. However, there are inter-layer communication to pass the blocks loaded from the memory layer to different PE layers or store results back to memory layer. Since there is only one I/S layer (i.e., Layer 0), all data loading/storing operations of PE layers $1 \dots L - 1$ will be executed one after the other. That is the load/store operations will be chained among PE layers.

Computing one $\beta \times \beta$ block, $C^{ij}(_, _)$, inside each PE of the same layer requires the blocks of A and B be properly aligned among PEs to obtain correct results. Since all PEs execute concurrently in an asynchronous mode, a barrier synchronization point is required immediately before commencing the local computations inside PEs, i.e., immediately after line 26 of Listing 1. This guarantees that all current blocks in PEs of current layer are properly placed in the correct PEs in order to get correct results of the blocks of C . The worst case scenario for this synchronization overhead is that the N^2 PEs of current layer execute the barrier

Table 94.1 This table shows the values selected for our analytical model parameters

| Parameter | Value |
|-----------------------------|---------------|
| PE frequency | 2.0 GHz |
| SDRAM latency, t_s^m | 10 ns |
| Message startup, t_s^h | 10 PE cycles |
| Message startup, t_s^v | 5 PE cycles |
| Per hop time, t_h | 15 PE cycles |
| Cache line, β | 64 Bytes |
| Synchronization cycles, S | 200 PE cycles |

operation one after the other. Therefore, a maximum expected parallel overhead of computing in one PE layer ℓ is estimated as:

$$T_o = n_1 / (L - 1) \times n_2 \times (S + N^2 - 1),$$

where, S is the number of CPU clock cycles to execute the instructions of a barrier operation in one PE. Since the division of the MMM problem among all PEs resulted in equal workloads in all PE layers and between PEs of the same layer, the above synchronization overhead can be minimized by activating all PEs to start their program execution by an initial synchronization operation.

94.4.2 Evaluation

The total parallel execution time of our algorithm is the sum of computing time, communication time, and parallel overhead time. For the local computation inside each PE, we used the conventional serial MMM algorithm for multiplying two $\beta \times \beta$ blocks in $2\beta^3$ PE clock cycles. We choose our analytical model parameters as shown in Table 94.1.

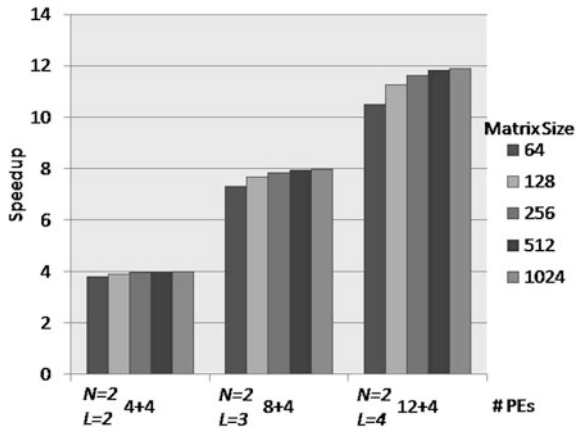
The execution of our algorithm (Listing 1) goes as follows: in the initial stage, we see that execution starts in Layer 0, then after the operations Align(block,direction) and Load(block, ℓ) of the column block $B^{uv}(:, _)$ it starts in Layer 1, and so on till Layer $L - 1$. After Layer 1 finishes the initial stage, it immediately starts to execute the main stage from Line 21. It is clear in the main stage (Lines 21–33) that our algorithm is load balanced since all PEs have the same workload to process. However, the execution time of each layer increases as we move down to Layer $L-1$ because the time of both operations Load(block, ℓ) and Store(block) increases as ℓ increases. This guarantees that the operations in Lines 24 and 31 of Listing 1 of all PE layers will not execute at the same time.

From the above discussion, we see the execution of the algorithm will be chained among all PE layers. We estimated the total parallel execution time of our algorithm as:

$$T_p = T_{initial} + T_{comp}^{L-1} + n_1 / (L - 1) \times (L - 2) \times T_{load}^0,$$

where, the times $T_{initial}$ and T_{comp}^{L-1} are given by:

Fig. 94.4 Measured speedup with increasing the number of PEs ($= N^2 \times L$) in the 3-D processor, L : number of PEs layers, each layer L has $N \times N$ PEs



$$\begin{aligned}
 T_{\text{initial}} &= (n_2 - 1) \times (\max(T_{\text{align}}, T_{\text{load}}^0) + T_{\text{send}}^v) \\
 &\quad + (L - 1) \times T_{\text{send}}^v + T_{\text{align}}, \\
 T_{\text{comp}}^{L-1} &= n_1 / (L - 1) \times (T_{\text{load}}^{\ell} + T_{\text{align}} + n_2 \times S \\
 &\quad + n_2 \times n_3 \times T_{\text{multiply}} + T_{\text{store}}^0), \\
 T_{\text{multiply}} &= N \times \max(2 \times \beta^3, t_{\text{send}}^v).
 \end{aligned}$$

Figure 94.4 shows the performance of our 3-D processor on performing the MMM operation. The matrices A , B , and C are chosen with square sizes. The speedup of the 3-D algorithm over the conventional standard matrix–matrix multiplication is measured. As we can see, increasing the number of PE layers leads to increasing speedup. For example, when $L = 4$ the speedup is approaching 12.0 which is the actual number of PEs used in computing because the rest of PEs ($=4$) are exploited in moving data and results between the rest of PE layers and the memory layer. It may be expected that as we increase the number of PE layers, memory operations become the bottleneck of the processor since all PE layers share one memory layer. However, this is not the case because our architecture overlap data movement and data redistribution with computing. Figure 94.4 also shows that increasing the matrices size from 64 going to 1024 raises the speedup since more data reuse in the PEs caches compensate the startup overhead of copying super-blocks of matrix B to all PE layer.

94.5 Conclusions

TSV's 3-D stacking using a mix of logic process chips and DRAM process chips permitted great processing capabilities and huge memory bandwidths to deliver high performance in many scientific and engineering applications. In this paper,

we presented a 3-D stacked many-core processor composed of PEs layers stacked with memory layer(s). The high memory bandwidth and the increased inter-layer and intra-layer PE connections permitted us to devise a novel highly efficient 3-D parallel algorithm for performing the kernel matrix–matrix multiplication (MMM). Our analytical evaluation of the 3-D processor on the MMM kernel showed a near linear speedup as we increase the number of PE layers. The overhead of data movement and redistribution is also overlapped with computing specially as the size of matrices increase.

One problem that we didn't take into our consideration in this paper is the amount of heat dissipated as we increase the number of PE layers. A possible solution that will be studied in our future works is studying variant folded organizations of the torus interconnection in each PEs layer as well as using power-aware routing algorithms such as the one presented in [13].

References

1. IBM Research (2007) 3-D Chips: IBM moves Moore's law into the third dimension. ScienceDaily 12 April 2007. <http://www.sciencedaily.com/releases/2007/04/070412132140.htm>
2. Xie Y (2010) Processor architecture design using 3D integration technology. In: Proceeding of the 23rd International Conference on VLSI Design, pp. 446–451
3. Fox G, Otto S, Hey A (1987) Matrix algorithms on a hypercube I: matrix multiplication. *Parallel Comput* 4:17–31
4. van de Geijn R, Watts J (1995) SUMMA: scalable universal matrix multiplication algorithm. The University of Texas, Technical Report TR-95-13, April 1995
5. Agarwal R, Gustavson F, Zubair M (1994) A high performance matrix multiplication algorithm on a distributed-memory parallel computer, using overlapped communication. *IBM J Res Dev* 38(6):673–681
6. Cannon L (1969) A cellular computer to implement the kalman filter algorithm, Ph.D. dissertation, Montana State University, 1969
7. Kung S (1988) VLSI array processors. Prentice Hall, Englewood Cliffs
8. Agarwal R et al (1995) A three-dimensional approach to parallel matrix multiplication. *IBM J Res Dev* 39(5):575–582
9. Ho C-T, Johnsson SL, Edelman A (1991) Matrix multiplication on hypercubes using full bandwidth and constant storage. In: The 1991 International Conference on Parallel Processing, pp. 447–451
10. Kumar V, Gupta A (1994) Analyzing scalability of parallel algorithms and architectures. *J Parallel Distrib Comput* 22(3):379–391
11. Grama A, Gupta A, Karypis G, Kumar V (2003) Introduction to parallel computing, 2nd edn. Addison Wesley, Reading
12. Park N, Hong B, Prasanna VK (2002) Analysis of memory hierarchy performance of block data layout. In: ICPP '02: Proceedings of the 2002 International Conference on Parallel Processing (ICPP'02), p. 35
13. Kdouh W, El-Rewini H (2011) Reliability-aware platform optimization for 3d chip multi-processors. *J Supercomput* 51:1–20. <http://dx.doi.org/10.1007/s11227-011-0577-5>

Chapter 95

Face: Fractal Analysis in Cell Engineering

K. P. Lam, D. J. Collins and J. B. Richardson

Abstract Most fractal compression schemes encode an image as a collection of transforms that analyse image detail at every scale, typically allowing a receiver to regenerate or synthesize the output using “instructions” from the transmitter. Conceived in the mid 1980s, such an analyse/synthesis approach to image transformation exploits the self-affine approximations as the basis functions for capturing image detail which is subsequently used in the reconstruction algorithm. The latter often entails the application of an iterative procedure which specifies a set of (growth) rules, not unlike the so called *L*-system (of Lindenmayer) constructed to describe the intricate developmental patterns of biological plants at multiple scales of resolution. Based on the computational science of fractals and image transforms, this paper furthers our earlier development of an investigative visualisation platform which sought to characterise the diversity of patterns observed in the columnar branching of cartilage cells during growth/repair, by providing objective quality measures of the structural organisation and biochemical composition of the repaired tissue. Our analysis was compared with previous studies on histological assessments of cartilages via polarised light microscopy, revealing promising results.

K. P. Lam (✉) · D. J. Collins

School of Computing and Mathematics, University of Keele, Staffordshire,
STAFFS ST5 5BG, U.K

J. B. Richardson

Robert Jones & Agnes Hunt Orthopaedic Hospital (RJA), Oswestry, SY10 7AG, UK

95.1 Introduction

The spontaneous appearance of complex structures during cell growth are of great significance to our understanding of how spatially ordered biological patterns such as those seen in the columnar branching of cartilage cells emerge from small, primary units. Traditionally, the histological characterisation of such bio-structural complexity has been based largely on subjective qualitative descriptions. A synergistic approach would study how pattern formation in living organisms which are composed of many subsystems can be understood through their interactions, whose dynamics are underpinned by physical laws, to generate spatial, temporal and functional structures. Such complex organisational structures often form visible patterns when viewed at a macro level [1]. Indeed, this approach was adopted in our earlier work where computerised image analysis had been applied in conjunction with digital video microscopy for the quantitative characterisation of cell growth and differentiation in such a biological context [2]. This was accomplished by exploring the contemporary branch of computational geometry/mathematics concerning fractals, with the principal aim of providing an objective description of the morphological and biochemical complexity of the systems. Using the well documented technique of fractal image transforms [3], the work examined the role of positional information and, more relevantly, investigated how they might regulate pattern formation at a macroscopic level from a pattern recognition perspective. Computationally, our approach to studying pattern formation is closely allied to the pioneering work of Barnsley, who conjectured that the mathematical theory of fractals could be applied to compress natural images; viz., the principal goal is to find resolution independent models, defined by finite and short strings (of zeros and ones), for real-world images [4]. Specifically, the work entails revisiting methods of the fractal image transformation and, more importantly, its underlying principles and techniques for image coding in the context of pattern descriptions. In particular, our implementation adopted the work of Fisher on Partitioned Iterated Function System (PIFS), which was itself based on Jacquin's first practical implementation of IFS following the seminal work of Barnsley on fractal image coding in the 1980s [5, 6]. Here, histological images that are encoded as PIFS possess details of self-similarity at different scales; however, the "components" defining such self-similarity are far from obvious and less well defined. Thus practical implementations of the fractal encoding procedure must generate an approximation to blocks of an image based on similarity measures that are mathematically well defined and biologically meaningful. From an image analysis standpoint, this is a principal advantage of the fractal encoding framework, particularly for image analysis, as the details and similarity at different scales are present in the often highly compressed fractal code [7].

Image classification is a popular and well researched field, with several well defined approaches for clustering and classification [8]. However, combining fractal encoding and classification procedures is an approach that has not been well documented; only a handful of papers were found; e.g., in [9]. More specifically,

standard or non fractal based image coding algorithms often result in a compressed data file, which can be used to reconstruct the image, or a close approximation of it. The data file itself holds little or no meaning; for example, in transform coding, its only use is to apply a computationally well defined inverse transformation to regenerate the image. By contrast, an image which has been encoded with a fractal compression system results in a different type of data file. Our work thus started by amalgamating the involved concepts, with the goal of developing a classification system that utilises fractal coding. A primary focus of the work reported here concerns a detailed investigation of how visual patterns pertinent to the organisation of cell colonies of cells in repaired cartilage tissue can be extracted and described by automated analysis of standard/routine histology from a spatial viewpoint. To achieve this, the role of positional information at the microscopic level was examined empirically with the goal of understanding how it might regulate pattern formation at the macroscopic level using a pattern recognition (and mining) approach. Based on *Fractal* geometry, the study sought to develop advanced image analysis techniques, coupled with data and rule mining principles, to describe and characterise such complexity of cell/tissue structures in an objective way,¹ hence the title “*FACE*”—*Fractal Analysis in Cell Engineering*.

95.2 Technical Background and Related Work

The mathematics of fractals has been well established since the late 19th century, yet it was only in the mid-1960s that a detailed study of the science of fractals was pioneered by Mandelbrot [10]. Since then, research in field has gained in depth and popularity, including visualisation of the so called *L-System* [11]) and, most relevantly, Barnley’s Iterated Function Systems (IFS) which demonstrated that many natural-looking objects can be obtained as the fixed point of certain types of function [4, 12]. Using IFS, the goal was to develop a (lossy) compression algorithm that takes advantage of the self-similarities in an image, creating an optimal forms of image compression that are comparable to JPEG [13], a long established and standard compliant compression technology which is still in use today. Algorithmically, the technique of IFS is central to fractal encoding, which can best be described using the concept of *fixed point functions* (or attractors); a fixed point function is one which fulfils:

$$f(x_0) = x_0$$

Here, if the function $f(x)$ is chosen to be a contractive affine transform such that:

$$f(x) = \alpha x + \beta \quad \text{where } |\alpha| < 1$$

¹ As digitally sampled microscopic images encode spatial information (objectively) in numerical values.

then $f(x)$ will be convergent to the value x_0 where

$$x_0 = \frac{\beta}{1 - \alpha}$$

Thus, given a value y , it is possible to create a fixed point function to generate this value from any arbitrary initial value. Similarly, it is possible to develop a system to recreate any set of predefined values (the attractors) from an arbitrary set of starting values; viz. an *Iterative Function System* (IFS). Indeed, most fractal rendering software applications utilise the technique of IFS to generate the fractal image.

From a computational perspective, images that are fractally encoded possess details of self-similarity at different scales and, as such, practical implementations of the fractal compression procedure generate an approximation to blocks of an image based upon their similarity. A principal advantage of such a fractal/encoding framework is that the details and similarity at different scales must be present in the compressed fractal code. Consequently, to compress and reconstruct an image using fractal techniques, one must construct an IFS whose attractor is the image itself; i.e., an attractor or fixed point is the set of convergent values of an IFS. In essence, the compressed data is merely the transformations (or “instructions”) required to regenerate this image from any arbitrary starting image; see the original work of Barnsley [12], and subsequently Fisher [5].² Here, our model/algorithm divided the image (I) into n sets of Range blocks (R) and Domain blocks (D) such that:

$$I = \bigcup_{i=1}^n R_i \quad \text{where } R = (D) \quad (95.1)$$

Where $W(x)$ is the map containing the set of fixed point transformations between D and R . Finding this set of transformations is aided by the *Collage Theorem*, as was originally proposed by Barnsley [12] and later implemented as a workable solution by Jacquin [6]. The latter was based upon the *Piecewise Transformation System* (PTS) model and, within this solution the whole of the image is composed of transformations of parts of itself. This is in contrast to Barnsley’s IFS that is typical of the *Self-Transformation System* (STS), where the entire image/function is mapped to each part of the image. In practice, the PTS system is more flexible model than the STS as natural looking images seldom contain any transformation of the entire image. As a result, it follows that the image may be better and more efficiently described by a piecewise or partitioned transformation; viz. a partitioned STS as described in [14]. From an image analysis viewpoint, Eq. (95.1) which specifies an encoding procedure necessitates an exhaustive search or matching process through a set of domain blocks $\{D\}$ to find

² Since Barnsley first presented the idea of fractal image encoding, several approaches have been proposed; [Jac92] and [DUB92].

the best matching pair of D_i and its transformation. More formally, the metric $\text{dist}(\mu \supset R_i, T_i \circ S_i(\mu \supset D_i))$ is a minimum where $\text{dist}(x, y)$ is distance or distortion metric, μ is the partitioned image, S_i consists of a contraction and translation factor and T_i is the block transformation. Based on Fisher's implementation [3], the domain to range block mapping as described in (95.1) is achieved through a *Partitioned Iterative Function System* (PIFS), a 2D (as in spatially 2D) analogue to the IFS. Essentially this is composed of the 3D input vector containing x , y , grey value (g) components and the affine transform in matrix depicted in (95.2) below:

$$w_i \begin{bmatrix} x \\ y \\ g \end{bmatrix} = \begin{bmatrix} a_i & b_i & 0 \\ c_i & d_i & 0 \\ 0 & 0 & s_i \end{bmatrix} \begin{bmatrix} x \\ y \\ g \end{bmatrix} + \begin{bmatrix} e_i \\ f_i \\ o_i \end{bmatrix} \quad (95.2)$$

Equation (95.2) defines the affine transform consisting of two orthonormal components; a spatial transform, designated as v_i in (95.3), and an IFS on the grey value designated as $f(g)$ in (95.4):

$$v_i(x, y) = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e_i \\ f_i \end{bmatrix} \quad (95.3)$$

$$f(g) = s_i - g + o_i \quad (95.4)$$

The spatial transform v_i given in (95.4) was chosen in our work reported previously [2] to achieve a 50 % reduction in size and a translation between the best matching blocks (discussed later). The scale coefficient of grey value (s_i) is a contrast scaling, while the o_i represents an overall brightness translation. The \mathbf{W} map consists of a set of PIFS so that the image is described by the union of the transforms; viz., analogous to the union of range blocks.

$$\mathbf{W}(\mathbf{I}) = \bigcup_{i=1}^n \mathbf{W}_i(x, y, g) \quad (95.5)$$

Given (95.5), it is relatively easy to compute the fixed point of the set; the specified set of transforms are applied repeatedly until $\mathbf{W}(\mathbf{I}) = \mathbf{I}$ at which point, \mathbf{I} will be the attractor. The Collage Theorem facilitates the computation of such transforms by formulating a method (based on the IFS) that approximates a set of values (of \mathbf{I}) through a mapping such that the distance between two sets is minimised. This was achieved in our investigation by dividing the image into blocks (as per the PIFS approach described earlier), calculating the transform required and, within a given tolerance, the associated error by which the best transform was then selected.

Figure 95.1 demonstrates how the concepts and principles discussed above were applied and tested against standard/real-world images. Starting with two seemingly random and arbitrarily organised images/blocks which were produced at coarse level (or large scale σ), each image was successively refined at lower scales by means of the transformations specific to the respective fractal code or

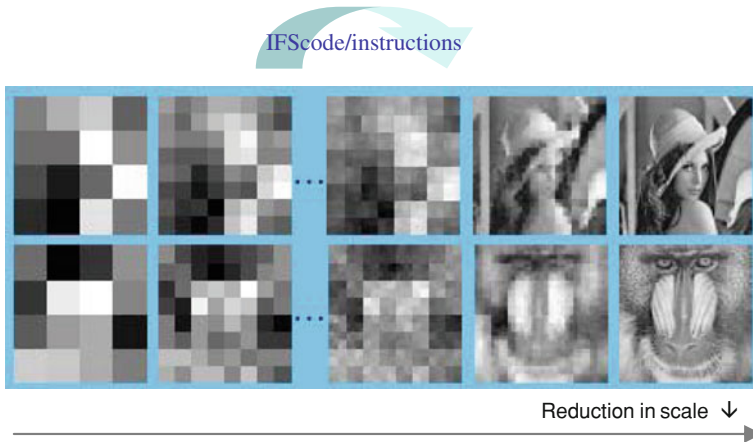


Fig. 95.1 The reconstruction of two well documented images; *Baboon* (bottom) and *Lenna* (top), both of size 512×512 . The final images on the far right of the figure represent the respective *attractors* which converge to the corresponding ‘instruction sets’ defined by the IFS

instructions extracted from the original images (see affine transformation defined in Eqs. (95.2)–(95.5) above). Here, the compressed data is a collection of transforms of known form, the reconstruction of the image requires storing the parameters pertinent to the specific IFS. Furthermore, the fractal code output from an encoding algorithm is akin to a set of “instructions” (rather than a dataset as in other encoding algorithms). Based on these instructions, it should be possible to synthesize the individual images and compare the fractal code generated from them, as the similarities between these images should result in, at least in theory, similar fractal code.

95.3 Fractal Code Classification

In practice, the choice of distortion or distance measure between two images is determined at the discretion of the encoder designer. Typically the L^2 metric is adopted, which is analogous to the Root Mean Square (RMS) error between the range block and the transformed domain block. This led to finding the best matches by minimising the metric R ;

$$R = \sum_{i=1}^n (s - a_i + o - b_i)^2 \quad (95.6)$$

resulting in the following conditions [3]:

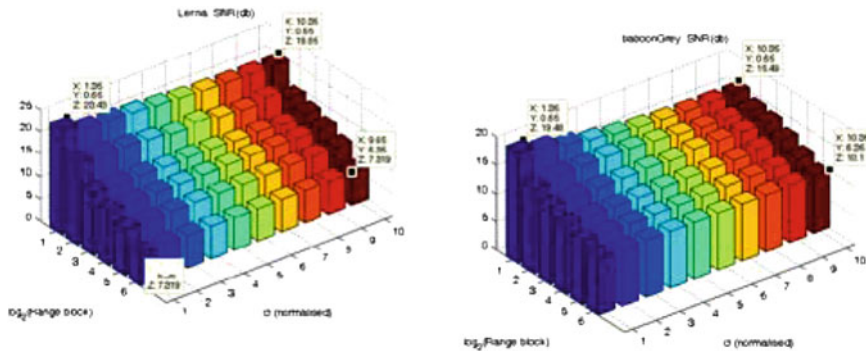


Fig. 95.2 Signal-to-noise ratios (SNRs) computed at different scales for the two images shown in Fig. 95.1—(left) Lenna and (right) baboon. Note also the *log-scale* applied, showing the highly non-linear characteristics of the refinement achievable at different scales (*Range block* sizes)

$$S = \frac{n \sum_{i=1}^n a_i b_i - \sum_{i=1}^n a_i \sum_{i=1}^n b_i}{n \sum_{i=1}^n a_i^2 - (\sum_{i=1}^n a_i)^2} \quad (95.7)$$

$$o = \frac{1}{n} \left[\sum_{i=1}^n b_i - s \sum_{i=1}^n a_i \right] \quad (95.8)$$

and giving:

$$R = \frac{1}{n} \left[\sum_{i=1}^n b_i^2 + s \left(s \sum_{i=1}^n a_i^2 - 2 \sum_{i=1}^n a_i b_i + 2o \sum_{i=1}^n a_i \right) + o \left(no - 2 \sum_{i=1}^n b_i \right) \right] \quad (95.9)$$

Equation (95.9) can be used in the encoder to reduce the number of calculations and, thus, achieve better run-time performance. More relevantly, it also provided the basis for measuring at different scales the overall quality of matches (\sim classification) of specific regions of interest in an image, using the so called “*scale-distortion*” analysis as summarised in Fig. 95.2 below.

To further reduce the amount of searching and number of comparisons required (as per Eq. (95.1)), a pre-processing and classification of domain blocks is first performed before computing the final transforms (Eq. (95.5)) as described in the preceding section. This is achieved by dividing the domain block into quadrants, and using the average value to classify their brightness. Taking into account the three planes of symmetry, a domain block can always be positioned in one of the configurations described in [3]; see Fig. 95.3. These three different classifications of domain block are both unique and all inclusive, thereby reducing the number of comparisons by a factor of three. In addition, second order statistics can also be computed for pixels within each quadrant to further classify potential matching domains more efficiently; i.e., given that there are $24(= 4!)$ possible orderings of



Fig. 95.3 The three possible layouts of the domain blocks [3]

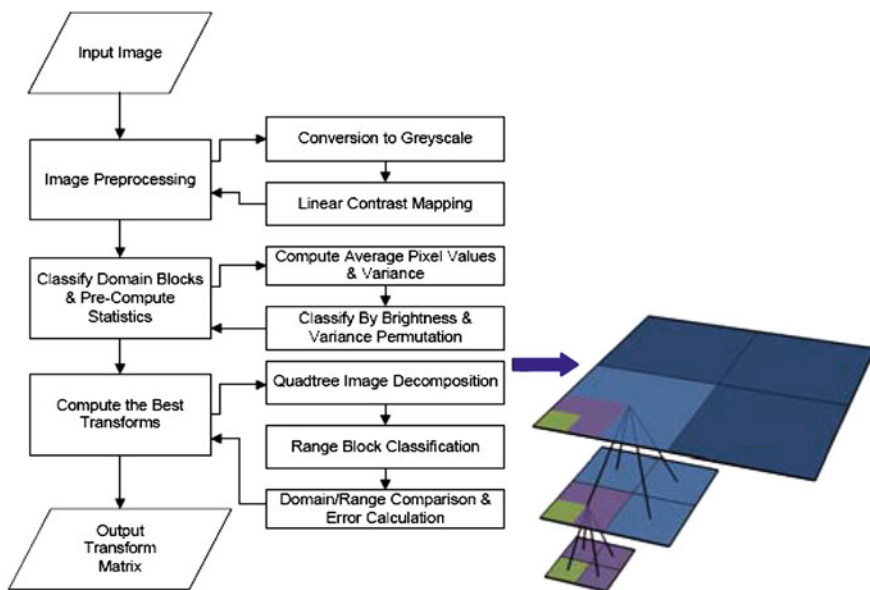


Fig. 95.4 **a** (Left) an overview of the fractal code/instructions extractor algorithm. **b** (Right) quadtree (recursive) image decomposition

the variance in each quadrant, any domain can be determined in constant time as 1 of 3 classes and 1 of 24 subclasses.

Finally, the main transforms computation was developed using a recursive algorithm; namely, the *quadtree* image decomposition as (first) described in [5], to partition the image/blocks as illustrated Fig. 95.4b. The original image was first divided into a square of size 2^n (where n is the smaller of $\lfloor \log_2 \text{width} \rfloor$ and $\lfloor \log_2 \text{height} \rfloor$). A minimum quadtree depth is then assigned (4 by default) and once this depth has been reached, the range block (the portion of the block being analysed) is classified. This classification is the same as that used for the domain blocks; i.e., the variance, quadrant brightness, orientation, etc. are used to assign a class. The range block is then compared to domain blocks of the same class and the error for the associated transform is calculated. This is repeated until all the domain blocks have been compared. In passing, it is noted that the use of recursion as an important feature of our implementation; a highly parallel/distributed implementation of such an algorithm has been developed successfully in our early

work (not reported here) Further, the transforms computed by using Eqs. (95.3)–(95.4), and the associated error using Eq. (95.5), were designed for standard image compression implementation (of Fisher) where the RMS metric as previously earlier (in Eq. (95.9)) was adopted. In general, this could be easily modified or adapted to a more context dependant metric for the specific application. For instance, if it was found that the most important factor for determining the quality of cartilage was the saturation of the individual quadrants, then such consideration could be factored into the error function to facilitate better transforms.

Summarising the descriptions thus far, Fig. 95.4 presents an overview of the fractal code/instructions extractor algorithm.

95.4 Experimental Systems

A modular and extensible visualisation platform was tailor made for researchers at RJAH to help assess the quality of cartilage repair following the ACI procedure [15, 16]. It allows users to dynamically extract the fractal code from an image and flexibly view the largest and lowest error transforms in a selected region. Constructed with an “easy-to-use” graphical interface, the experimental system automatically labels the selected regions of interest based on the underlying fractal code extracted to assist researchers to study, analyse and compare the different parts of an image in detail during routine histology. To date, similar work on objective quantification has used traditional techniques, including polarised light microscopy and *Fourier Transform Infrared Imaging Spectroscopy* (FTIR) [16] to assess quality through vibrational excitation of the collagen and proteoglycan molecules.

To facilitate biologically meaningful characterisation of the fractal code/transforms computed from the input image, our MATLAB implementation stores and returns the transforms in a matrix. This matrix is then sorted by the size of the blocks (inversely proportional to the depth) and then in increasing order, thereby creating a list of the most accurate transforms in order of size. The first twelve of these transforms were displayed and overlaid on top of the original image. This facilitates a preliminary “data mining” (i.e., by visual inspection) and interpretation of this by our collaborators in Oswestry, using their expert knowledge in the field. In particular, the preliminary transform display, as shown Fig. 95.5, was found to suggest that from the cartilage/cell image vertical bands of similarity may be important. This finding was confirmed by clinicians who had previously concluded that humans notice vertical bands in histology samples [15, 17]. Such positive correlation between the fractal transforms and the current qualitative description was encouraging. To this end, three other test images were acquired and the individually cropped regions were passed through the fractal encoder for analysis. This analysis highlighted that regions of “shiny” (or good) cartilage were covered by the smallest transforms (i.e., a region of high detail). This could potentially be used to identify regions of different quality. From an image analysis

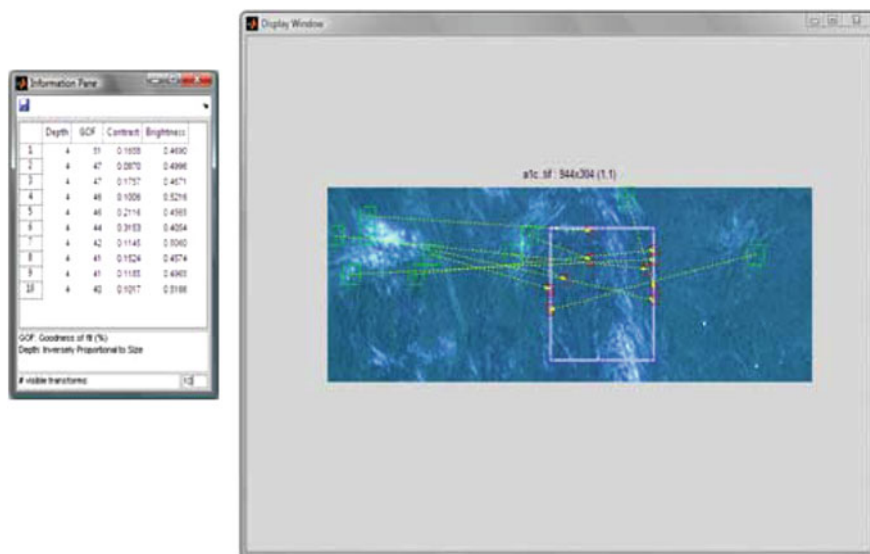


Fig. 95.5 The ROI matched with the multiscale (texture) properties of the fractal code extracted from the cell cartilage image

perspective, a considerable level of understanding of the underlying cell growth mechanism in repair tissue is crucial to enable a *transform viewer* to be constructed with the goal of interpreting the largely abstract mathematical transformations. Figure 95.5 depicts the main user interface of the latest revision of the visualisation system developed, summarized as follows:

- The image view-plane on the right shows the top 12 (default) most similar regions/areas which have been located and ranked by the program. This information is also automatically expanded in a table with details shown on an information window/panel generated separately and placed initially on the left side of the image view plane.
- The first column of the table shows the scale information of the individual matches, the 2nd column the so called “Goodness-of-Fit” (on a scale of 1–10, 10 being the 100 % match based on the criteria adopted), the 3rd column shows the contrast value (\sim feature sharpness) and the 4th column gives the averaged brightness of the regions. These values/statistics are already normalised in order that the identified regions can be measured and compared in a meaningful way—i.e., like-for-like comparisons.
- A mouse is used to move and/or change the size of the so called ROI (region of interest) window (highlighted as a rectangular white box within the image view plane) to study other areas of interest within the image viewplane. Doing so automatically and synchronously updates the relevant statistics displayed on the information panel. Similarly, the number of rows included in the information

panel can also be changed at any time using the input textbox at the bottom right.

- The information panel can also save the statistics displayed at any time, for example, to enable further investigation.

In passing, the demonstrator constructed was a fully functioning, royalty-free experimental platform which was designed to allow non-expert users to dynamically extract the fractal code from a digitised image and flexibly view the largest and lowest error transforms in a selected region. In addition to its ability to effectively compare transforms between different regions of the same image, the production code of the system (currently applicable to *Windows* only) was also equipped with the facility to combine multiple images whereby local similarities between the fractal code extracted can be discovered and (hopefully) utilised to classify the images of interest.

95.5 Experiment/Results and Discussion

The RJA Orthopaedic Hospital in Oswestry have an archive of approximately 200 cartilage biopsies taken from patients who have received cell transplants in the knee joint since the late 1990s [15]. From these biopsies, digitised images were collected using a high power digital camera (Hamamatsu) and microscope (Leica). The images were of sections of the repair tissue that had been stained with standard histological methods; i.e., H&E and toluidine blue viewed both with transmitted light and polarised light. For these patients, both the Lysholm score (of functional outcome) in the short and medium term (1–9 years) are known. As proof of concept, a selection of these digitised images of histology was used to test the efficacy of the visualisation platform built in terms of its ability to quantify revealed visual patterns. A key objective was to investigate if the computed fractal transforms could discern or recognise differences in the appearance of cartilage repair tissue when digitised images are collected using transmitted polarised light. Here, articular cartilage when seen through polarised light has been termed hyaline-like or fibrocartilaginous, but this is a highly subjective judgement largely based on the uniformity of appearance; see Fig. 95.6a–c for illustration. Using the same approach as previously reported in [15], our experiment examined whether these differences can be assessed using established pattern recognition and mining techniques [18] and, if so, how well this distinction compares with previous assessments done by the human eye. In particular, a principal goal was to identify and (subsequently) study characteristics of the fractal code extracted for the apparently uniform and vertical orientation of collagen in the deep zone that, when view under the polarised microscope, is integrated with the underlying calcified layer of cartilage (thus suggesting successful replacement of tissue in this region). Similarly, the orientation of the collagen fibres which lie parallel to the hyaline-like articulating cartilage in the surface zone resembles that found in normal joints [15].

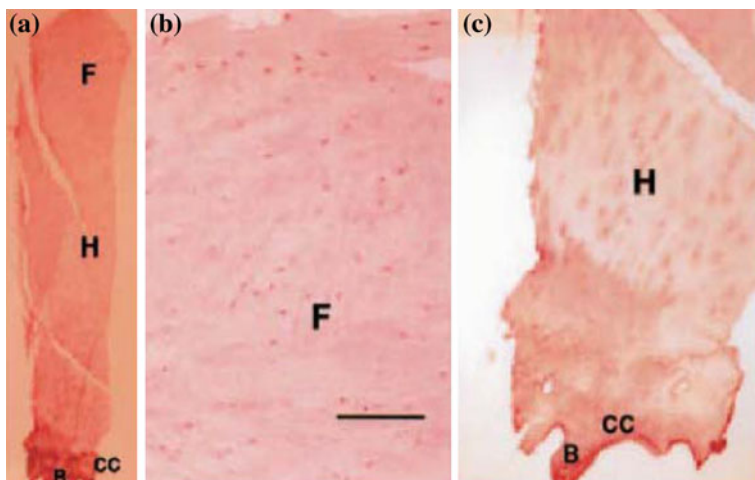


Fig. 95.6 **a–c** From left to right. Photomicrograph (a) showing a full-depth core biopsy of repair tissue taken 12 months after implantation. **b, c**—At a higher magnification showing **b** upper and **c** mid zones (bars = 100 μ). Keys: F, fibrocartilage-like tissue; H, hyaline-like tissue; CC, calcified cartilage; B, subchondral bone. Reproduced with author's permission from [15]

Analytically, a close examination of the computed fractal/IFS code reveals that, for our encoder to be effective, the relevant images or regions of interest must be composed of features at scales that are also present at coarser scales up to a rigid translation and an affine transform of intensities. This is illustrated in Fig. 95.7, which depicts the typically hyaline rich features in the deeper zone. The mappings within the selected region (blue enclosing box) are directional, with arrows (yellow) pointing from each domain block (green box) to the respective range block(s) (red boxes). They were overlaid on the top of the image to aid visual inspection.³ Here, it is shown that, (i) most of the self-similar regions have a largely uniform appearance, (ii) the domain blocks are localised within a close proximity in the deep zone, and (iii) the corresponding domain and range blocks were statistically self similar, with relatively low distortion/error at a fine scale (depth = 4). From a signal/image processing perspective, this is the “self-transformability” assumption described by [6], where it was shown that such an assumption generally holds for images composed of isolated straight lines and/or constant regions (since these features are self-similar). Thus the ability of the fractal/IFS encoder which we developed to represent straight edges (parallel alignment of collagen fibres), constant regions (smooth uniform surfaces) and slow-varying/constant gradients (hyaline appearance) effectively is important; as most standard transform coders fail to take advantage of these types of spatial structures. Indeed, previously described wavelet transform techniques that have

³ To reduce the characteristic pink/red colour of immunostaining, a de-saturation process has been applied to the selected region.

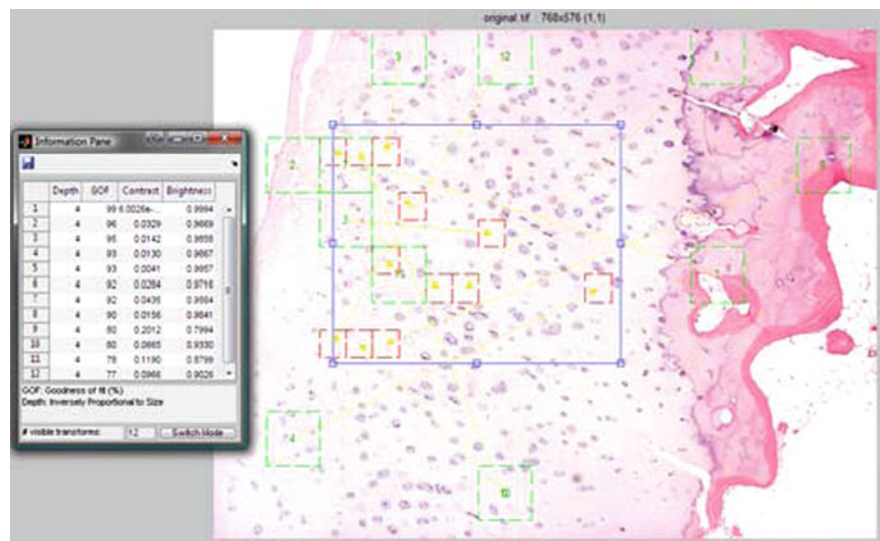


Fig. 95.7 Identifying visually correlated patterns through a ‘snap-shot’ of the dynamic events which occur in the repair process of cartilage. At 12 months the zonal heterogeneity of repair tissue is apparent

achieved particularly good compression results have done so by augmenting scalar quantization of transform coefficients with a zero-tree vector which could efficiently encode locally constant regions; see [19, 20].

In summary, our analysis confirmed a demonstrably higher proportion of hyaline-like cartilage than the relatively small amount previously reported in failed ACIs, in agreement with the results reported earlier using polarised microscopy.

95.6 Conclusions

Traditionally, our understanding of cell growth is all too often compartmentalised into distinct disciplines spanning microbiology, biochemistry and genetics. These provide complementary yet disparate information concerning how cells grow, evolve and reproduce and, more relevantly, how they may be coordinated at the physiological level. The image classification approach adopted in this research sought to significantly enhance our understanding of this, and has gone some way towards integrating physiological function with cell pattern formation; i.e., cell growth was examined from the viewpoint of computational fractal geometry using a “*structure–function*” approach. Surprisingly, however, little effort has since been expended on applying such reasoning to the study of biological cell re/production in general and characterisation of cartilage cell growth in particular. This work has attempted to address this deficit through utilising recent advances in multi-scale

fractal transformation research. Thus far the work has continued to promise the possibility of clinically useful objective analytic tools. To this end, the experimental system currently provides objective measures of cartilage cell quality as part of a comparative large-scale randomised clinical trial on the efficacy of the *Autologous Chondrocyte Implantation* (ACI) procedure.

Acknowledgments The authors would like to acknowledge the financial support of the EPSRC (UK) funded 3ME Initiatives at Keele. They would also like to thank *The Nuffield Foundation* (UK) for supporting related work on MATLAB in the area of visualisation. We are further indebted to Professor Sally Roberts and her team for provision of images and general support at RJAH, Oswestry.

References

1. Meinhardt H (1982) Models of biological pattern formation. Academic Press, London
2. Emery R, Lam KP, Collins DJ, Richardson JB (2009) FACELIFT—leveraging image fractal transforms, (Poster) IEEE international conference of information visualisation (IV09), July 2009
3. Fisher Y (1994) Fractal image compression, theory and application, Springer, New York
4. Barnsley M, Lyman PH (1993) Fractal image compression. AK Peters Ltd
5. Fisher Y, Jacobs EW, Boss RD (1992) Fractal image compression using iterated transforms, Image and Text Compression, Kluwer Academic Press, pp 35–61
6. Jacquin A (1992) Image coding based on a fractal theory of iterated contractive image transformation. IEEE Trans. Image Process 1:18–30
7. Russ JC (2007) The image processing handbook, 5th edn. Taylor and Francis
8. Gonzalez R (2006) Digital image processing. Addison-Wesley, New York
9. Linnell TA, Deravi F (2003) Novel fractal domain features for image classification, international conference on visual information engineering
10. Mandelbrot BB (1963) Fractal aspects of the iteration of $z \rightarrow \lambda z(1-z)$ for complex λ and z . Ann New York Acad Sci 351(1)
11. Lindenmayer A (1968) Mathematical models for cellular interaction in development I. Filaments with one-sided inputs. J Theor Biol 18:280–289
12. Barnsley M (1998) Fractals everywhere. Academic Press
13. JPEG homepage. <http://www.jpeg.org/jpeg/index.html>
14. Monro DM, Dubridge F (1992) Fractal approximation of image blocks. Proc IEEE ICASSP 3:485–488
15. Richardson RB, Caterson B, Evans EH, Ashton BA, Roberts S (1999) Repair of human articular cartilage after implantation of autologous chondrocytes, JBJS, 81-B:1064–1068
16. Roberts S, Richardson JB et al (2006) Assessment of cartilage repair tissue following autologous chondrocyte implantation by histochemistry and Fourier transform infrared imaging spectroscopy. OARSI,
17. Richardson J (2009) Private communication. RJAH
18. Witten IH, Eibe F (2005) Data mining, practical machine learning tools and techniques, 2nd edn. Elsevier, San Francisco
19. Davis G (1998) A wavelet based analysis of fractal image compression. IEEE Trans Image Process 7(2)
20. Koenderink JJ (1987) The structure of images, biological cybn's., vol. 50. Springer, Berlin, pp 363–37087

Chapter 96

A First Implementation of the Delay Based Routing Protocol

Eric Gamess, Daniel Gámez and Paul Marrero

Abstract In this paper, we introduce the first implementation of the Delay Based Routing Protocol (DBRP), a flexible distance vector routing protocol with a delay based metric, that supports authentication and encryption. Besides carrying routing information, DBRP also transports information for common services such as the IP addresses of the DNS servers. It operates over the data link layer and is centered on Type-Length-Value (TLV) tuples, which make it easy to extend to other existing or future network protocols. Our implementation is developed in C++, and offers a GUI developed in Qt for easy configuration. We validate this first implementation under several test scenarios and have obtained very encouraging results. Additionally, the exploratory study of our application that we conduce with a group of students indicates that the protocol can be easily understood and configured through the GUI.

96.1 Introduction

The Delay Based Routing Protocol (DBRP) is a new distance vector routing protocol that was introduced in [1]. It natively supports IPv4 and IPv6 [2–4], which is fundamental since the central pool of IPv4 addresses administrated by the

E. Gamess

Laboratorio de Comunicación y Redes, Central University of Venezuela, Caracas, 1040, Venezuela

E. Gamess (✉) · D. Gámez · P. Marrero

Escuela de Computación, Central University of Venezuela, Caracas, 1040, Venezuela

e-mail: eric.gamess@ciens.ucv.ve

D. Gámez

e-mail: daniel.gamez@gmail.com

P. Marrero

e-mail: paul.marrero@gmail.com

Internet Assigned Numbers Authority (IANA) has been depleted on February 3, 2011. Unlike other routing protocols which only transport routing data, DBRP also carries information for common services, such as the IP addresses of the DNS servers for a specific network. Security is addressed in DBRP, and network administrators can choose among different hash and cypher algorithms for a better authentication, integrity, and privacy. DBRP is based on Type-Length-Value (TLV) tuples, which make it very flexible and easy for extension. That is, by defined new TLV tuples, new common services can be added to DBRP, as well as new hash and cypher algorithms. Moreover, DBRP can be easily adapted to other routing protocols by defining new TLVs.

In this paper, we present the first implementation of DBRP called tinyDBRP. It is totally written in C++ and divided in two parts: an easy-to-use GUI-oriented configuration tool and a daemon. Through the different modules of the configuration tool, users can specify the network and DBRP parameters that are saved in the configuration file. At initialization time, the daemon reads the configuration file and runs accordingly.

The rest of the paper is organized as follows. In [Sect. 96.2](#), we present an overview of DBRP. In [Sect. 96.3](#), we introduce our implementation of the new routing protocol (tinyDBRP). We intensively validate our implementation in different test scenarios in [Sect. 96.4](#). In [Sect. 96.5](#), we discuss the result of a survey. Related work is presented in [Sect. 96.6](#). Finally, [Sect. 96.7](#) concludes the paper and presents the extension that we plan to do for DBRP.

96.2 An Overview of DBRP

This section presents a brief overview of DBRP. More in-depth information about the protocol can be found in [\[1\]](#).

96.2.1 *Metric of DBRP*

The main drawback of the metrics used in actual routing protocols, such as RIP [\[5\]](#) (hop count), OSPF [\[6, 7\]](#) (inversely proportional to the bandwidth of the links), and EIGRP [\[8\]](#) (composed metric involving bandwidth, delay, load, and reliability), is the limitation to factors that are only intrinsic to links. None includes aspects related with the routing devices per se, which are important since a router with limited resources will not have the same performance as a router with good resources. Hence, DBRP has a new simple metric that not only considers the links involved in the route, but also the routing devices. In DBRP, a weight is associated to all the links and routers in the network. The total metric of a particular route is the sum of the weight associated to each link and each router in the path. The weight of a router with few resources or with a lot of traffic must be high. The

Fig. 96.1 Position of DBRP in the OSI model

| DBRP |
|-----------------|
| Data Link Layer |
| Physical Layer |

weight of links is proportional to the total delay (serialization and propagation) to transmit typical frames. For each route, DBRP also tracks the number of hops and the path-MTU. In case of having two or more paths to a destination network with the same metric, the tie will be handled with the smallest hop count. If the hop count is still a tie, then load balancing must be done.

96.2.2 DBRP in an Ethernet Frame

As shown in Fig. 96.1, DBRP works on top of the data link layer. Thus, it is independent of the network layer and can be used to route any network layers. DBRPv1 is aimed for IPv4 and IPv6.

In broadcast multi-access networks, DBRP uses multicast to restrict the processing of its messages to DBRP-enabled routers only. Figure 96.2 shows a DBRP message in an Ethernet frame. The Destination MAC Address can be a unicast or multicast address. The Source MAC Address is the MAC address of the interface used to send the message. The *Ethertype* value 0×7777 indicates that a DBRP message is encapsulated in the Ethernet frame. The Cyclic Redundancy Check (CRC) is an error-detecting code designed to detect accidental changes in the transmission at the data link layer level.

96.2.3 TLV-Based Architecture

To make it more flexible, the architecture of DBRP is based on Type-Length-Value (TLV) tuples. This design dramatically facilitates the extension of DBRP. That is, the integration of new features in DBRP can be made by creating new TLVs. In Fig. 96.3, the general structure of a DBRP message is depicted. The message starts with a common header, followed by TLVs.

96.2.4 PDUs in DBRP

DBRP has three types of messages (update, request, and shutdown) also called Protocol Data Units (PDUs).

Fig. 96.2 A DBRP message in an Ethernet frame

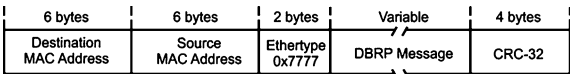
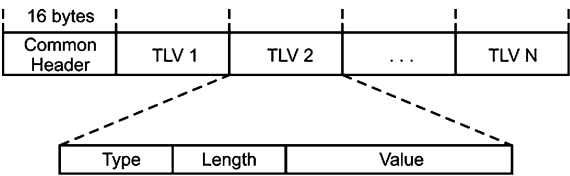


Fig. 96.3 General message of DBRP



- **Update:** similarly to other routing protocols, DBRP uses periodic update messages to propagate routing information. Unlike other routing protocols, DBRP also periodically sends information of common services through update messages. Update messages are sent as layer-2 multicast periodically (synchronous updates) or as unicast (asynchronous updates) in response to a request message. When a change occurs in the network, DBRP sends triggered update messages to accelerate the convergence process.
- **Request:** a router sends request messages to inform its neighbors of the protocols (IPv4 and/or IPv6 in DBRPv1) that it supports. It is a way to solicit routing and common services information for the locally supported network protocols.
- **Shutdown:** unlike other routing protocols, DBRP has shutdown messages to inform neighboring routers that the protocol has been disabled. The goal of the shutdown messages is to speed up the convergence process. The shutdown messages can be used in two different situations (specifically and globally). The specific usage is done when disabling DBRP on a particular DBRP-enabled interface. In this case, the router sends shutdown messages on that interface to inform its neighbors that it will not be part of the DBRP network through that interface anymore. It also has to send triggered updates on the other DBRP-enabled interfaces to inform about the change. On the other hand when DBRP is disabled at the level of a network protocol (IPv4 or IPv6), it sends shutdown messages on all of its DBRP-enabled interfaces.

A common header (see Fig. 96.4) is defined in DBRP for the three PDUs. It is the first part of the messages and is followed by TLVs which vary depending of the message's type. The first field, called *Version*, must contain the version of DBRP (currently version 1, also known as DBRPv1). *Code* identifies the PDU type. *Domain Identifier* is currently not used in DBRPv1. It will be employed in the next version of the protocol and will allow the division of huge networks in domains, for scalability purposes. In DBRPv1, all the routers belong to the same domain. For security purpose, users can enable different modes of operation in DBRP: (1) no authentication and no encryption, (2) authentication but no encryption, and (3) authentication and encryption. The A-flag (Authentication) and E-flag (Encryption) indicate the security mode currently used. *Number of TLVs* indicates the number of TLVs included in the messages. *Checksum* provides a basic verification that the information has been transmitted correctly. It is a way to detect

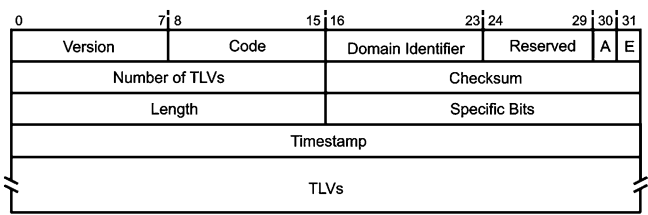


Fig. 96.4 DBRP common header

transmission errors due to noise. If the checksum fails, the message must be discarded. *Length* is the total length in bytes of the PDU. The usage of the field *Specific Bits* varies depending of the PDU’s type. The goal of the field *Timestamp* is the protection against replay attacks. It is defined as the number of seconds elapsed since midnight Coordinated Universal Time (UTC) of January 1, 1970.

96.2.5 Authentication and Encryption

DBRPv1 implements three authentication algorithms (Clear Text, MD5 [9], and SHA-1 [10]) and three encryption algorithms [11] (DES, 3DES, and AES). There are three modes of operation: (1) no authentication and no encryption, (2) authentication but no encryption, and (3) authentication and encryption. In the first mode (no authentication and no encryption), authentication and encryption are disabled. Therefore, it has no security overhead and can be used in small networks where threads are unlikely or security measures are taken at another level. The second mode (authentication but no encryption) authenticates but does not encrypt. It is the recommended mode for most networks. In the last mode (authentication and encryption), both authentication and encryption are enabled. It can be set up in networks where security threads are common. An Authentication TLV can be inserted in all the DBRP’s messages (update, request, and shutdown). However, the Encryption TLV can only appear in an update message, since it is the only one that transport sensitive data.

96.2.6 Common Services

DBRPv1 has two Common Services TLVs (IPv4 DNS TLV and IPv6 DNS TLV). The idea is to carry, up to the routers that are connected to a specific network, information that can help in the configuration. In this case, DBRP carries the IP addresses of the DNS servers. The main goal is to simplify the network administration, that is, the DNS information can be propagated using DBRP and locally fed into a DHCP server that will distribute it to the computers connected to the

LAN. If the DNS servers are changed, the new IP addresses of the servers will be propagated through DBRP and reach the DHCP server that will automatically update this information at the level of the computers connected to the LAN.

96.3 Our Implementation of DBRPv1

tinyDBRP, our implementation of DBRPv1, is divided in two entities: a daemon and an easy-to-use GUI-oriented configuration tool called *DBRP Settings*. The daemon is in charge for managing the routing information and tables. The objective of *DBRP Settings* is the easy modification of the configuration file, read by the daemon at initialization time. Seven modules are available for configuration namely: (1) Interfaces, (2) Routing Settings, (3) Authentication, (4) Encryption, (5) Tables, (6) DNS IPv4 Services, and (7) DNS IPv6 Services.

tinyDBRP was developed in Debian GNU/Linux using C++ for both x86 and x86_64 architectures. We implemented the communication between routers with raw sockets, which makes tinyDBRP independent from the network layer. We also used the following libraries in the development:

- Qt: a cross-platform application framework that is widely used for developing application software with a GUI. Qt is a free and open source software distributed under the terms of the GNU Lesser General Public License.
- Qt-Designer: it is a Qt complement for designing and building GUIs from Qt components. With Qt-Designer, users can easily compose and customize widgets and dialogs in a what-you-see-is-what-you-get (WYSIWYG) manner, and generate the associated code.
- OpenSSL [11]: is an open source library that implements the SSL (Secure Sockets Layer) and TLS (Transport Layer Security) protocols. It provides basic cryptographic functions and various utility functions.
- Wireshark: originally named Ethereal, Wireshark is a free and open source packet analyzer. It allows users to capture the network traffic. It is widely used in education and by network specialists for troubleshooting, analysis, and software and communications protocol development.

96.3.1 DBRP Daemon

The daemon implements DBRP and is responsible for propagating routing and common services information, as well as maintaining the protocol's tables. The daemon reads the initial configuration from a text file called *dbrp.conf*. The file can be written with a text editor or generated by *DBRP Settings*. This file indicates which network interfaces will be used by DBRP and which network protocols will be routed. For each network interface, two threads are created for sending and receiving

messages, respectively. Another thread modifies the routing table of the operating system when DBRP messages arrive. It was necessary to synchronize the routing information handled by the daemon with *DBRP Settings*. To achieve this synchronization, additional threads were implemented in each of the parties. A first thread that runs in the daemon is in charge to serialize and encapsulate the routing tables and send them to *DBRP Settings* via Unix sockets for local communication. A second thread that runs in the configuration tool is responsible to receive the encapsulated routing tables, deserialize them, and finally update the information in *DBRP Settings*.

96.3.2 Interfaces Module

The Interfaces module is shown in Fig. 96.5. It allows users to configure the network and DBRP parameters for all interfaces. That is, users can specify the IPv4 and IPv6 addresses, the netmask length, the status (up or down), and the delay associated to DBRP for each interface. The module will validate the correct format of the introduced data and indicate possible overlapped networks. Also, this module permits the creation and deletion of virtual interfaces, also known as dummy interfaces, which are very useful for debugging, learning, and teaching purposes. Additionally, the activation and deactivation of DBRP at the interface level are managed here.

96.3.3 Routing Settings Module

In this module, users specify the *Router Delay* and *Domain ID*. *Router Delay* is assigned in nanoseconds to reflect the congestion that experiences a router depending on its power. *Domain Identifier* is defined in DBRPv1 but not actually meaningful. It will be used in DBRPv2 for scalability purpose by segmenting a large network in smaller networks, called domains. In DBRPv1, all routers must belong to the single domain; hence, they must have the same value for *Domain Identifier*.

96.3.4 Authentication Module

DBRPv1 implements three authentication algorithms (Clear Text, MD5 [9], and SHA-1 [10]). For each hash algorithm, users can easily specify the list of associated keys. Keys can be added, removed, modified, or reordered as needed as depicted in Fig. 96.6. For a successful communication, each router must have the same list of keys for the selected algorithm.

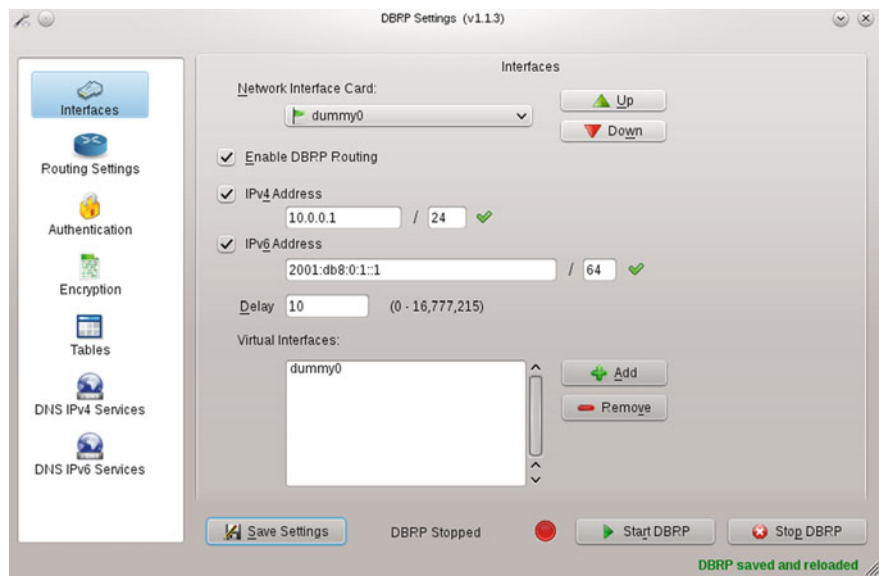


Fig. 96.5 Interfaces module

96.3.5 Encryption Module

The standard interior routing protocols (e.g., RIP [5], OSPF [6, 7], IS-IS [6, 12], and EIGRP [8]) do not allow encryption, which can be a limitation with the growing security issues. DBRPv1 does permit it through a choice of three encryption algorithms (DES, 3DES, and AES) as shown in Fig. 96.7. In the case of 3DES, users must specify a set of three sub-keys for each actual key. For a successful communication, all the routers of the routing domain must have the same key list for the selected algorithm. Keys can be added, removed, modified, or reordered as needed. The only encrypted data is the one in the Update PDU, since it contains sensitive routing and common services information.

96.3.6 Tables Module

In this module, users can display the IPv4 and IPv6 routing tables as well as the Common Services DNS IPv4 and DNS IPv6 tables, by selecting each one from a combo-box as shown in Fig. 96.8. These tables are updated regularly by the protocol in execution. The IPv4 and IPv6 routing tables show all directly connected networks and those learned through DBRP, as well as their prefix length, metric, and network interface. For the common services table, the module shows the prefix, prefix length, priority, and DNS servers list.

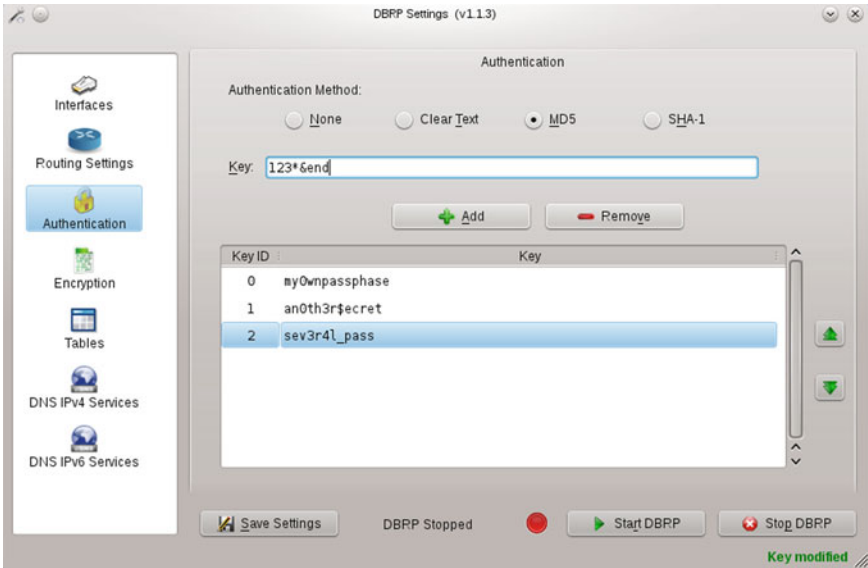


Fig. 96.6 Authentication module

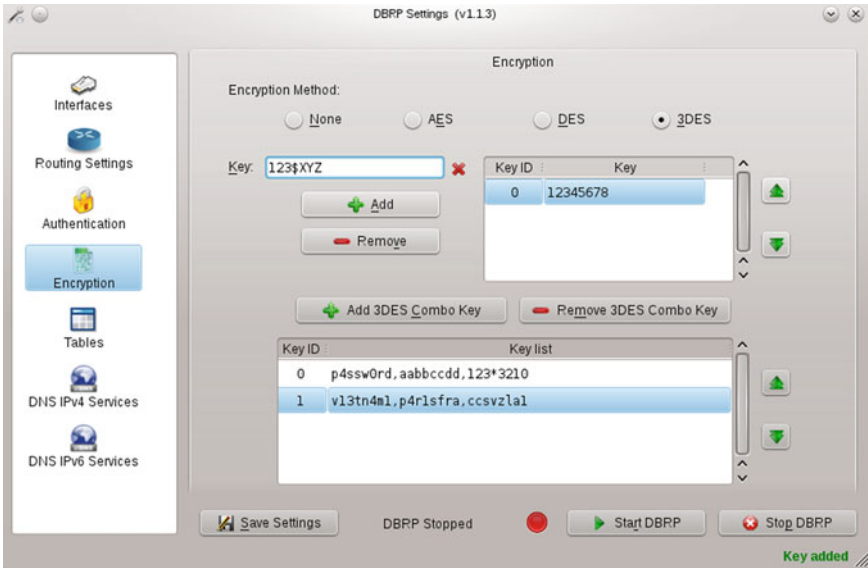


Fig. 96.7 Encryption module

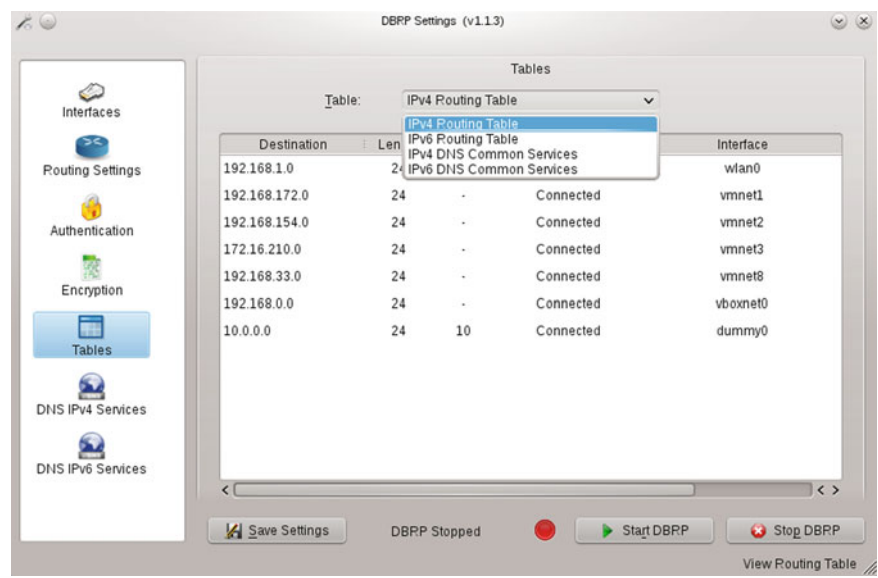


Fig. 96.8 Tables module

96.3.7 DNS IPv4 Services and DNS IPv6 Services Modules

DBRPv1 transports common services such as the list of DNS servers destined to specific networks. With this module, it is possible to configure the destination network (prefix ID and prefix length) to which DNS servers are directed. For a better configuration control, the module validates the format of the entered data. For each network that must receive DNS information, users can establish the priority of DNS servers, add and delete DNS server addresses to/from the list, edit a DNS server from the list by double clicking on it, change the order of DNS servers, as shown in Fig. 96.9.

96.4 Validation Tests

This section presents the experiments that we did to evaluate tinyDBRP and a plugin developed to analyze DBRP traffic with Wireshark.

96.4.1 Test Scenarios

In order to validate DBRP operation, we proposed three scenarios in which certain aspects were evaluated. For each scenario, the following activities were done:

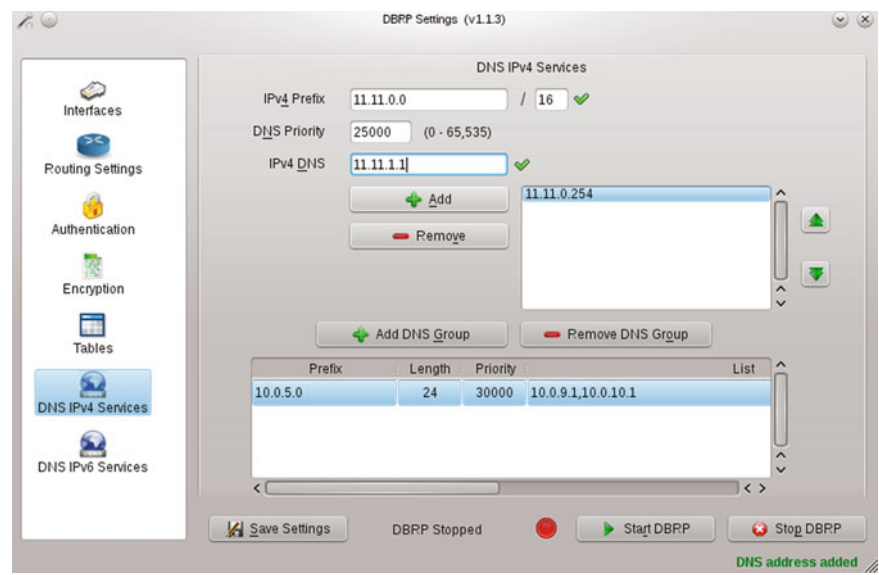


Fig. 96.9 DNS IPv4 services module

- Modifying IP addresses and network masks of virtual and physical interfaces.
- Enabling and disabling IPv4 and/or IPv6 stacks in virtual and physical interfaces.
- Configuring different authentication and encryption keys and algorithms.
- Modifying the value of the metric in physical and virtual interfaces.
- Changing the topology by putting up and down physical and virtual interfaces unexpectedly.
- Changing the topology by stopping routers unexpectedly.

Figure 96.10 shows the first scenario which consists of a ring topology. We chose this topology since it has a loop and is common in production networks. In the tests, we incremented the size of the ring from 5 to 10 routers.

Figure 96.11 shows the second scenario with BMA (Broadcast Multi-Access) networks. In the test, we incremented the number of routers in the BMA networks from 3 to 7.

Figure 96.12 shows the third scenario where a simple loop was introduced. In the three test scenarios, tinyDBRP showed the expected behavior.

96.4.2 Wireshark Plugin

In order to make a detailed analysis of network traffic generated by DBRP, we developed a plugin for Wireshark [13] to allow the dissection of DBRP's

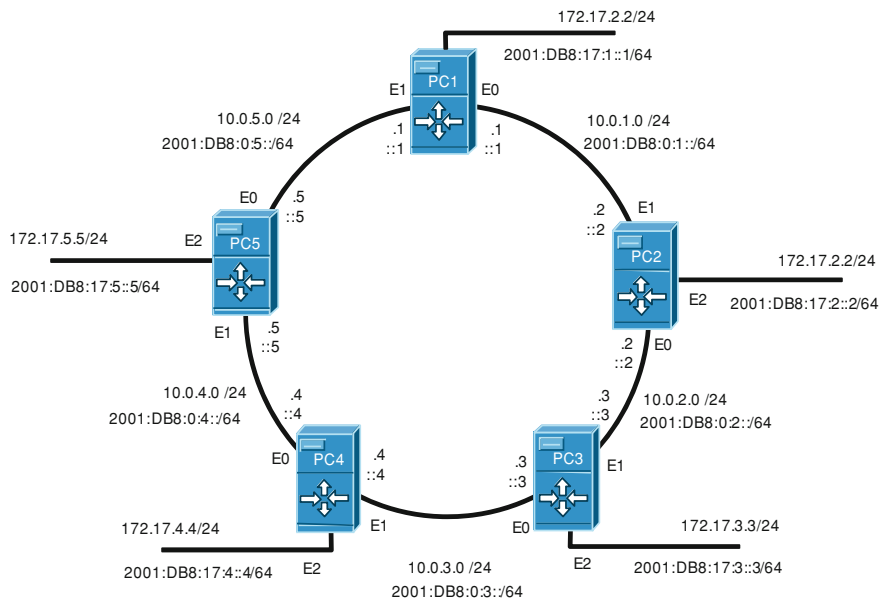


Fig. 96.10 Ring topology

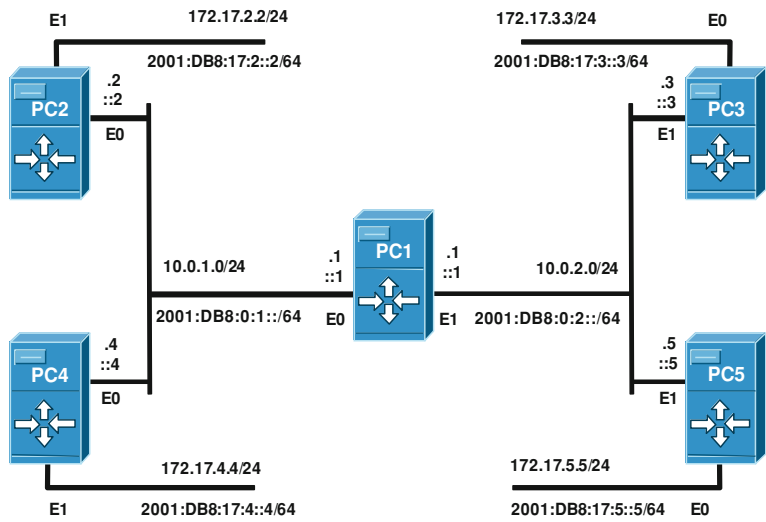


Fig. 96.11 BMA topology

messages, as shown in Fig. 96.13. A DBRP message is filtered through the *Ethertype* field of Ethernet with value 0x7777. Wireshark shows the fields of the PDU as a hierarchical tree, with a brief description. The use of this plugin

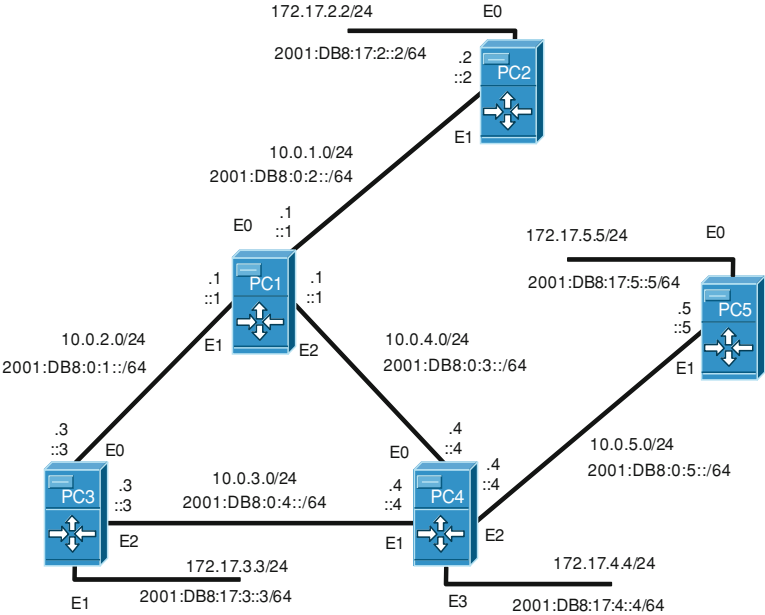


Fig. 96.12 Simple loop topology

dramatically helped us in the implementation of the protocol and is very useful for teaching and learning purpose.

96.5 Survey Study

An exploratory study of tinyDBRP was conducted to evaluate the effectiveness of the first implementation of DBRP in a group of 40 students with different knowledge in terms of networking concepts, ranging from no knowledge at all to advanced skills. The study was carried out in the School of Computer Science of our University (Universidad Central de Venezuela, Caracas, Venezuela). We created a virtual machine for each virtual router to run the complete experiments in a unique PC. We used VirtualBox,¹ a famous virtualization solution that is freely available as open source software. We used the scenarios of Figs. 96.10, 96.11, and 96.12. For each exercise, students had to configure the different routers using *DBRP Settings*, capture PDUs (Update, Request, and Shutdown), change the authentication and encryption parameters, study the propagation of routing information and common services (IP addresses of DNS servers), and understand how convergence was reached after a change in the network topology. The

¹ <https://www.virtualbox.org>

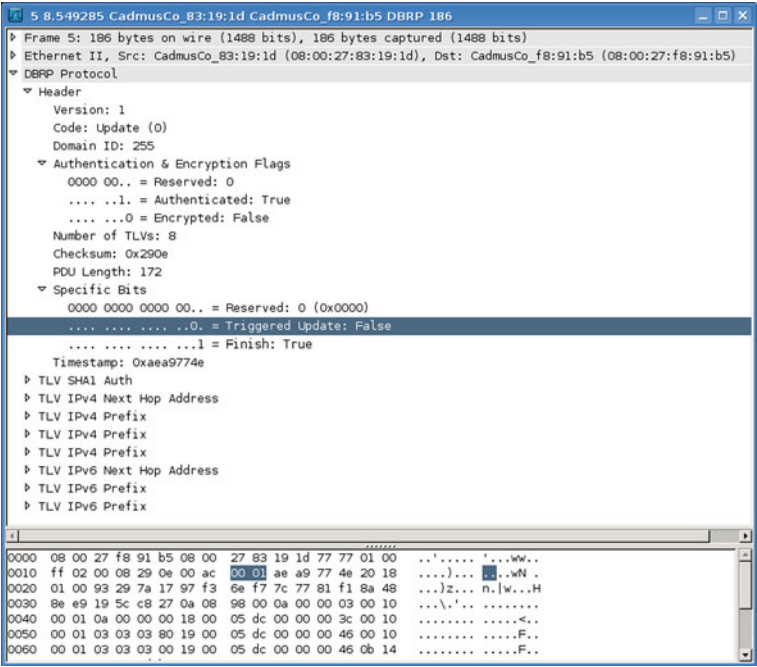


Fig. 96.13 Capture of an update PDU with the Wireshark plugin

students had to rate different aspects of the protocol and the seven configuration modules with a score ranging from 1 to 5 marks, where 1 mark was the lowest score acceptance and 5 marks the highest.

Figure 96.14 shows the results of the evaluation at the level of the protocol for the three scenarios (Ring, BMA, and Simple Loop). Question 1 evaluated the simplicity of the protocol. Question 2 is focused on PDUs (design, scalability, and capture with Wireshark). Question 3 was related to the expected behavior of DBRP with dynamic changes, and finally the goal of question 4 was to assess the convergence time.

Figure 96.15 shows the result of the evaluation related to the 7 modules (Interfaces, Routing Settings, Authentication, Encryption, Tables, DNS IPv4 Services, and DNS IPv6 Services) of *DBRP Settings*. Question 1 was related to the look-and-feel and element distribution in the modules. Question 2 evaluated the validation done by the modules on data entered by the users. Question 3 assessed the help messages, and finally, the goal of question 4 was to estimate how intuitive the usage of the modules was.

The results obtained in the survey are very encouraging and seem to indicate that tinyDBRP is easy to configure and shows a rapid convergence after a network topology change.

Fig. 96.14 Results of the survey for the protocol

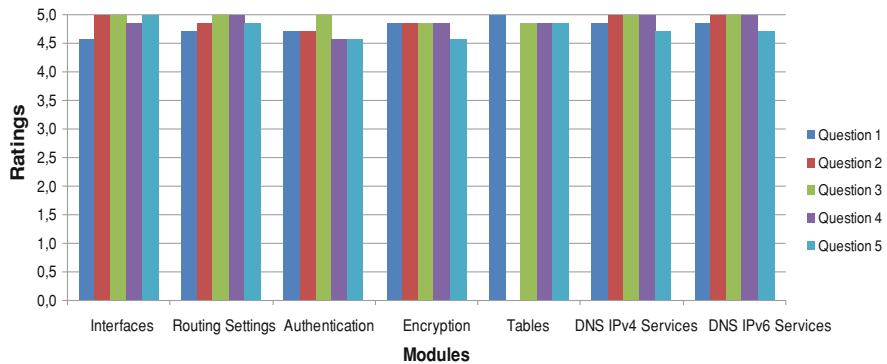
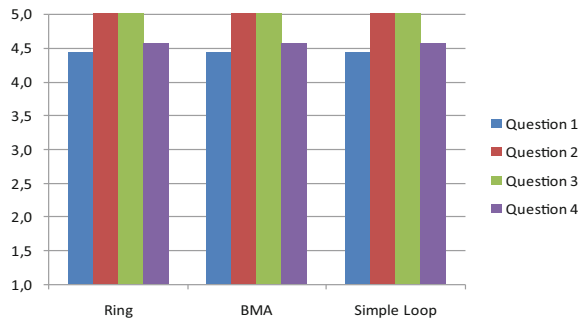


Fig. 96.15 Results of the survey for the modules

96.6 Related Work

There is no direct related work since DBRP is a new routing protocol recently proposed [1]. Our implementation is the very first implementation of the protocol. We wish that some other research group will develop another implementation so we can realize some compatibility tests.

96.7 Conclusions and Future Work

In this paper we introduced tinyDBRP, a first implementation of the Delay Based Routing Protocol (DBRPv1) that was recently proposed in [1]. tinyDBRP is divided in two entities: a daemon and an easy-to-use GUI-oriented configuration tool. We also developed a Wireshark plugin for the dissection of the PDUs, which dramatically helped in the implementation of the protocol. The validation tests and the survey study that we did seems to indicate that DBRP can be used as an alternative to RIP, OSPF, and IS-IS, which is easy to setup. tinyDBRP and the

Wireshark plugin can be downloaded from SourceForge (<http://sourceforge.net/p/dbrp>).

As future work, we plan to further investigate the scalability of DBRP. To achieve our goal, we propose to simulate large networks since real validations are too costly in time and resources. We will also study the possibility of dividing large networks in domains, to limit the propagation of update messages. For this reason, the field *Domain Identifier* is defined in the Common Header of DBRPv1 but not used yet.

Acknowledgments We want to thank the CNTI (Centro Nacional de Tecnologías de Información) which partially supported this research.

References

1. Gamess E, Marrero P, Gámez D (2011) A delay based routing protocol with support for common services. In: Proceedings of the 6th international IFIP/ACM Latin American networking conference 2011 (LANC '11). Quito, Ecuador, October 2011
2. Davies J (2008) Understanding IPv6, 2nd edn. Microsoft Press, USA
3. Deering S, Hinden R (1998) Internet protocol, Version 6 (IPv6) Specification. RFC 2460
4. Gamess E, Morales N (2007) Implementing IPv6 at Central University of Venezuela. In: Proceedings of the 4th international IFIP/ACM Latin American networking conference 2007 (LANC '07), October 2007
5. Malkin G (1998) RIP Version 2. RFC 2453, November 1998
6. Doyle J (2005) OSPF and IS-IS: choosing an IGP for large-scale networks, 1st edn. Addison-Wesley Professional,
7. Moy J (2008) OSPF complete implementation, 1st edn. Addison-Wesley Professional,
8. Expósito J, Trujillo V, Gamess E (2010) Easy-EIGRP: a didactic application for teaching and learning of the enhanced interior gateway routing protocol. The 6th international conference on networking and services (ICNS 2010), Cancun, Mexico. May 2010
9. Rivest R (1992) The MD5 message-digest algorithm. RFC 1321, April 1992
10. Eastlake D, Jones P (2001) US secure hash Algorithm 1 (SHA1). RFC 3174
11. Davies J (2011) Implementing SSL/TLS using cryptography and PKI, 1st edn. Wiley,
12. Gredler H, Goralski W (2004) The complete IS-IS routing protocol, 1st edn. Springer,
13. Lamping U (2010) Wireshark developer's guide for Wireshark 1.7. Free software foundation

Chapter 97

Different Aspects of Data Stream Clustering

Madjid Khalilian, Norwati Mustapha, Md Nasir Sulaiman
and Ali Mamat

Abstract Nowadays the growth of the datasets size causes some difficulties to extract useful information and knowledge especially in specific domains. However, new methods in data mining need to be developed in both sides of supervised and unsupervised approaches. Nevertheless, data stream clustering can be taken into account as an effective strategy to apply for huge data as an unsupervised fashion. In this research we not only propose a framework for data stream clustering but also evaluate different aspects of existing obstacles in this arena. The main problem in data stream clustering is visiting data once therefore new methods should be applied. On the other hand, concept drift must be recognized in real-time. In this paper, we try to clarify: first, the different aspects of problem with regard to data stream clustering generally and how several prominent solutions tackle different problems; second, the varying assumptions, heuristics, and intuitions forming the basis of approaches and finally a new framework for data stream clustering is proposed with regard to the specific difficulties encountered in this field of research.

M. Khalilian (✉) · N. Mustapha · M. N. Sulaiman · A. Mamat
Faculty of Computer Science and Information Technology, University Putra Malaysia,
Bintulu, Malaysia
e-mail: khalilian@ieec.org

N. Mustapha
e-mail: norwati@fsktm.upm.edu.my

M. N. Sulaiman
e-mail: nasir@fsktm.upm.edu.my

A. Mamat
e-mail: ali@fsktm.upm.edu.my

97.1 Introduction

During the last decade many applications have been developed that they should manage massive amount of data which causes limitation in data storage capacity and processing time. Furthermore, many applications must operate in real-time to achieve their objectives. As an important case for these kinds of application, network intrusion detection system (NIDS) can be pointed where it generates a huge data and this data should be process in real time to discover suspicious data. However, we identify some difficulties against this problem:

- 1 Stream data may only be visit once.
- 2 Algorithm should be operated in resources constraints.
- 3 The number and shape of clusters may be unknown in advance and the characteristics of clusters may change over time.
- 4 Random noise must be managed.
- 5 Different granularities of clustering can be revealed.

It is desirable to have algorithms which are able to detect clusters of objects with evolving intrinsic with considering this point that visiting of data is possible once. In addition, resource constraints and outliers detection are others aspects. Therefore, main objective for this paper is proposing a frame work to overcome these problems. Moreover, the parameters, extracting general components and determining suitable quality measurements will be studied.

97.2 Background

If we want to categorize data stream clustering methods, we will recognize two main aspects for grouping approaches, which one group refers to type of solution includes task oriented and data oriented. Another group of approaches refers to solutions techniques which are component based and non-component based methods. In continue we review these techniques with their pros and cons.

97.2.1 *Component Based Methods*

Generally, these methods are based on two main components: on-line and off-line components; consequently, stream clustering take places in two steps on-line and offline. This kind of framework was proposed for the first time by [1] and in continue they have applied this framework to solve other problems in data stream clustering [2–6]. As it mentioned before main problems in data stream clustering are visiting data once and concept drift. Thus, Micro clustering and Macro clustering are utilized in two main components. It also employs a pyramid structure for

organizing macro clusters during the time. Because of this, it is possible to answer user's question during tilted time. However, experimental results have demonstrated acceptable accuracy and efficiency. In absence of certainty, accuracy would be decreased; therefore, [4] proposed the UMicro algorithm for clustering uncertain data streams. It has the clear effectiveness with comparison of CluStream. In addition, high dimension data and different data type such as categorical data are some minor problems around data stream clustering, for this purpose they improved primary framework [2, 3, 6]. Generally speaking, approaches which are applied K-Means or K-Medians suffer from lack of accuracy when there are a lot of outliers. Beside, K-Means is also sensitive to value of outliers. These methods are not suitable for discovering clusters with non-convex shapes or clusters of very different size. In addition, number of clusters should be determined as value of parameter K. Aforementioned weaknesses motivated researchers to employ some other techniques, e.g. [7] have developed a connectivity based comprehensive points to cluster data stream. Online Component includes:

- 1 Adding arrival point to the sparse Graph
- 2 If it is reciprocally connected to representative vertex then joins an existing clusters otherwise it is considered as exemplar or predictor (it is based on $NN(V_i)$)

And off-line component includes:

- Using AVL tree structure for search representative vertices
- Using usefulness parameter to update of repository based on decay concept.

$Usefulness(r_i, count) =$

$\log(\lambda) \cdot (currentTime - creationTime(r_i) + 1) + \log(count + 1)$

Accuracy is outstanding in their research but it exhibits low performance. Another disadvantage refers to use a repository for previous data; thus, it is unable to give us a history in different scale time. In sum main problems as follows:

- Managing lots of pointers in memory.
- Violent memory constraint condition.
- Complexity in its programs.
- Lots of parameters which must be determined.
- Using fixed value for decay in fading function for offline component. It employs a priority FIFO queue to manage evolving data (may be a cluster archived whereas in near future some new data points arrive).

Reference [8] proposed a new framework for online monitoring clusters over multiple evolving streams by correlations and events. The streams are smoothed by piecewise linear approximation, and each end point of the line segment can be regarded as a trigger point. At each trigger point, for clusters that have trigger streams, they update the weighted correlations related to trigger streams in clusters. Whenever an event happens, the clusters are modified through efficient split and merge processes. In [9] a new entropy based method has been developed for

mixed numeric and categorical data stream clustering. They also use online and off line components to process data.

Reference [10] has presented a clustering system for streaming time series and uses a top-down strategy to construct a binary tree hierarchy of clusters, with the goal of finding highly correlated sets of variables. A common measure of cluster quality is the cluster's diameter, which is defined as the highest dissimilarity between objects of the same cluster. The system evolves by continuously monitoring the clusters' diameters.

The examples are processed as they arrive by using a single scan over the entire data. The system incrementally computes the dissimilarities between time series, maintaining and updating the sufficient statistics at each new example arrival, updating only the leaves. The splitting criterion is supported by a confidence level given by the Hoeffding bound, which is detected when the system has gathered enough information to confidently define the diameter of each individual cluster. The system includes an agglomerative phase, based on the diameters of existing clusters, also supported by the Hoeffding bound. The aggregation phase enables the adaptation of the cluster structure to smooth changes in the correlation structure of time series.

97.2.2 Non-Component Based Methods

BIRCH can be considered a primitive method in this area [11]. In fact it has been designed for traditional data mining but it is suitable for very large data base so it has been applied for data stream mining. This method introduces two new concepts: micro clustering and macro clustering. Based on these two concepts it could overcome two main difficulties in agglomerative method in clustering: scalability and the inability to undo what was performed in the previous step. It works in two steps: first it scans data base and builds a tree which includes information about data clusters. In second step BIRCH refines tree by removing sparse nodes as outliers and creates original clusters. The main disadvantage of this method is the limitation in capacity of leaf. If clusters are not spherical in shape, BIRCH does not perform well because it uses the notion of radius or diameter to control the boundary of a cluster.

STREAM is the next main method which has been designed especially for data stream clustering [12]. In this method K-Medians is leveraged to cluster objects with SSQ criterion for error measuring. In the first scan objects grouped and medians of each group is gathered and associated them a weight with regard to the number of objects in the cluster. In second step these medians is clustered until top tree. We can realize two main disadvantages for this method: time granularity and data evolving.

In general, whenever there are n_i medians at level i they are clustered to form level $(i + 1)$ medians as it is depicted in Fig 97.2.

Fig. 97.1 STREAM clustering algorithm

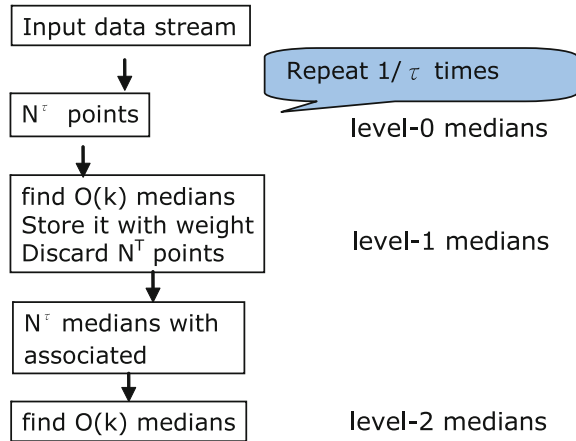
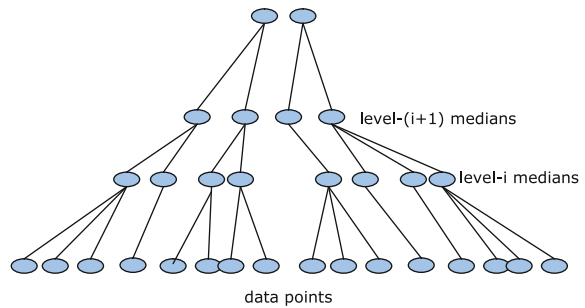


Fig. 97.2 Structure of data in STREAM process



There are some other approaches in which is not utilized online and off-line components while they process data stream [13–16].

97.3 Data Stream Clustering Problems

Generally speaking, we have two groups of problems: major and minor problems. Major problems include problem A and B:

- 1) Problem A :Scan data once
 - a) Solution A1: using online, offline components (e.g. most methods such as CLUstream).
 - b) Solution A2: Data sampling e.g. STREAM
- 2) Problem B :Evolving data
 - a) Solution B1: Using fading model (decay value).
 - b) Solution B2: A tree structure has been employed e.g. BIRCH.

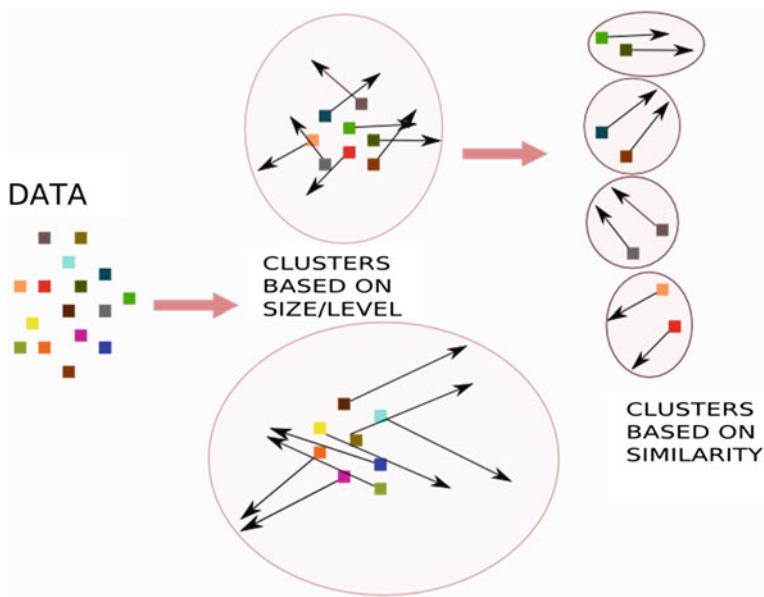


Fig. 97.3 Dividing arrival data to same size subsets

Fig. 97.4 Quality comparisons for algorithms

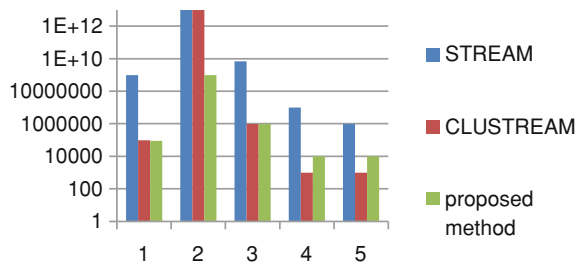
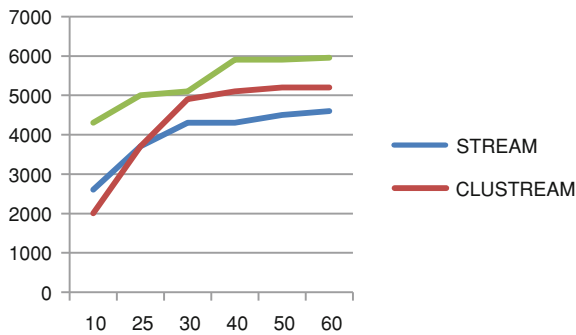


Fig. 97.5 Speed up comparison for algorithms



3) Solution A1:

- a) Pros: Having summary of data (global view)
- b) Cons; Resource constraints, speed up.

4) Solution A2:

- a) Pros: increasing in speed up.
- b) Cons: decreasing in quality.

5) Solution B1:

- a) Pros: Managing evolving data efficiently.
- b) Cons: Finding suitable value for threshold.

6) Solution B2:

- a) Pros: Don't need to determining extra parameters e.g. threshold value.
- b) Cons: Inflexibility

In addition we have some minor problems in data stream clustering:

- High dimensional data
- Detecting noise and outliers
- Space constraints
- Uncertainty data
- Different data types
- Spherical shaped clusters vs. arbitrarily shaped clusters
- Number of determined parameters

97.4 Proposed Framework

We briefly describe the design of proposed framework in this study. Our proposed framework includes two main components: on-line and off-line. In on-line module we identified five sub-modules as below:

- Data Preprocessing: having a good result is related directly to input data. Variables with different data type like categorical, numerical and ordinal should be determined and converted to numerical for using K-Means. On the other hand missing values have to be processed and replaced with suitable values. Attributes which are effective in length of vector should be selected and determined. Managing mix type datasets can be done in this module.
- Data Normalizing: data value in different range caused some difficulties in our method, so we need to normalize all variables that participate in measuring vector length. If attributes' values haven't been normalized, we won't be able to compare and use length of vector in correct way.

- **Data Dividing:** We divide arrival data in stream into some subsets by K-Means algorithm based on length of vector which is equivalency relation. Length of vectors are input for K-Means algorithm and output will be some partitions which elements inside them are equivalent and ready to clustering. In other word all samples in one partition have almost same size but might be dissimilar.
- **Subsets clustering:** after finding subsets, clustering algorithm is applied on each subset and outcomes final group. Although samples in different subsets may be similar based on COSIN criterion but they are in different levels [17].
- **Clusters Validity:** quality of clusters should be demonstrated. K-Means uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm moves objects between clusters until the sum cannot be decreased further. The result is a set of clusters that are as compact and well-separated as possible. You can control the details of the minimization using several optional input parameters to K-Means, including ones for the initial values of the cluster centroids, and for the maximum number of iterations. To get an idea of how well-separated the resulting clusters are, we can make a silhouette plot using the cluster indices output from K-Means. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. This measure ranges from +1, indicating points that are very distant from neighboring clusters, to 0, indicating points that are not distinctly in one cluster or -1, indicating points that are probably assigned to the wrong cluster. A more quantitative way to compare the two solutions is to look at the average silhouette values for the two cases. Average silhouette width values can be interpreted as follows: [18]

| | |
|----------|---|
| 0.7–1.0 | A strong structure has been found |
| 0.5–0.7 | A reasonable structure has been found |
| 0.25–0.5 | The structure is weak and could be artificial |
| <0.25 | No substantial structure has been found |

After validating micro clusters by mentioned criteria, we summarize statistics about each cluster including: number of elements in each cluster, mean, variance, compactness and separateness for each cluster.

97.5 Divide and Conquer K-Means Algorithm

Our proposed algorithm for on-line component includes two main steps:
Subsets Dividing:

1. compute length of vector for all samples in each group of data
2. for $I = 2$ to k find clusters with max value for average of silhouette, If this value is less than .25 then ignore subset dividing and return

3. return I as number of subsets and divide problem space into I subsets

Subsets Clustering:

1. if $I = k$ then return space dividing as final clustering result
2. for each subsets s_i :
 - a) $L_i = 2$; K-Means (s_i, L_i)
 - b) Compute silhouette value for clusters inside subset
 - c) If mean of silhouette value is less than .25 for one cluster inside subset and total number of clusters $\leq k/I$ then cluster again for this sub cluster

97.6 Experimental Results

We compare our method According to its quality and speed up using a specific data set. In this section data set is described and proposed algorithm is compared with STREAM and CLUStream considering quality and speed up of clustering.

97.6.1 Data set Description

The 1998 DARPA Intrusion Detection Evaluation Program was prepared and managed by MIT Lincoln Labs. The objective was to survey and evaluate research in intrusion detection. A standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment, was provided. The 1999 KDD intrusion detection contest uses a version of this dataset. The raw training data was about four gigabytes of compressed binary TCP dump data from seven weeks of network traffic. This was processed into about five million connection records. Similarly, the two weeks of test data yielded around two million connection records.

It is important to note that the test data is not from the same probability distribution as the training data, and it includes specific attack types not in the training data. This makes the task more realistic. Some intrusion experts believe that most novel attacks are variants of known attacks and the “signature” of known attacks can be sufficient to catch novel variants. The datasets contain a total of 24 training attack types with an additional 14 types in the test data only. The research intends to compare efficiency of proposed frame work with other methods in this area under different era, and KDD Cup 99 is too huge, various data is distributed unevenly, consequently the research will sample 10 % (494000 sample and 47 types of network connection characteristic in each kind of network connection record) and test dataset.

We should also apply main task of preprocessing includes: managing missing value, normalizing data, managing different data types and converting for using by the algorithm of detection, selecting the most important of attributes (reduction) and grouping data (converting to stream data).

97.6.2 Experimental Setup

In all experiments we use MATLAB software as a powerful tool to compute clusters and windows XP with Pentium 2.1 GHZ. The K-Means, hierarchical clustering and proposed algorithms are applied on the above mentioned data set. As a similarity metric, Euclidean distance has been used in each algorithm.

97.6.3 Quality and Speed up Compression

In this section we demonstrate our experiments result for clusters quality. Efficiency improvements are illustrated according to following graphs from quality and speed up points of view which are based on sum of square criterion and number of processing points per unit time, respectively.

Our contributions in this research are:

- Firstly, better quality in comparison of previous works and traditional methods. Managing evolving data for dramatic changes i.e. ID dataset.
- Secondly, using vector model as a base for data stream clustering; therefore, reducing in size and complexity of problem.
- Thirdly, Good scalability in terms of stream size, dimensions and number of clusters.
- Lastly, Improvement in speed up and reliability is considerable because of employing vector model.

97.7 Conclusions

In this paper we have demonstrated some difficulties in data stream clustering where its data mostly are high scale and dimensions; consequently, new methods need to be developed for processing these huge data sources. Furthermore concept drift is nature of data and should be managed by new methods. On the other hand, efficiency in terms of accuracy is one of the most critical measurements which are mostly defined by compactness and separateness for those data that their labels are unknown (for known labels we can use precision and recall). Therefore, we need to design efficient algorithms whereas scan data once and extract hidden patterns inside it. Evolving data, visiting data once, accuracy and space limitations are

major issues in data stream clustering. However, devising new framework with combining of advantages in two main approaches can overcome most drawbacks.

Acknowledgments This work was supported by grant 03-04-10 875FR from the Basic Research Program of the University Putra Malaysia.

References

1. Aggarwal CC, Han J, Wang J, Yu PS (2003) A framework for clustering evolving data streams, VLDB Conference, 81–92
2. Aggarwal CC, Yu PS (2006) A framework for clustering massive text and categorical data streams, ACM SIAM Data Mining Conference
3. Aggarwal CC, Han J, Wang J, Yu PS (2004) A framework for projected clustering of high dimensional data streams. In: Proceedings of the thirtieth international conference on very large data bases, vol 30. Publisher VLDB Endowment, pp 852–863
4. Aggarwal CC, Yu PS (2008) A framework for clustering uncertain data streams. In: Proceedings of the 24th international conference on data engineering, CDE 2008, pp 150–159
5. Aggarwal C (2009) A framework for clustering massive-domain data streams. In: Proceedings of the 2009 IEEE international conference on data engineering, pp 52–61
6. Aggarwal CC (2009) On high dimensional projected clustering of uncertain data streams. In: Proceedings of the IEEE 25th international conference on data engineering, ICDE 2009, pp 1152–1154
7. Lühr S, Lazarescu M (2009) Incremental clustering of dynamic data streams using connectivity based representative points. *Data Knowl Eng* 68:1–27
8. Yeh MY, Dai BR, Chen MS (2007) Clustering over multiple evolving streams by events and correlations. *IEEE Trans Knowl Data Eng* 10(3):1349–1362
9. Wang S, Fan Y, Zhang C, Xu HX, Hao X, Hu Y (2008) entropy based clustering of data streams with mixed numeric and categorical values. *International Journal of Contemporary Hospitality Management* 24(3):430–450
10. Rodrigues PP, Gama J, Pedroso JP (2007) Hierarchical clustering of time-series data streams. *IEEE Trans Knowl Data Eng* 14(2):615–627
11. Tian Z, Ramakrishnan R, Miron L (1996) BIRCH: An efficient data clustering method for very large databases. In: Presented at ACM SIGMOD Conference on Management of Data, pp 103–114
12. Callaghan LO, Mishra N, Meyerson A, Guha S, Motwani R (2002) Streaming-data algorithms for high-quality clustering. In: Proceedings of IEEE international conference on data engineering
13. Kiselev I, Alhajj R (2008) An adaptive multi-agent system for continuous learning of streaming data. *IAT* 148–153
14. Cormode G, Muthukrishnan S, Zhuang W (2007) Conquering the divide: Continuous clustering of distributed data streams. In: Proceedings of international conference on data engineering
15. Jiang W, Brice P (2009) Data stream clustering and modeling using context-trees
16. Zhou A, Cao F, Yan Y, Sha C, He X (2007) Distributed data stream clustering: a fast EM-based approach. *Information Science*, 176(14):1952–1985
17. Khalilian M, Mustapha N, Sulaiman MN, Boroujeni FZ (2009) K-means divide and conquer clustering. In: Proceedings of international conference on computer and automation engineering. ICCAE '09, Thailand, pp 306–309
18. Kononenko I, Kukar M (2007) *machin learning and data mining*. Horwood Publishing, Chichester

Chapter 98

Teaching Computer Ethics Via Current News Articles

Reva Freedman

Abstract Many engineering and Computer Science students need to take a course in computer ethics, and many others could benefit from such a course as an aspect of good citizenship. However, many students cannot appreciate a course organized along the lines of traditional categories of ethics. Additionally, undergraduates generally prefer articles that are relevant to their own lives. This paper provides a discussion of new and current articles in areas of computer ethics that we believe are relevant to students. We have classified the articles by features rather than by a preexisting classification. We expect that the use of current news articles will increase students' understanding of this material and enhance their interest in ethical dilemmas inherent in modern computer systems.

98.1 Introduction

Many engineering and Computer Science students are required to take a course in computer ethics, and many non-majors would also benefit from a general education course in computer ethics as an aspect of good citizenship. However, due to a lack of background in philosophy, many students have difficulties with the common course organization based on philosophers' traditional categories of ethics. Additionally, many students prefer readings that are relevant to their own lives.

This paper provides a discussion of new and current news articles in areas of computer ethics relevant to undergraduate students. We have classified them by common features rather than by a preexisting classification. In cases where the same algorithm can also be used for socially useful purposes, we discuss both

R. Freedman (✉)

Department of Computer Science, Northern Illinois University, DeKalb, IL 60115, USA
e-mail: egamess@gmail.com

aspects. The topics were chosen based on the availability of current, accessible news articles that were largely self-contained. Topics include data mining and reidentification, user interface issues, tracking people via computer, and taking advantage of new facilities of computer systems in novel ways. We expect that the use of current news articles and a classification scheme using common features will increase students' understanding of ethical issues in Computer Science and enhance their interest in the course.

98.2 Data Mining

98.2.1 *Credit Denial Via Data Mining*

Data mining, also called machine learning or pattern recognition, involves using algorithms to find relationships in large databases [1]. Reference [2] discusses the use of data mining by American Express in 2009. American Express had been looking for relationships between consumer behavior and that of other customers who had fallen behind in repayment. In addition to the stores where a customer shopped, their database included data such as local house prices, the type of mortgage lender a consumer was using, and whether its small-business card customers belonged to an industry that was facing hard times. In some cases, American Express used the derived relationships to cut the credit line of customers who had excellent repayment records. Customers received a letter stating that "other customers who have used their card at establishments where you recently shopped have a poor repayment history with American Express." Many consumers objected, including one who provided online [3] a list of the stores and restaurants that he had patronized.

After receiving further unfavorable publicity, American Express decided to stop using this type of spending pattern as a criterion for making such decisions [2, 4]. In fact, this idea created enough bad publicity that Congress required that the executive branch provide a report detailing the degree to which banks assess customers' creditworthiness based on where they shop [5].

Instead, American Express decided to state that they use hundreds of data points as input to credit decisions, and that total debt as a percent of income is the primary determiner of their credit decisions. Using those factors, they are well within the law, which requires only that (a) if credit is denied, the issuer must give you the most important three data points used, and (b) protected categories, such as race and sex, cannot be used. Still, American Express has discussed with investors how much more data they are using than in previous years.

With regard to industries, a dentist will probably do better in the American Express scoring system than a person who owns a construction company. American Express has publicly stated that people with more than one house and a mortgage on each used to be a good bet, but no longer. If you are using a subprime

lender or one that has gone bankrupt, your credit line may be affected even though you have no control over which company your bank sends your mortgage to [6].

Students find this material interesting because getting and using credit cards is important to them. Many of them feel that judging a person by the behavior of people with similar characteristics is intrinsically unfair, and are surprised to learn that it is legal. Others believe that companies should be free to do whatever they want, and they are surprised to learn that categorizing customers by protected categories such as race and sex is illegal.

In addition to traditional credit data, companies can obtain data not specifically related to credit through other types of data collection agencies. These companies collect and sell data on marital status, recent births, education history, type of car, cable service plan and magazine subscriptions. In addition to obtaining data from these companies, companies can also outsource part or all of their data mining operation to companies such as Zoot Enterprises.

Students are usually surprised to learn that some credit card issuers target people who have had earlier financial problems [7]. Although companies previously used data mining to decide which consumers to target, they are now also using similar algorithms to decide exactly which wording to use in each solicitation letter. The goal is to personalize each letter as much as possible without turning people off, by, for example, including a picture of their home in a letter about refinancing a mortgage.

98.2.2 Medical Insurance Denial Via Data Mining

Drugstores, prescription benefit managers and data mining companies such as IMS Health and Verispan sell data about individual users of prescription drugs to drug manufacturers [8]. Although these companies claim that identifiers are always removed before such sales are made, Sweeney [9] and others have shown that simple deidentification is not sufficient to maintain the privacy of individuals.

According to the Wall Street Journal [10], health insurers are buying extensive data from data gathering companies, including data on online shopping, catalog purchases, magazine subscriptions, recreational activities and use of social networking sites, and using it for data mining. The data gathering companies mine public records for hunting permits and boat registrations. They attempt to gather data on lifestyles and health conditions directly from individuals by running surveys whose true purpose is disguised. From social media, they collect information on individuals' favorite social networks, how much they use it, and the types of fan pages they participate in. For example, if you affiliate with a cancer research group, they might assume you have a family history of cancer even though they have no direct evidence.

They buy data from online publishers about whether a user reads financial or sports articles, similar to the data online marketers use to target online ads. Using currently available data, insurers can learn about your commuting time, the

amount of time you spend watching TV, which sports (if any) you engage in, and the type of food you eat. The insurers assume that the amount of exercise a person gets and the amount of fast food that person eats are related to the development of diseases such as diabetes or heart disease.

In addition, they also collect data about individuals' financial status from credit reports, such as "foreclosure/bankruptcy indicators" and credit status. About twenty years ago, researchers for companies that sell car and home insurance discovered correlations between people's credit histories and claims. People with better credit are less likely to file claims. Now medical insurers are doing the same thing.

98.2.3 Socially Useful Aspects of Data Mining

Of course data mining can be used for socially useful purposes, such as in medicine and other scientific research. In addition, data mining can also be used to mitigate the harm caused by the ease of posting false information online. Reference. [11] shows how the latest techniques in text mining can be used to identify fake reviews.

98.3 Reidentification

98.3.1 Lack of Privacy Caused by Reidentification

When databases are publicly released, they are often *deidentified* with the goal of maintaining privacy for the people described therein. A common form of deidentification involves assigning an identification code to each individual so that conclusions can still be drawn but individuals cannot be identified. However, it is often surprisingly easy to *reidentify* individuals by matching up information about people in the database with equivalent information from outside the database. For example, Sweeney [9] showed that one could deduce private medical information about the governor of Connecticut stored in a supposedly deidentified database based on only a few facts about him available from the voter roll. Sweeney later developed the theory of *k*-anonymity [12]. A database provides *k*-anonymity protection if the information for each person contained in the database cannot be distinguished from at least $k-1$ other people in the database.

For a contest among computer science researchers, Netflix gave out a dataset containing 500,000 anonymous ratings of movies in its catalog. Entrants were asked to devise a recommender system that would perform at least 10 % better than Netflix's existing system. However, it turned out that if one knew a few additional facts about an anonymous member of the database, one could break the

anonymity. In [13], two researchers from the University of Texas did exactly that, obtaining the additional data from IMDB, the Internet Movie Data Base. Netflix forced to cancel a second contest to avoid a potential lawsuit and problems with the FTC [14].

Reference [15] explains how the first Netflix contest worked and gives some background about the second. It is an upbeat article with no hint that the second contest would be canceled for reasons of privacy.

In a similar case, AOL had released a dataset containing the searches conducted by 358,000 unnamed people [16, 17]. The data included 20,000,000 search records collected over three months. Although the users were not named, each was given a unique identifier. The files included a timestamp for each record and the URL chosen by the user after searching. Not only could any student of the dataset track the searches of an individual user over time, but many users revealed personal data about themselves through their searches, including their location and medical problems. In this case, breaking the anonymity of these users could be done with ordinary journalistic techniques and did not require data mining.

A related case involves a program that claimed to identify whether a person on Facebook is gay [18]. Being “outed” as gay would probably cause some persons harm but not others. Reference [19] points out the poor quality of the study both in terms of underlying theory and the lack of a rigorous evaluation. Still, being called gay could harm a person regardless of whether the program was accurate or not.

98.3.2 Socially Useful Aspects of Reidentification

Just as with data mining, reidentification can also be used for socially useful purposes. The Wall Street Journal has done an extensive study on fraud identified through Medicare billings. Although doctors’ names are encrypted in the database, outliers can be reidentified by use of factors such as location, type of practice, and data from court files. However, journalists are prohibited from publishing the name of a doctor unless it is obtained from another source [20]. A followup article [21] gives references to all articles in the series.

Similar rules apply to the National Practitioner Data Bank, a federal database of database of doctor malpractice and disciplinary cases [22]. Recently the rules were changed so that journalists must agree not to use other publicly available data to attempt to identify doctors [23]. One expects regulations to evolve in this area, since this rule may be a restriction on freedom of speech. In addition, once the data is available, it will be difficult to preclude its publication.

In addition to the use of text mining to identify fake reviews, several journalists have succeeded in tracking down the authors of fake reviews using traditional research techniques. Reference [24] summarizes the problem of fake reviews on sites such as Amazon, Yelp and TripAdvisor. In [25] the author tracks down a fake reviewer on Yelp. Reference [26] tracks down a case on Amazon where multiple fake reviews had been posted by a company for its own benefit.

98.4 User Interface Issues

98.4.1 *No Human in the Loop*

The term “no human in the loop” has been used to refer to online systems where it is difficult or impossible to contact a representative of the owner directly to correct an error.

In [27], James Fallows explains the steps he had to go through to recover his wife’s Gmail account after the password was stolen by an unknown person in another country. To some extent, he was only able to do this because, as a well-known author, he was able to speak to individuals at Google, while an ordinary person would have been able to contact them only by email. In [28], Fallows offers his opinion on the two most important steps people must take to protect their Gmail accounts. Reference [29] provides a summary of the problem and shows that both Yahoo Mail and Hotmail suffer from similar problems.

Another Google product, Google Maps, can cause problems for business owners due to the lack of a human in the loop [30]. Whether a store has been reported closed to Google through error on the part of a user or through malevolence, it can cause a serious loss of business for the owner.

In [31], Randall Stross, another well-known author, confronts a bank that has overcharged his wife via an automatic payment. Like Fallows, he also contemplates whether his status as a reporter has helped him resolve the situation.

If you are unlucky enough to have a name similar to someone on the Transportation Security Authority’s no-fly list, it is possible that you will be stopped every time you try to board an airplane [32]. The U.S. government is trying to improve the situation by allowing affected travelers to apply for a “redress number,” a secret code that they can enter on their reservations to indicate that they are not the banned person. However, at the point when you are turned back, the individual traveler has no recourse, and the system was only developed because so many people, including some prominent citizens, were unfairly rejected.

98.4.2 *Unsafe User Interfaces*

In 2010, the New York Times ran a series of articles [33–36] on people who had been injured or killed by radiation overdoses in hospitals. In general, these problems involved problems with both the human and computerized aspects of complex systems. Faults were found in hardware and software design and in hospital practices. Problems included poorly designed hardware without safety interlocks, poorly designed interfaces that did not prevent technicians from making serious or fatal errors, and flaws in the training and oversight of technicians.

What is especially depressing about this type of injury is that this problem is not new: one of the best known cases, overdoses due to the Therac-25 machine, happened 25 years ago [37].

98.4.3 Use of Unconscious Processes in Ad Design

Just as major food corporations contribute to obesity by creating artificial foods with combinations of fat and sugar that our bodies crave, the Disney corporation is studying unconscious mental processes to achieve greater penetration of ads. They have fifteen scientists who study eye tracking, heart rate, skin temperature and facial expressions (via the study of facial muscle tensions) in order to produce web ads that users will pay more attention to [38].

To date they have learned that “flyout” ads that appear next to the a media player work as well as transparent ads that appear over the content, but disturb users less. They have also learned that although you may find the TV news ticker intrusive, keeping it running during commercial breaks does not take away from the commercials but helps retain viewer attention. Of all the time spent watching the ad, only about 13 % of it is spent watching the news ticker. The Disney team has also learned that consumers learn to ignore transparent “watermark” logos that overlay part of a web page.

98.5 Tracking People Via Computer

98.5.1 Tracking Workers Via Computer

Using computers, employers can track employees at a level that was not previously possible. For example, Disney [39] has installed large overhead monitors in its laundry rooms. The system, called the “electronic whip” by some employees, shows the piecework speed of each employee.

On the other hand, customers might prefer to see service speeds measured. Everyone who has ever shopped at a grocery store or stood in line at an airport knows that some employees are faster than others. In this type of situation, the fact that another human being is involved in the transaction puts a natural brake on possible speedup.

98.5.2 Tracking Drivers Via Computer

GPS systems and automated toll systems such as EZPass have made it possible to track drivers, whether as employees, customers or family members [40]. Trucking companies

have used this information along with log data collected electronically in the truck to determine whether drivers have been speeding or skipping legally required rest periods. Package delivery companies such as UPS and FedEx also utilize electronic tracking of drivers, generally without controversy.

On the other hand, the use of similar information by rental car companies has occasioned more criticism. Although there have been isolated instances of rental car companies using GPS data to determine whether drivers have violated the rental contract by speeding or taking a vehicle out of state [41, 42], there has been enough negative feedback from courts and the public that the practice has not caught on.

98.6 Taking Advantage of Aspects of the Computer

98.6.1 Taking Advantage of Computer Speed

The New York Stock Exchange used to be a place where people traded stock using brokers as intermediaries. Now, most of the profit comes from computerized stock trading operations that can make trades a millisecond before any human can respond [43].

98.6.2 Content Farming: Taking Advantage of Google

Content farming is the practice of creating spam web sites whose only purpose is to show up high in the search rankings and carry ads. Content farming takes advantage of details of Google's algorithms along with the fact that there is no human in the loop. Google has been slow to deal with content farms ranking higher than links with higher quality information.

Reference [44] explains how content farming works and how it can be used to make money. The cartoon at [45], which illustrates several different kinds of content farming, will make more sense to students once they know what content farming is.

98.6.3 Taking Advantage of Online Availability

The online availability of court records, such as real estate records and divorce settlements, has made it much easier for people to access these records. In many cases, the records were already available through a trip to the courthouse followed by manual copying, but online access has made it easier for people to obtain them

rapidly and in quantity. In this case, the records were not actually private although they were in practice. Similarly, before the advent of the web, people did not have to worry any public act or even semi-public act (e.g., at a party), no matter how minor, could be archived online.

98.7 Related Work

Bynum, in his survey of types of Computer Science ethics classes [46], mentions the possibility of a course based on news articles, calling it a “para” computer ethics class rather than a “theoretical” one (emphasis his).

Chapter 2 of [47] describes many cases involving privacy. A book about the “digital explosion,” it explains for the general public various aspects of current networked computer systems, and thus includes many topics other than ethics.

Of course many textbooks on computer ethics have been published from a variety of points of view. Although there is no room to compare them here, a thirty-year retrospective has been published by Tavani [48].

There is very little overlap between the cases covered in this paper and those covered by such standard textbooks such as [49, 50]. In addition to containing more recent cases, we do not include topics such as cryptography and intellectual property that are not usually the concern of undergraduate students.

Reference [51] introduces students to social impact analysis, an approach to studying the complex interactions between people and computers in socio-technical systems. This approach is highly compatible with the articles presented in this paper.

98.8 Conclusion

In summary, this paper has discussed a variety of contemporary news articles showing ethical issues in Computer Science. The articles have been classified by similarity rather than by abstract principles. The number and complexity of ethical issues that citizens need to deal with is growing rapidly as computer systems become more complex and pervasive in our lives. We believe that the use of current news articles in an ethics course will increase students’ interest in and understanding of ethical issues in Computer Science.

References

1. Witten I, Frank E, Hall M (2011) Data mining: practical machine learning tools and techniques, 3rd edn. Morgan Kaufmann, Waltham
2. Lieber R (2009) American Express kept a (very) watchful eye on charges. New York Times, Jan. 30, 2009, p. B1. <http://www.nytimes.com/2009/01/31/your-money/credit-and-debit-cards/31money.html?pagewanted=all>
3. Johnson K (2009) Beware: these stores could harm your credit! (Part II). Jan. 30, 2009. <http://www.newcreditrules.com/newcreditrulescom/2009/01/beware-these-stores-could-harm-your-credit-part-ii-.html>
4. Johnson K (2009) American Express says it has changed its discriminatory policy, but don't be fooled. Jan. 30, 2009. <http://www.newcreditrules.com/newcreditrulescom/2009/01/american-express-to-change-discriminatory-policy-sort-of.html>
5. Johnson K (2009) Credit Card Act contains amendment inspired by this campaign. May 29, 2009. <http://www.newcreditrules.com/newcreditrulescom/2009/05/despite-the-fact-that-the-credit-card-act-does-not-outlaw-credit-card-redlining-i-am-proud-of-this-campaigns-unyielding.html>
6. Johnson K (2009) Woman denied credit due to blacklisted mortgage company: Bank of America. Jan. 30, 2009. <http://www.newcreditrules.com/newcreditrulescom/2009/03/woman-denied-credit-due-to-blacklisted-mortgage-company.html>
7. Stone B (2008) Banks mine data, and woo troubled borrowers. New York Times, Oct. 22, 2008, p. B1. www.nytimes.com/2008/10/22/business/22target.html?pagewanted=all
8. Freudenheim M (2009) And you thought a prescription was private. New York Times, Aug. 8, 2009, p. BU1. http://www.nytimes.com/2009/08/09/business/09privacy.html?_r=1&sq=prescription%20private&st=cse&scp=1&pagewanted=all
9. Sweeney L (1997) Weaving technology and policy together to maintain confidentiality. J Law Med Ethics 25(2–3):98–110
10. Scism L, Marentmont M (2010) Insurers test data profiles to identify risky clients. Wall Street Journal, Nov. 19, 2010. <http://online.wsj.com/article/SB10001424052748704648604575620750998072986.html>
11. Ott M, Choi Y, Cardie C, Hancock J (2011) Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 2011, pp. 309–319. <http://www.aclweb.org/anthology/P11-1032>
12. Sweeney L (2002) *k*-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557–570, 2002, <http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity2.pdf>
13. Narayanan A, Shmatikov V (2008) Robust de-anonymization of large datasets (How to break anonymity of the Netflix prize dataset) 2008. http://arxiv.org/PS_cache/cs/pdf/0610/0610105v2.pdf
14. Lohr S (2010) Netflix cancels contest after concerns are raised about privacy. New York Times, March 12, 2010, p. B3. <http://www.nytimes.com/2010/03/13/technology/13netflix.html>
15. Lohr S (2009) Netflix awards \$1 million prize and starts a new contest. Sept. 21, 2009. <http://bits.blogs.nytimes.com/2009/09/21/netflix-awards-1-million-prize-and-starts-a-new-contest/>
16. Barbaro M, Zeller T (2006) A face is exposed for AOL searcher no. 4417749. New York Times, Aug. 9, 2006. <http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all>
17. Hansell S (2006) AOL removes search data on vast group of web users. New York Times, Aug. 8, 2006. <http://www.nytimes.com/2006/08/08/business/media/08aol.html?fta=y&pagewanted=all>

18. Jernigan C, Mistree B (2009) Gaydar: facebook friendships expose sexual orientation. *First Monday* 14(10) (Oct. 5, 2009). <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2611/2302>
19. Slattery B (2009) 'Gaydar': Does Facebook know if you're gay? Sept. 21, 2009. http://www.pcworld.com/article/172342/gaydar_does_facebook_know_if_youre_gay.html
20. Schoofs M, Tamman M (2010) Confidentiality Cloaks Medicare Abuse. *Wall Street Journal*, Dec. 22, 2010. <http://online.wsj.com/article/SB10001424052748704457604576011382824069032.html>
21. Carreyrou J (2011) Access to widen on medicare data. *Wall Street Journal*, Dec. 8, 2011. <http://online.wsj.com/article/SB10001424052970204319004577084883951644966.html>
22. Wilson D (2011) Senator protests agency decision to remove doctor data online. <http://prescriptions.blogs.nytimes.com/2011/10/07/senator-protests-agency-decision-to-remove-doctor-data-online/>. Accessed 7 Oct 2011
23. Wilson D (2011) Agency restores doctor discipline files—with a catch. <http://prescriptions.blogs.nytimes.com/2011/11/09/agency-restores-doctor-discipline-files-with-a-catch/>. Accessed 9 Nov 2011
24. Streitfeld D (2011) In a race to out-rave, 5-star web reviews go for \$5. *New York Times*, Aug. 19, 2011, p. A1. <http://www.nytimes.com/2011/08/20/technology/finding-fake-reviews-online.html>
25. Segal D (2011) A rave, a pan, or just a fake? *New York Times*, May 21, 2011, p. BU7. <http://www.nytimes.com/2011/05/22/your-money/22haggler.html>
26. Pilon M (2009) A fake Amazon reviewer confesses. July 9, 2009. <http://blogs.wsj.com/wallet/2009/07/09/delonghis-strange-brew-tracking-down-fake-amazon-raves/>
27. Fallows J (2011) Hacked! *The Atlantic*, Nov. 2011. <http://www.theatlantic.com/magazine/archive/2011/11/hacked/8673/>
28. Fallows J (2011) Quick points on Gmail security, Oct. 12, 2011. <http://www.theatlantic.com/technology/archive/2011/10/quick-points-on-gmail-security/246562/>
29. Yufeczi Z (2011) Small Dictators, Big Bots. Sept. 22, 2011. http://www.slate.com/articles/technology/future_tense/2011/09/small_dictators_big_bots.single.html
30. Segal D (2011) Closed, says Google, but shops' signs say open. *New York Times*, Sept. 5, 2011, p. A1. <http://www.nytimes.com/2011/09/06/technology/closed-in-error-on-google-places-merchants-seek-fixes.html>
31. Stross R (2009) Our payments were automatic. Stopping them wasn't. *New York Times*, July 26, 2009, p. BU3. <http://www.nytimes.com/2009/07/26/business/26digi.html>
32. Singel R (2008) Feds set to take over airline watch list checking, again. Sept. 9, 2008. <http://www.wired.com/threatlevel/2008/09/gov-set-to-take/>
33. Bogdanich W (2010) Radiation offers new cures, and ways to do harm. *New York Times*, Jan. 23, 2010, p. A1. <http://www.nytimes.com/2010/01/24/health/24radiation.html?pagewanted=all>
34. Bogdanich W (2010) As technology surges, radiation safeguards lag. *New York Times*, Jan. 26, 2010, p. A1. <http://www.nytimes.com/2010/01/27/us/27radiation.html?pagewanted=all>
35. Bogdanich W, Rebelo K (2010) A pinpoint beam strays invisibly, harming instead of healing. *New York Times*, Dec. 28, 2010, p. A1. <http://www.nytimes.com/2010/12/29/health/29radiation.html?pagewanted=all>
36. Bogdanich W (2010) Case studies: when medical radiation goes awry. *New York Times*, Jan. 27, 2010. <http://www.nytimes.com/2010/01/27/us/27RADIATIONSIDEBAR.html?pagewanted=all>
37. Leveson N, Turner C (1993) An investigation of the therac-25 accidents. *IEEE Comput* 26(7):18–41
38. Barnes B (2009) Lab watches web surfers to see which ads work. *New York Times*, July 26, 2009, p. B1. <http://www.nytimes.com/2009/07/27/technology/27disney.html?pagewanted=all>
39. Lopez S (2011) Disneyland workers answer to 'electronic whip'. *Los Angeles Times*, Oct. 19, 2011. <http://www.latimes.com/health/la-me-1019-lopez-disney-20111018,0,4593135.column>

40. MobileCast: Collecting Delivery Information in Real-Time. <http://www.roadnet.com/pub/products/MobileCast/>
41. Ramasastry A (2005) Tracking every move you make: can car rental companies use technology to monitor our driving. Aug. 23, 2005. <http://writ.news.findlaw.com/ramasastry/20050823.html>
42. Company Uses GPS to Detect Out-of-State Driving, Adds on Big Fee. Sept.9-18, 2002. <http://www.flyertalk.com/forum/budget/176910-company-uses-gps-detect-out-state-driving-adds-big-fee.html>
43. Duhigg C (2009) Stock traders find speed pays, in milliseconds. New York Times, July. 23, 2009, p. A1. <http://www.nytimes.com/2009/07/24/business/24trading.html>
44. Sullivan D (2010) The Google Sewage Factory, In Action: The Chocomize Story. July 27, 2010. <http://searchengineland.com/google-sewage-factory-the-chocomize-story-47403>
45. McFadden B (2011) The strip: content farm. July 16, 2011. www.nytimes.com/interactive/2011/07/17/opinion/sunday/20110717_McFadden_Cartoon.html
46. Bynum T (1991) Computer ethics in the computer science curriculum. In: National Conference on Computing and Values, 1991. http://www.southernct.edu/organizations/rccs/oldsite/resources/teaching/teaching_mono/bynum/bynum_human_values.html
47. Abelson H, Ledeen K, Lewis H (2008) Blown to bits: your life, liberty, and happiness after the digital explosion. Addison-Wesley, Boston
48. Tavani H (1999) Computer ethics textbooks: a thirty-year retrospective. ACM SIGCAS Comput Soc 29(3):26–31
49. Baase S (2003) A gift of fire: social, legal and ethical issues for computers and the internet, 2nd edn. Pearson Education, Upper Saddle River
50. Spinello R, Tavani H (2001) Readings in CyberEthics. Jones and Bartlett, Sudbury
51. Teaching Tools for the Computing Curriculum. <http://computingcases.org/>

Chapter 99

Designing and Integrating a New Model of Semi-Online Vehicle's Fines Control System

Anas Al-okaily, Qassim Bani Hani, Laiali Almazaydeh,
Omar Abuzaghleh and Zenon Chaczko

Abstract In this paper we suggest to develop a vehicle's speed and fines control system to manage and control different aspects of fleet and cruise management system. The system developed to be sponsored by the government, which is represented by Department of Motor Vehicle (DMV) and should be operated by them. The purposes of the proposed project include speed, passengers' safety and vehicle readiness and related fines associated with driving practice such as wearing seat belt and speed limits. The system can be implemented by developing a software system inside a chip supported by recent related technologies such as GPS, GPRS and cameras, then installing the chip into the vehicle. The final outcome will be levying penalties respective to the driver's mistakes and offences; in addition new era of communication between DMV and driver, vehicle and driver, driver and DMV will be followed.

A. Al-okaily (✉)

Department of Computer Science, University of Connecticut, Storrs CT, USA
e-mail: aaa10013@engr.uconn.edu

Q. B. Hani · L. Almazaydeh · O. Abuzaghleh (✉)

Department of Computer Science, University of Bridgeport, 26 Park Avenue,
Bridgeport CT 06604, USA
e-mail: oabuzagh@bridgeport.edu

Q. B. Hani

e-mail: qbanihan@bridgeport.edu

L. Almazaydeh

e-mail: lalmazay@bridgeport.edu

Z. Chaczko

University of Technology, Sydney, Australia
e-mail: Zenon.Chaczko@uts.edu.au

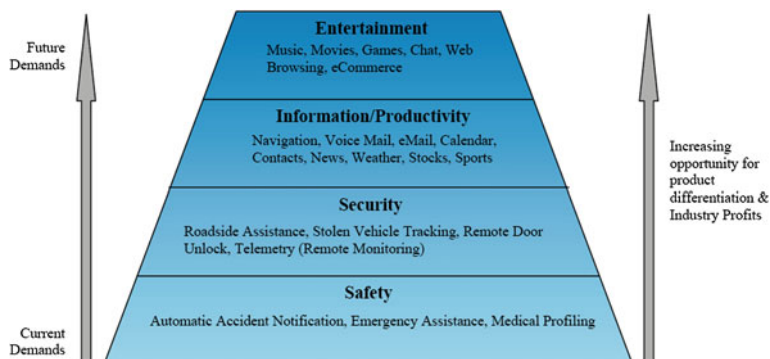


Fig. 99.1 The evolution of consumer telematics interest on USA

99.1 Introduction

Ever since life existed on earth, people used to travel long distances, therefore they increasingly developed methods and tools to save time and efforts in travelling. Animals were the first method used by people as mobile tool. Later, carriages (Rickshaws) were pulled by animals—such as horses—were developed, and were used for very long time.

A century ago; the first car was invented by Richard Dudgeon in 1866, had a steam engine, and the steam engine was costly. So owning and using such car by the public people was hard. In 1920s Henry Ford invented the first affordable car. For the first time Ford's car allowed and motivated different public people levels to have a real opportunity to have a car; since then, the world witnesses tremendous improvement and development in this new way of life.

As a consequence, for the massive usage and development in car domain, governments started and still adopted more and more regulations and management laws in different domains of cars practices. Besides that, private sectors and manufactures have increased the importance of the developments and enhancements on functional aspects and services provided in cars, in order to ease and improve the usage of cars.

Nowadays there are a serious application and implementation on the roads; road of Telematics (Telecommunication Informatics) applications, as shown in Fig. 99.1 [1]. Several companies such as General motors, Ford and even Microsoft have entered this road by developing several applications and implementing several services.

Vehicles are now being transformed by a wireless revolution that will substantially enlarge the Telematics market over the next decade. But carmakers are unlikely to win much of the revenue from this expanding market unless they aggressively shape the environment so that Telematics applications can succeed commercially. [1].

So, naturally different domain and fields of science will participate and incorporate in the introduction and development of Telematics domain. As IT industries

have been participating and incorporating on the developments in most aspects of science industries. Main IT companies which may involve in such project include: navigations companies, cell phone companies, network communications companies, and most of companies related to the IT projects development and software implementations. Therefore IT will have a part or even a main part in the Telematic's applications and services.

99.2 Related Work

Reference [2] published about car black box, the system is similar to the airplane black box, and its big picture is shown in Fig. 99.2. The black box is mainly used to record information related to accidents. The box records a driving data, visual data, collision and position data; before and after the accident. So these output information can be easily used for the analysing and investigation related to the accidents and help to settle many disputes and ambiguity related to the accidents. The car black box is equipped with a wireless communication which can send accident location information to emergency centre.

Car black box's function involves:

- Data collection
- Driving data, such as speed, brakes and seatbelts status is recorded
- Visual data
- Collision data
- Position data
- Accident analysis data and wireless communication activities data are recorded.

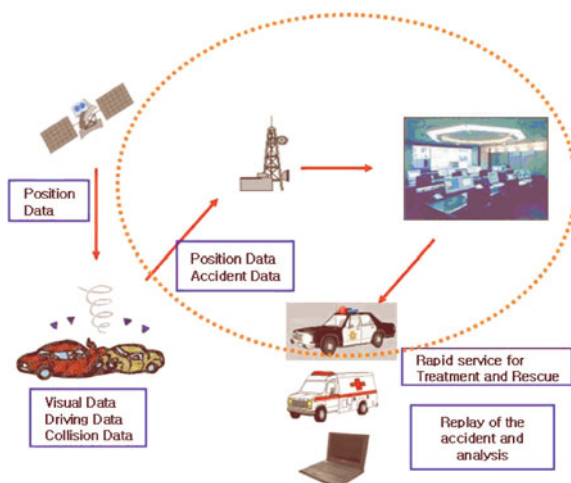
Reference [3] discussed a new system, intelligent adaptation system (ISA), i.e. A combination of technological systems that support drivers in their choice of Travel speeds. While there are varying terms used worldwide to describe the forms of ISA systems, Australian road agencies actively working in ISA have recently agreed to adopt the following common language:

- Advisory ISA—systems that remind drivers of the prevailing speed limit and exert no control over the vehicle.
- Supportive ISA—systems that provide some degree of vehicle-initiated limiting of speed, but allows the driver to override the system.
- Limiting ISA—systems that include vehicle-initiated speed limiting that cannot be overridden (usually accompanied by an emergency failure function).

ISA was trailed in several countries around the world before 2002. Sweden in 1997, Netherlands 1999, United Kingdom 1998, Denmark 2001 and Finland 2002. The result of these trials shows several results (this till the end of 2002):

- Most of the result shows that drivers have positive attitude towards ISA.

Fig. 99.2 Car's black box top model



- Many of the drivers also reported that the lower speeds when using ISA did not lead to longer travel times.
- The compulsory system seemed to be the most effective means of speed reduction out of the three systems; it was this system that was rated as least acceptable by the drivers. There are some technical video presented in RTA websites, for more explanations.

According to the parliament of NSW in Australia, This \$1 million intelligent speed adaptation project (trial) will involve 100 cars, up to four global positioning satellites and cutting-edge in-card technology similar to that of satellite navigation devices. This project will take place between 2008 and 2009 in some areas in NSW. In [4] the result indicates that “The NSW ISA Trial has demonstrated that Advisory ISA technology has the potential to deliver considerable road safety benefits, by reducing the level and duration of speeding amongst the majority of participating drivers. Results from the attitudinal research showed that the Advisory ISA technology was generally well received and accepted by those participating in the trial.”

Another system (SatNav), connected via GPS just tells you, whenever the driver exceeds the speed limits, to conduct driver-self actions, that means the system does not take any action; just telling that you have exceeded the speed limit [5]. In contrast, another system connected within the vehicles, and controls the vehicle's speed by automatically applying a sequence of brakes or by switching off the vehicle engine when the vehicle's speed goes over limit [6].

Reference [7] has a patent of GPS based monitoring system; in form three of the patent the author claim some of VFCP's functions, but with some differences. In his system, he claims that a speed penalty can be issued whenever the driver exceeds the speed limit without the driver's involvement, a beeper to alarm the driver of the exceeded speed, display screen to be used, GPS module, and GPRS

module to communicate. The system should have geographic map of the speed zones; Accident alerting messages to be sent to DMV; biometric pin number for authorizations. But the proposed system has several differences, there are several parts of VFCP that make it different than the author's system, the author's system can switch off/on the vehicle in case of exceeding the speed limit; while in the proposed system its more likely to be remote controlling rather than physical interfering with the engine management in the speed related issues; the author's system use smart card to record all full details about the vehicle movement, whereas in the proposed system does not, recording all locations and logistic data about the driver movement, since it has big consideration regarding people privacy; there is no direct connection to be held between driver and DMV or message to be sent and to be showed on the vehicle display screen; there no speaker, moreover no belts or light connection with the system, moreover, there is no usage of cameras and recording operations.

Referring to BMW web site; BMW has developed many technologies regarding vehicle communication and safety aspects. One of several developments is storing the services status in the vehicle key, and then these data can be accessed by the services centres during the next appointment. Another one, similarly to GMC Company, when the airbags activated, an integrated sensors automatically trigger an emergency call and notify the nearest emergency centre of the vehicle's location using the GPS navigator. This has been known as emergency-call E-call which is the call that a vehicle sent in case of emergency situation occurred in the vehicle such as the airbag activation. In addition, a mobile phone preparation with a Bluetooth interface and hand free controls, has been developed which can be activated via the steering wheel and voice control features. For the speed issue, BMW had developed technology, which is Active Cruise control, by which, an automatic reduction of the vehicle speed can be done if a slower vehicle appears in the head street; as well accelerate back when the street is free. Moreover, there is an advance safety electronics system, the system use optical-fibre network to coordinates the system's responses in an emergency situation; such as in collision, the system coordinate the inflation level of the airbags and if necessary deactivate the fuel pump and disengages the battery. In conclusion, BMW has come up with the most technological updates. According to VFCP specification some of these technologies are similar with the mentioned system, but at a secondary level of its aims; since the main aims and objectives of the proposed system is to deal with fines related to DMV.

Referring to Mercedes (http://www3.mercedes-benz.com/international_home/en/), there were not significant technologies mentioned as the ones in BMW. The most technologies used regarding using GPS navigation is the implementation of GPS navigator called COMAND; COMAND functions is almost normal functions and does not related to the speed issues and even to safety issues. On the other hand, Ford vehicles in their website describes some technologies similar to the BMW ones; they got crash sensor which sends GPS signals, in addition they got **anti-theft system** which has been developed to help in preventing the engine from being started unless a code key should be entered. In conclusion, there is no direct

relation with the technologies that Ford and Mercedes have implemented with the main aims of VFCEP.

Reference [8] proposes in details a new approach toward vehicle speed control. The new approach is called speedNet, according to the authors speedNet is “a GSM Network variant that has the ability to actively track mobile network users and their velocity, predict their mobility patterns and control the speed of the vehicles in a speed limit critical area by sending a control command to the Vehicular Velocity Limiting Systems (VVLS) of the vehicle(s) in concern.” SpeedNet use GSM network which considered as second generation mobile system (more secure and digitally encrypted); VVLS (vehicular velocity limiting system) which is the system by which the vehicle speed can be limited and controlled, the function of VVLS system is to set the vehicle’ speed on a specific speed (in SpeedNet VVLS set the vehicle’s speed as the maximum allowable limit). According to the authors, SpeedNet is a customized wireless mobile network where base station and user station are both mobiles; the main idea is, tracking the mobile user by the base station in order to do speed policing and control the traffic.

This simulation described as the following, each base station tracks a set of speed limited zone, the two nearest base station from the user mobile (car) tracks and monitors the location and speed of vehicle when the vehicle is in speed critical zone. In other words, if the vehicle is in a critical speed zone, the base station sends the speed limit to the vehicle VVLS to set the speed limit for the vehicle; then if the vehicle exceeds the limited speed, VVLS force the vehicle to decelerate its speed into the allowable speed. In conclusion, this system is just a proposed system and according to its functionality, the system has large differences with VFCEP.

In [5] they have developed a neurofuzzy approach in the design of road bumps for the control of vehicle speeds. This is a new contribution to the body of knowledge on road bump design. The approach is adopted in order to improve on the use of fuzzy logic, which attempts to capture imprecision and uncertainties. The study is a practical case example used to demonstrate the feasibility of applying the proposed methodology.

99.3 Problem Identification

In order to explain project rationale several problems and issues are needed to be described and smoothly linked and structured. There are three issues that are needed to be figured to support the rationality of applying such system in the real life. Motor vehicle thefts issue; road accident issues including excessive speed and non-belt wearing, economical issues and technical issues. According to US census bureau www.census.gov, the following statistics were released to 2007 CASE STATISTICS.

Firstly, due to the unavailability of full information and statistics which this paper concern about for a recent a year; only 2007 statistics is available in full details. Secondly with initial and a rough calculation, VFCEP is assumed to contribute in reducing at most 45 % of related damages: which include:

Table 99.1 2007 statistic of fatal crashes

| Item | 2004 | 2005 | 2006 | 2007 |
|-----------------------------------|--------|--------|--------|--------|
| Fatal crashes, total | 38,444 | 39,252 | 38,648 | 37,428 |
| One vehicle involved | 21,836 | 22,678 | 22,701 | 22,054 |
| Two or more vehicle involved | 16,608 | 16,574 | 15,947 | 15,194 |
| Persons killed in fatal crashes\1 | 42,836 | 43,510 | 42,708 | 41,059 |
| Occupants | 37,304 | 37,646 | 36,956 | 30,401 |
| Drivers | 26,871 | 27,491 | 27,348 | 26,480 |
| Passengers | 10,355 | 10,069 | 9,507 | 8,977 |
| Others | 78 | 86 | 101 | 98 |
| Non occupants | 5,532 | 5,864 | 5,752 | 5,504 |
| Pedestrians | 4,675 | 4,892 | 4,795 | 4,654 |
| Pedalcyclists | 727 | 786 | 772 | 698 |
| Other/unknown | 130 | 186 | 185 | 152 |

Source US senses bureau www.senses.com

- Crashes fatalities: referring to Table 99.1, the proposed system would reduce speeding practices to up to **45 %**; **saving the life of 13,040 lives = 5,868 persons** in USA, only for 2007 metrics. For Belt statistics: 56 % of people were killed in car accidents, were not using the belt, so $56 \% \times 41,059 = 22,993$ persons were possibly killed due to not wearing seatbelts. The proposed system will save lives for: **$45 \% \times 22,993 = 10,346$ persons.**
- Crashes Injuries: referring to www.census.gov 37 % of people were injured in car accidents, this can come up with the following calculation: $37 \% \times 2,491,000$ (crashes injuries in 2007) = 921,670 persons were possibly injured due to speeding factor. So the proposed system might help in saving **$45 \% \times 921,670 = 414,751$ persons** in USA, only for 2007. Belt statistics: 56 % of people were injured in car accidents, were not using the belt, this can come up with the following calculation: $56 \% \times 2,491,000 = 1,394,960$ persons were injured. Hence the proposed system will save: **$45 \% \times 1,394,960 = 627,732$ persons.**
- Economical impacts: According to NHTSA “Speeding is one of the most prevalent factors contributing to traffic crashes”. The economic cost to society of speeding-related crashes is estimated by NHTSA to be **\$40.4 billion** per year. So for 2007, **$45 \% \times 40.4 = 18.2$ billion**, in USA, per year.

99.4 Purposes

Vehicle Fines Control Project (VFCP) has many goals related for both drivers, and transportation authority, Department of Motor Vehicle (DMV). VFCP aims to develop a system which has microprocessor for operations, memory for recording and storing, navigations module, mobile/cell chip for communication, linked screen, microphone and Speakers, cameras, and the related software and control

devices. These components will be interconnected and physically packed on a convenient box. The box should be easily installed into the vehicle in a very secure and reliable way.

The project adopter and sponsor is DMV. This project mainly provides online information and supports to control and prevent some illegal driving practices such as excessive speed and non-wear belts. So the main purposes are:

DMV authority may gain several benefits and control issues such as:

- To control all vehicles legal and illegal actions over an entire country or a state. This may include the non-wear belts and exceeding speed practices.
- To save the troubles and drawbacks of using—the costly and ineffectively—traditional road penalty tools and methods. Such as speed cameras and direct police surveillances.
- To apply online fines in case of illegal driving practices; such as, exceeding the speed limits and not wearing seat belts.
- To notify the driver of exceeding the speed limit, unplugged belts and late registrations date.
- To assist DMV with locations queries for any suspected vehicle; such as stolen vehicles. This can be done by sending queries message into the system, then the system provide the relevant information needed.
- To allow DMV to communicate by voice or text messages with the vehicle driver in case of doing illegal actions; such as illegal parking, exceeding the speed limit, late registration. That's can be done, by Sending warning messages into the proposed system; then the received message cab be shown on the system screen or sound a voice message via the speakers; hence the driver can read or listen to it.
- To create new ways of faster emergency knowledge, in case of crashes or collisions, the system can provide DMV with related information; by sending emergence events and information to DMV.
- To integrate many commercial services, which are applicable in separate ways in the vehicle, then combines them in one secure and integrated system. Some examples of commercial system are; services developed by some companies—such as GMC company—which sends emergency message whenever the air-bags released, and service which just gives the driver roads speed limits info such as TomTom navigator.

Vehicle drivers may gain several important benefits from VFCEP application such as:

- To let the driver communicate with his own vehicle using a predefined text messages which can be sent by the driver's mobile phone; VFCEP can respond to the communication (instructions such as lock/unlock) after reading and analysed the messages based on built in messages-responder.
- To implement a security subsystem supported with a Pin number or fingerprint authentication access; any invading or hacking of the system will have a direct communication with the police or the vehicle owner, via emergency messages.

- To support the driver with live video of the rear and front of the vehicle on the VFCP screen monitor, using the linked cameras in front and rear, as well as to support other logistic information.
- To mandatory record some logistics information and/or video and audio records, as result this information can help and assist the driver with information which may needs. In addition these recorded data can be reviewed or used on some investigation by DMV. In other words, it can be considered as the black box for the vehicle, as forth for the plane; which can be evidence/proof against illegal actions, therefore DMV can view and review the data and records on the black box, in case of any fines investigation or checking.

99.5 Privacy Issue

Adapting “vehicle fines control project” do not require a full details of the owner/driver’s private information; the private information that might be needed is logistic ones, which is according to the system is mainly the vehicle plate number, vehicle’s registration number, vehicle model, vehicle colour, the driver/owner mobile number.

99.6 Proposed Sloution

Nowadays, according to the current technologies available, tools in the markets; as well as to the facilities and IT applications in the market. In my point of views, all tools, methods and technologies project need to be accomplished are available and ready for implementations.

What the project need in terms of devices, tools and hardware? The following descriptions state almost all of that:

- First of all, small microprocessor chip with its memory, which used as RAM and buffering, this unit is used to control and operate all software and hardware operations.
- Medium size memory, about 8 GB memory disk is quit flexible to store targeted information, this information may include: records data, images data, archiving data and video data if needed (video data captured from the connected cameras). In addition, the software codes and implementation which control and operate the functionality of the VFCP.
- Cell/phone (GPRS) module connected with the system. This unit obviously is used to send and received text messages, as well as to make or receive calls if they were needed.
- GPS navigator device connected with system. It used for locating the vehicle’s position as well as to read the speed of the vehicle. This unit need to provide the

system with location and the speed information; it must be accurate brand and may be a specific built-design for the proposed system.

- Sensors connected with the belt and light and the vehicle's lock management. These sensors used to notify and trigger a specified action needed, for example if the belt is linked or not, airbags releasing and vibration or collision sensors.
- Cameras, linked with the system, three cameras, one on the vehicle front and on the rear, the third inside one. The driver and passengers. These cameras connected with the system for recording purposes, as well as linked with the system's screen to assist the driver in monitoring the front and the rear of the vehicle.
- LCD Screen, that shows logistics information for the driver, in addition it used to show the received message or notes left by the system for any update information or miss messages.
- Speaker and microphone; the speaker is for producing any warning or alerts or directions etc. microphone, for communication via the system cell/phone in case the police need to make a call the driver. (Here instead we can create a built-in voice messages, hence producing the voice messages via the system instead of showing them on the screen, since this will not interrupting the driving).

Now, there are several phases in the project that are required to be clearly defined and implemented:

- First of all, we need to find a module (microchip) that can connect all devices together in order to control these devices and integrate them; ultimately write the software for implementation.
- Secondly, GPRS module need to be implemented, create functions that will achieve parts of the requirement such as write functions for sending and receiving GPRS messages.
- GPS navigators which can be connected with the system, and allow the software to fetch the required information which is mainly position and speed.
- Cameras should be connected with the system and have the ability to capture and save images and video from them, in addition show the images captured on the system's screen.
- The system should have a screen which will show the images received from the camera and shows logistic information.
- The 8 GB memory's role is to record and save needed information and data, as well achieving purposes.
- Speaker and microphone should be plugged into the system in order to be used for sounding or voicing.
- Sensor management: this includes the sensing methodologies from the belt and light, and need to provide the system with the sensing-output data.
- The Bluetooth device should be implemented to use for remote communication (sending and receiving data, from and into the system); these operations mainly have to be done by the DMV officers, that is, whenever the DMV officer need to get a data or information or make an update for the system software.

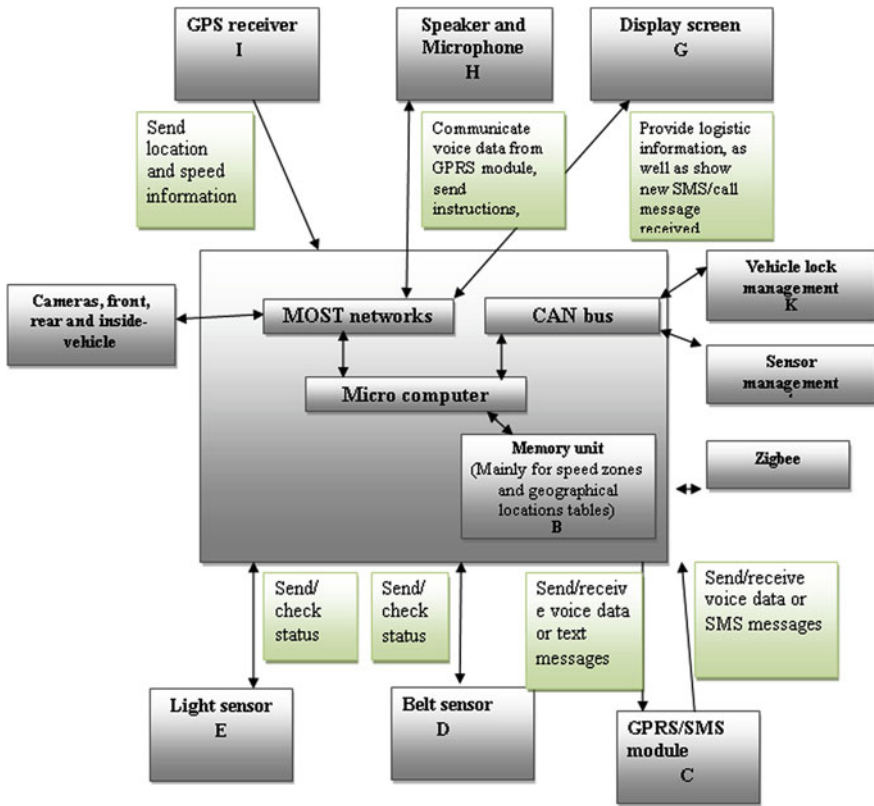


Fig. 99.3 The final conceptual framework design of VFCP system

The following description describes different components of VFCP architectures (Fig. 99.3):

These units should be linked and connected together, and at the end packed on a reliable and secure box, the box size must be quite small to ease the installing operation into the vehicle, of course with outputs plugs. These plugs connect the box with sensors and have other relative connections; such as USB reader and writer. Read memory archive; write updating operation such as Map updating. For more secure reading and writing operation, Bluetooth secure connection can be built in and created and can be accessed by the DMV officer remotely without the need to uncover the box place; so it make it easy and fast way for the officers to perform the updating and reading/writing operation.

The following diagram depicts the big picture of the proposed system's development and deployment (Fig. 99.4).

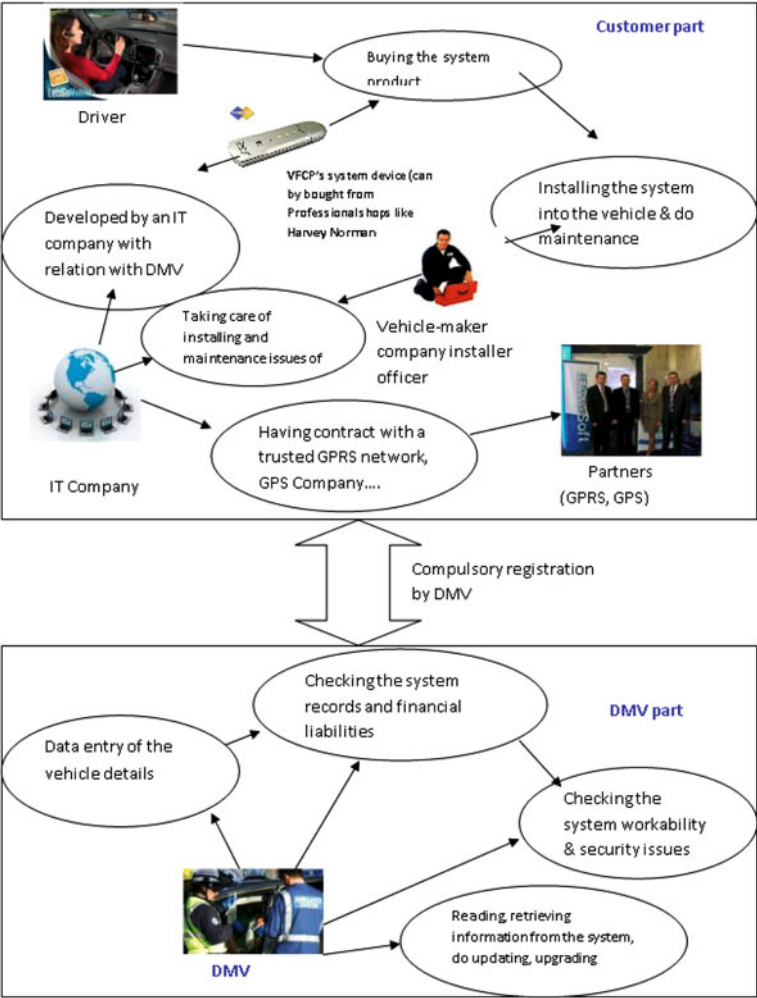


Fig. 99.4 The final conceptual framework design of VFCP system

99.7 Conclusion

A new design and integrated model of semi-online vehicle fines control system, which provide new integrated elements and adopt a new ways of online services to any department of traffic authorizations for any country. One of its significance is to reduce about 45 % of speeds and belts illegal practices, as results reduce the same percentage of deaths and injuries and economical impacts that are caused by such practices. It will help in saving 10,346 persons' lives; per year; 627,732 persons of injuries, per year; and 18.2 billion, in USA, for 2007.

References

1. François B, Andreas C, Philip R (2001) The road ahead for telematics. McKinsey & Company, New York
2. Filipova-Neumann L, Welzel P (2010) Reducing asymmetric information in insurance markets: cars with black boxes *Telemat Inf* 27(4):394–403
3. Australia's GPS-based speed control system may have solution to Blueline ills, (2008)
4. Results of the NSW Intelligent Speed Adaptation Trial (2010). Retrieved from www.rta.nsw.gov.au
5. Oke SA, Johnson AO, Salau TAO, Adeyefa AO (2007) Application of Neurofuzzy in the development of road bump designs. *Pac J Sci Technol* 8(1):73–79
6. Satnav that tells you not to break the speed limit. (2008) Daily mail. Retrieved from <http://www.dailymail.co.uk/sciencetech/article-565885/Satnav-tells-break-speed-limit.html#ixzz142m8Uy18>
7. Petrik S (2006) Australia Patent no. Australia patent office, A. P. Office
8. Pubudu P, Andrey S, Nirupama B, Tony P (2006) Speed control and policing in a cellular mobile network: SpeedNet. *Comput Commun* 29:3633–3646

Author Biographies

Anas Al-okaily PhD candidate in computer science at University of Connecticut, I worked as full-time/part-time lecturer in my country Jordan 2009/2010. My research interest include: Software Engineering, mobile application, sequential and parallel deterministic algorithms. I finished my master degree from University of Technology, Sydney, Australia 2009; my bachelor degree from Jordan University of Science and Technology, Jordan, 2005.

Laiali Almazaydeh is a Ph.D. student of Computer Science and Engineering at the University of Bridgeport. She received a B.S. in Computer Science from Al Hussein Bin Talal University and an M.S. in Computer Information Systems from The Arab Academy for Banking and Financial in 2003 and 2007, respectively. Her current research focuses on the wireless sensor networks and human computer interaction.

Mr. Abuzaghleh is a Ph.D. candidate in computer science and engineering at the University of Bridgeport. He has research interests in the areas of distributed database systems, data mining, and cloud computing. Mr. Abuzaghleh received the B.Sc. degree in computer science and applications from Hashemite University in 2004, the MS Degree in computer science from university of Bridgeport in 2007. Mr. Abuzaghleh is currently pursuing his Ph.D. degree in computer science and engineering at university of Bridgeport.

Dr. Zenon Chaczko is currently the Senior Lecturer of Software Engineering at UTS, and an active member of CRIN. After 25 years of R/D experience in ICT and marine systems industry as well as concurrent 4 years of P/T lecturing at the University of Technology Sydney, he moved to full time academic position at UTS in 2002. Since then he has been Program Head of Information and Communications Technologies. He is an experienced lecturer and researcher, consistently receiving excellent teaching results and reviews from students. He is an expert software and system engineer, researcher and supervisor, having supervised/co-supervised many candidates to completion in the last 8 years. He has completed his PhD in Engineering at UTS. His specialisation is anticipatory (AI) and biomimetic middleware systems for Wireless Sensor Networks. He is an author and co-author of several innovative AI theories and computational models

Chapter 100

State Diagnosis of a Lignite Deposit by Monitoring its Surface Temperature with a Thermovision Camera

Alina Dinca

Abstract The paper presents results of a research project [1] which was extended over two months and had as goal to prevent the self ignition phenomenon in a lignite deposit or stockpile. The idea was to monitor the surface temperature of the stockpile with a thermovision camera and compare the results with the ones provided by monitoring the temperature inside the stockpile with temperature sensors attached to an acquisition system. This paper focuses on finding out whether the state of a lignite stockpile can be diagnosed correctly by monitoring its surface temperature with a thermovision camera in order to prevent the self ignition phenomenon. The author actually tries to answer the question: does the surface temperature of a lignite stockpile monitored with a thermovision camera offer enough and correct temperature information to be able to prevent an upcoming fire?

100.1 Introduction

The research idea for this paper came out of the need to keep the lignite safely deposited over undetermined periods of time, until it is used as fossil fuel for power plants.

Lignite which is extracted from open mining pits is the main resource in matter of fuel for the power plants in the South-West Region of Romania, as well as in other countries. Lignite or coal in general, needs to be deposited for a while in stockpiles after it has been extracted and before being used as fuel. Stockpiles

A. Dinca (✉)
Department of Computer Science, Northern Illinois University,
Geneva Street, no. 3, Targu-Jiu, Romania
e-mail: alina@utgjiu.ro

should be built after a set of rules like their height mustn't exceed 10 m, lignite mustn't be deposited when it rains or when humidity is high, its granulation shouldn't exceed 10–20 cm, etc., all these in order to keep lignite deposited in best conditions [2, 3].

Lignite, coal or any other material deposited in large amounts tends to overheat within the stockpile [4]. Where there is heat, oxygen and some material to be burned, an upcoming fire should be taken into account! When the lignite is deposited following the proper rules, some of them mentioned above, heating or self heating appears very slowly.

As a process, self heating appears due to lignite oxidation [5]: because the lignite is not perfectly settled in the stockpile, air and water can run between the pieces. The oxygen contained in the air can make the lignite produce heat and since the heat cannot be exchanged with the environment, it is enhanced within itself and exchanged with neighbors, creating hot spots and leading to self ignition.

Usually, both chemical structure of lignite and rules mentioned above that are not always or rigorously followed as such, lead at one point in time or another to self heating and to self ignition of lignite, which have major consequences over human health, environment and production:

- After lignite reaches a certain temperature, its caloric power becomes lower; the higher temperature gets, the lower its caloric power becomes, and could decrease with 50–70 % [3], so a certain lignite quantity would produce in these conditions less energy than that produced by the same quantity that doesn't overheat;
- Increasing temperature within the stockpile, makes lignite release toxic gases, such as carbon monoxide (CO) during self heating process and sulfur dioxide (SO₂) during the self ignition process, which are harmful for humans living in the surroundings of the deposits, and also for environment [6];
- Toxic gases released by self heating and self igniting lignite also directly affect the surrounding environment and indirectly affect it by the extra excavations that are needed in order to replace the lost lignite.

So if lignite self heats, its caloric power decreases; if it self ignites, it affects a considerable amount of lignite in its surroundings, making it lost for the energy production. Also, human health and environment suffer.

These are the reasons why the thermal state of a lignite stockpile must be known at any moment in time in order to avoid self heating and most important, self ignition.

ENELEX, FLIR Systems integrator, managed to monitor successfully a stockpile and to prevent self ignition at the Nastup Mines Corporation in Tusimice, Czech Republic [7], which encouraged us to try a similar approach.



Fig. 100.1 special thermometer of 1.5 m

100.2 Approach: Monitoring a Lignite Deposit

100.2.1 Current Monitoring Methods

Current methods to monitor the temperature in lignite stockpiles and diagnose their state are:

- Periodically check suspicious areas with a special thermometer of 1.5 m (Fig. 100.1): each day for deposits with temperature over 35 °C, and from 6–25 days for deposits with temperature under 35 °C;
- Periodically check with metallic bars for depths higher than 1.5 m.

Current methods to keep temperature in desired ranges are:

If temperature indicated by thermometers or metallic bars is over a desired value or if an employee working at the lignite stockpile observes steam getting out of a stockpile's part, which is already too late, because deeper in the stockpile under the steamy area temperature is over 60 °C and may rise up to 70 °C, they call it in and: (1) either that part of the stockpile is sent to production or (2) that part of the stockpile is being moved from one place to another with an excavator (Fig. 100.2), operation which can take up to 1 or 2 days, depending of the area the heating is extended, until production requires it.

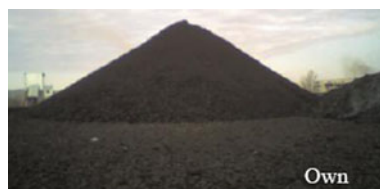
100.2.2 The Ideal Case

In order to define the state of a lignite deposit, let's imagine we have such a deposit, as presented in Fig. 100.3 below, and an imaginary thermometer that

Fig. 100.2 Excavator removing hot spots



Fig. 100.3 A side of a lignite stockpile



would gather temperature data from each point of the stockpile and display it every time when a temperature value changes with a given constant as an average maximum of temperature over an area (user defined dimensions) and the x, y, z coordinates of its location. There can be an unknown number n of areas within the stockpile where temperature can rise independently, areas which can become hot spots with different maximum temperature simultaneously.

The variation in time of heat inside a lignite deposit follows the pattern: it reaches 35°C slowly (in about two months), it continues heating until 45°C , it passes through the hot spot creation state, and in about 15 days, around that hot spot temperature can reach the self ignition value of 65°C .

So:

State 1—when the imaginary thermometer shows areas with a maximum temperature under 35°C , the stockpile is under no danger;

State 2—when the thermometer shows for certain areas an average temperature of $55\text{--}65^{\circ}\text{C}$, there is a hot spot already;

State 3—when the thermometer shows for certain areas an average temperature of over 65°C , the self ignition appears shortly.

In the case presented, it would be ideal that the temperature in a lignite stockpile should only be found in State 1. The moment when State 1 goes towards State 2, authorized personnel should be warned to take proper actions in order that the temperature of the stockpile to reach State 1 again.

Fig. 100.4 Same as in Fig. 100.3, in infrared

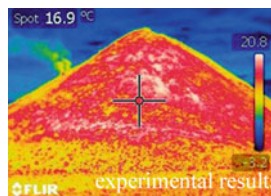


Fig. 100.5 Flir T200 thermovision camera



100.2.3 Our Approach: Thermovision Monitoring

In Fig. 100.3 the picture shows a side of a lignite deposit. Figure 100.4 shows the same side of the lignite stockpile in infrared. The picture in Fig. 100.4 is taken with a Flir Systems thermovision camera, model T200 (Fig. 100.5).

A thermovision camera absorbs the thermal energy or radiation emitted by the surface of the “viewed” or monitored objects and transforms it into color maps or thermographic images in which white means hottest and black means coldest. The colors between white and black represent values between the maximum of hottest and coldest [8].

In the research project [1], I wanted to find out whether the state of a lignite deposit can be determined by monitoring its surface temperature with a thermovision camera.

For that, we used only 5 % of a freshly created 10 m high lignite stockpile as in the scheme in Fig. 100.6. The surface of the chosen 5 % of the stockpile was monitored with the thermovision camera from 8 locations (from 1 to 8 in Fig. 100.6a), so that the eastern and western sides to be monitored from top and bottom, northern only from bottom and three points on top of the stockpile from one position. The 4, 5 and 6 positions were chosen to monitor the place where the other member of the project’s team put some pipes in which they put oil and sinked thermal sensors in order to check the temperature at different depths in the

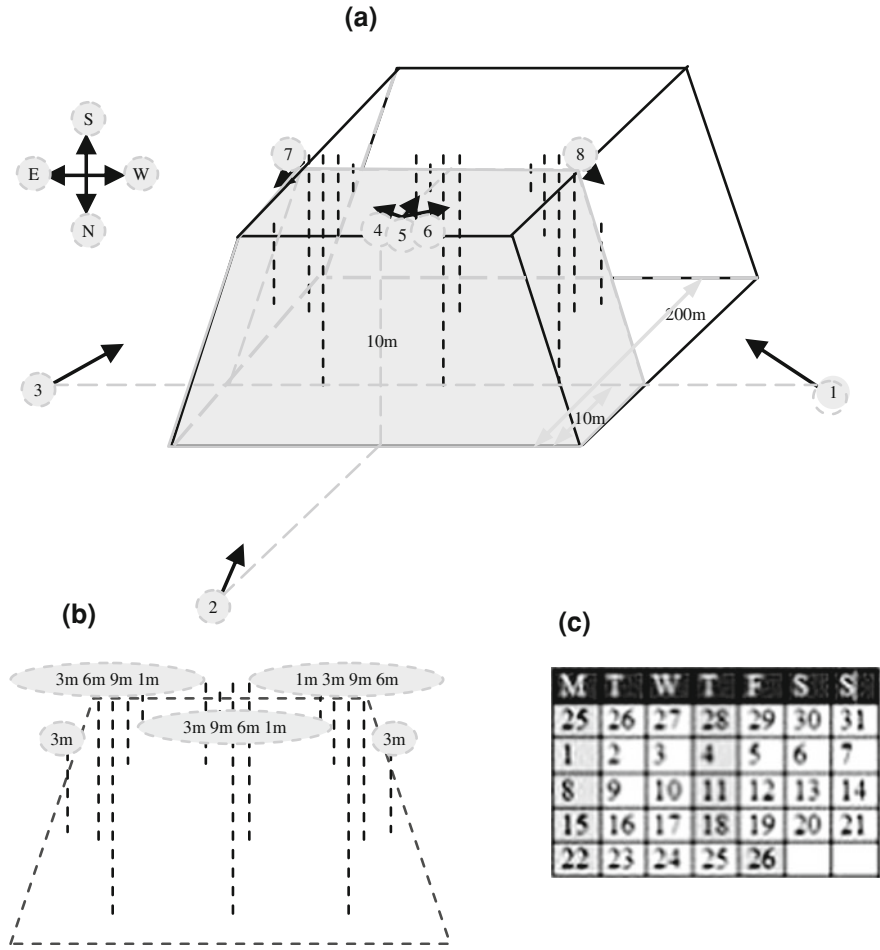


Fig. 100.6 **a** The surface of the chosen 5 % of the stockpile was monitored with the thermovision camera from 8 locations, **b** side view of the surface, **c** Recorded data

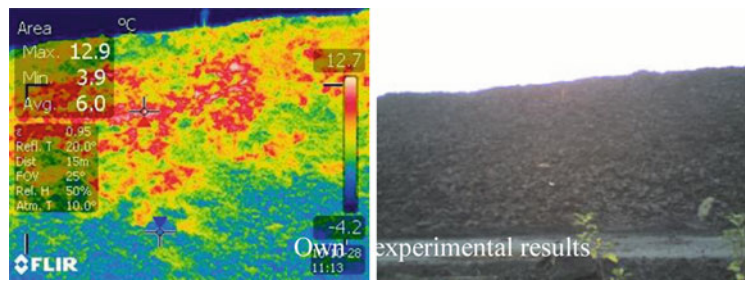


Fig. 100.7 Images took from location 1: *left*-thermogram, *right*-visible

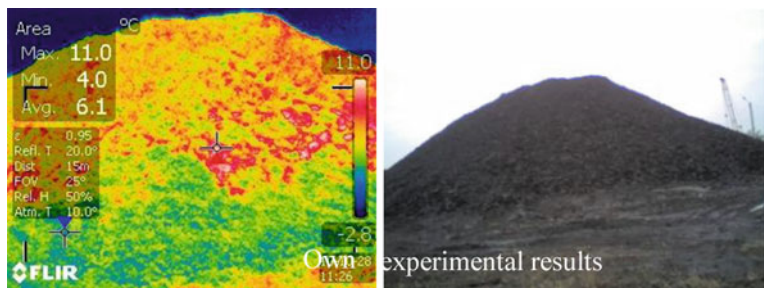


Fig. 100.8 Images took from location 2: *left-thermogram, right-visible*

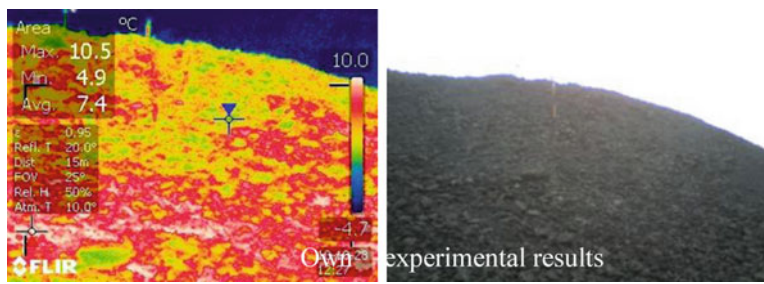


Fig. 100.9 Images took from location 3: *left-thermogram, right-visible*

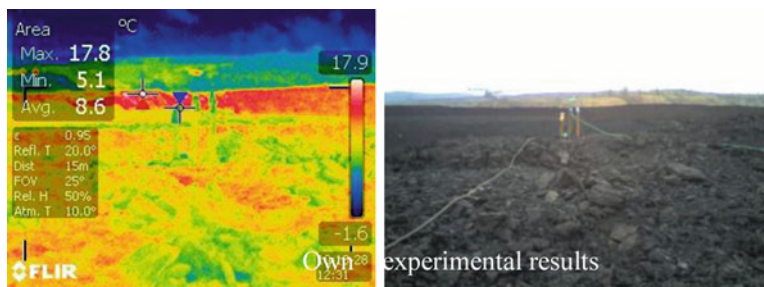


Fig. 100.10 Images took from location 4: *left-thermogram, right-visible*

deposit. In fact, all 8 locations were chosen for the camera to cover areas around the pipes. Also, an outdoor thermometer was put on sight for air temperature.

In Fig. 100.6b is a section of the monitored stockpile which shows the position and dimensions of each pipe and in Fig. 100.6c is the calendar for the months October and November 2010 when the monitoring took place. With light grey are the days (25, 28 of October 2010 and 1, 4, 8, 11, 15, 18, 22, 26 of November 2010) in which I monitored the stockpile with the thermovision camera from the 8 positions mentioned above.

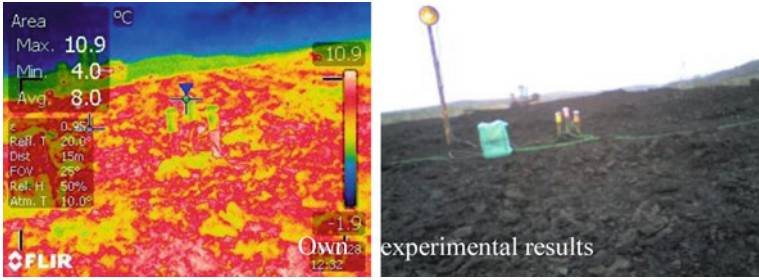


Fig. 100.11 Images took from location 5: *left*-thermogram, *right*-visible

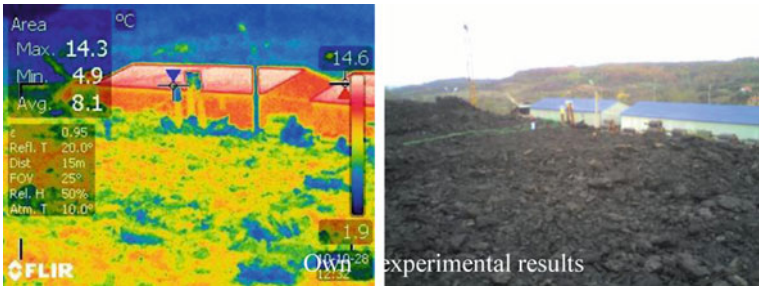


Fig. 100.12 Images took from location 6: *left*-thermogram, *right*-visible

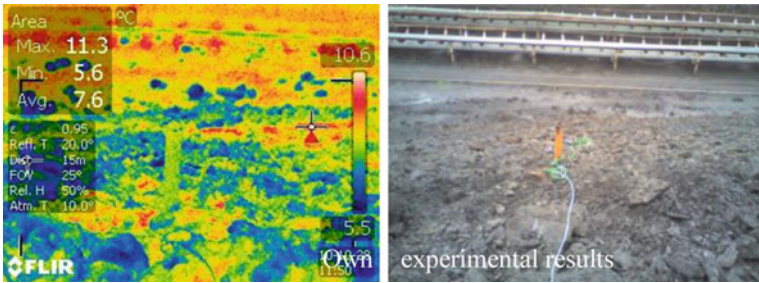


Fig. 100.13 Images took from location 7: *left*-thermogram, *right*-visible

In each of the days above, I monitored the chosen part of the stockpile with the thermovision camera. I took some thermal images at different moments of time, but only between 10:46 A.M. and 15:50 P.M.

In the following images (Figs. 100.7, 100.8, 100.9, 100.10, 100.11, 100.12, 100.13, 100.14, 100.15) are few of the thermograms taken from each of the 8 positions, shown here to exemplify the monitoring method with the thermovision camera. Also their corresponding digital image is at its right. Data of the thermograms is 28.10.2010.

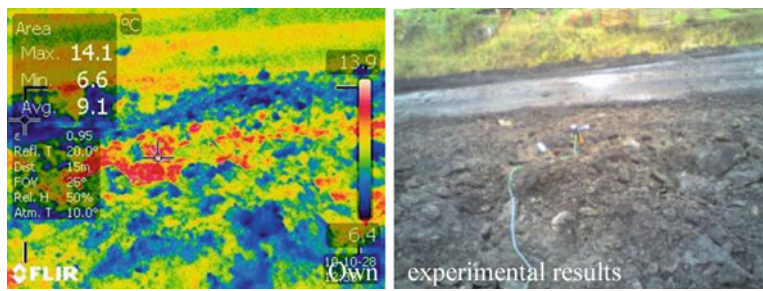


Fig. 100.14 Images took from location 8: *left*-thermogram, *right*-visible

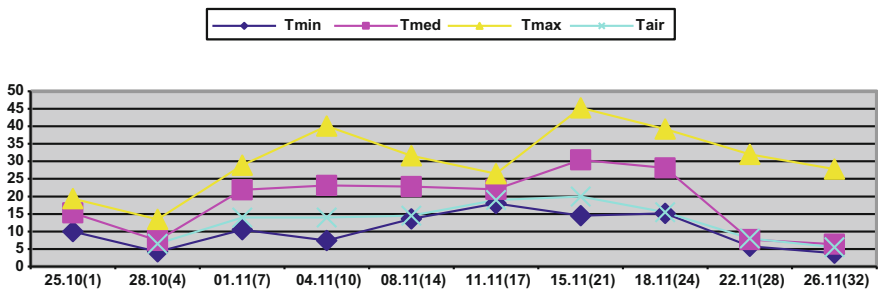


Fig. 100.15 Variation of temperature in time—view from the bottom—western side

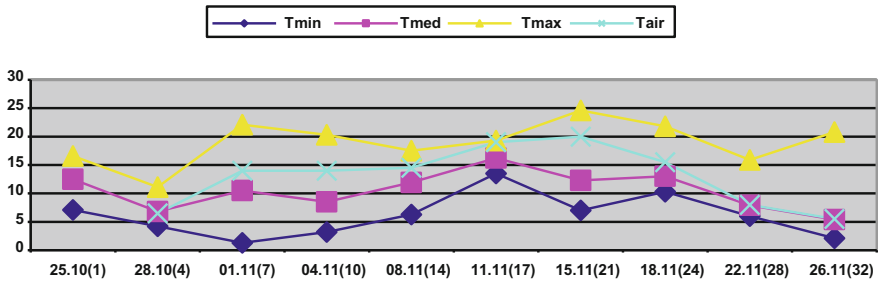


Fig. 100.16 Variation of temperature in time—view from the bottom—northern side

On each image, there is a predefined area for which I recorded the maximum, minimum and average temperature. All these data can also be seen on the thermograms below.

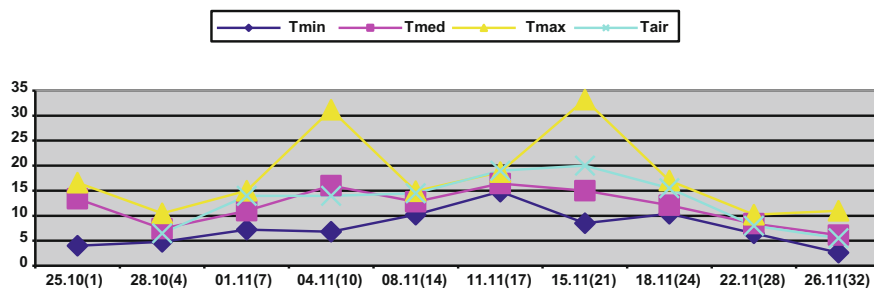


Fig. 100.17 Variation of temperature in time—view from the bottom—Eastern side

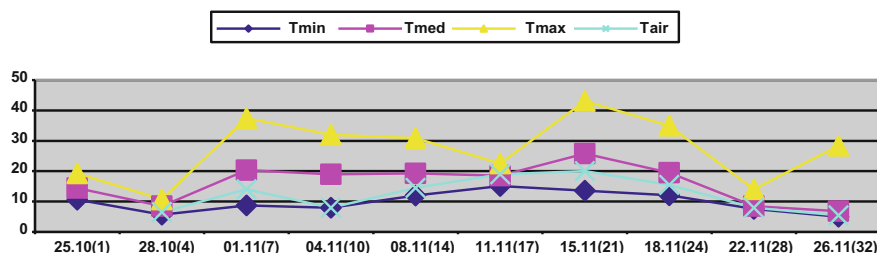


Fig. 100.18 Variation of temperature in time—view top—pipes—eastern side

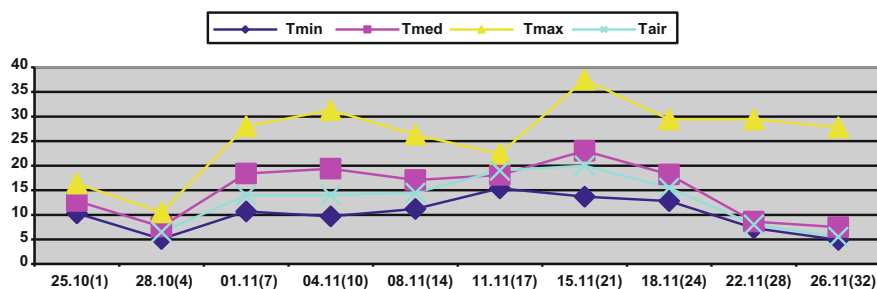


Fig. 100.19 Variation of temperature in time—view top—pipes—center side

100.3 Diagnosing the State of the Lignite Deposit

In order to determine the state of the lignite deposit, in the following graphics (Figs. 100.15, 100.16, 100.17, 100.18, 100.19, 100.20, 100.21, 100.22) I put values of temperature in time on the predefined area: the minimum (dark blue), maximum (yellow) and average value (pink) of temperature from each day of monitoring with the thermovision camera and also the air temperature (light blue) taken with the outdoor thermometer. With purple there are two hot spots.

From the graphics above I have drawn the following conclusions:

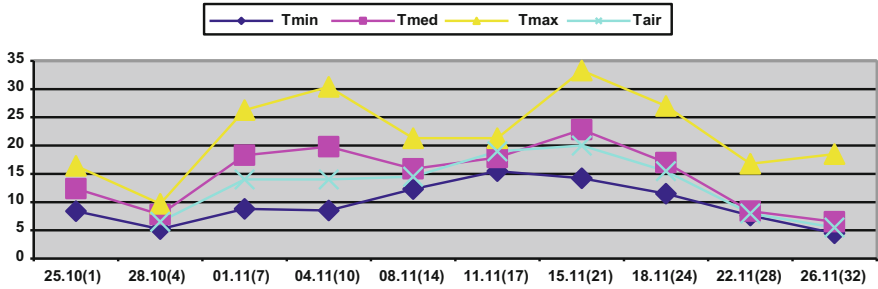


Fig. 100.20 Variation of temperature in time—view top—pipes—western side

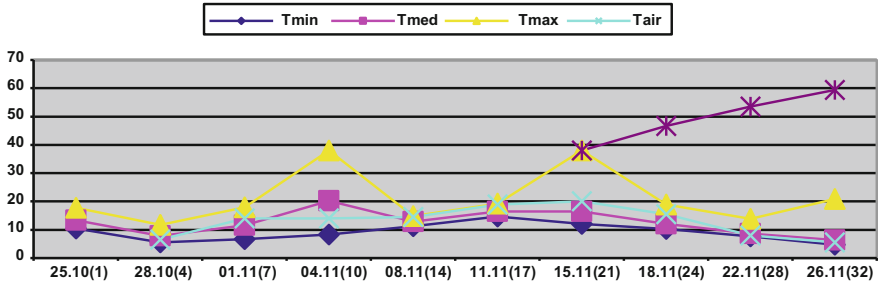


Fig. 100.21 Variation of temperature in time—view from the top—Eastern side

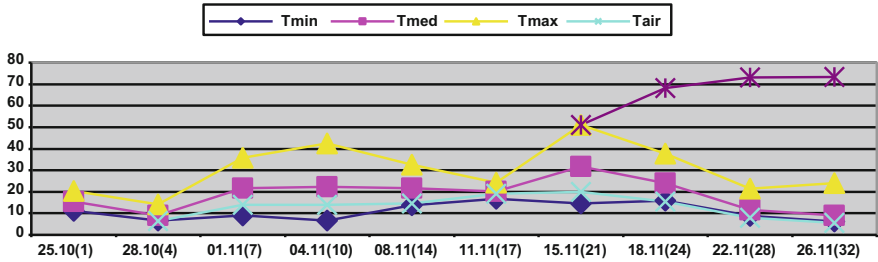


Fig. 100.22 Variation of temperature in time—view from the top—western side

- temperatures of the surface of the western side are higher than northern and eastern where the lowest values are;
- the maximum value of temperature is always higher than the air temperature;
- in the 22nd day of observation, on a small area on the top, close to the edge of the eastern side, temperature is 38 °C—over the value of State 1; also, in the 22nd day of observation, on another area from the top, but close to the edge of the western side, a hot spot appears of 51 °C—value close to State 2, and in less than two weeks, both of the temperatures mentioned above increased rapidly to almost 60 °C respectively to 73 °C.

100.4 Conclusions

No correlation could be established between the temperature at the surface of the lignite stockpile that was monitored with a thermovision camera and the data from the temperature sensors within the stockpile. The diagnosed state for the area in the stockpile that I used in this paper's research was State 1 to State 2, and only a small area of it in State 3, which was partially incorrect, as long as some areas within the stockpile, at different depths, were actually being in advanced State 3 and the thermovision camera couldn't receive any thermal signals from them. The temperature sensors identified that on the eastern area, at 8 m depth, the temperature increased in the last 2 weeks of observation at over 80 °C. Also, on the western area, at 6–7 m depth, another sensor detected a temperature increase at over 70 °C, both values undetected by the thermovision camera. If there were no temperature sensors within the deposit, my diagnosis would have been incomplete and incorrect.

As a final conclusion I realized that in the conditions mentioned above, the state of a lignite stockpile cannot be diagnosed correctly or completely by monitoring the surface temperature of the stockpile with a thermovision camera because lignite is a low temperature transmitter and the surface temperature of a stockpile is more likely to be influenced by the environmental factors such as wind, sun, rain, rather than by a strong heat-emitting hot spot at a 3–9 m depth within the stockpile which seems not to radiate enough to influence it.

I still believe that a continuous (on-line) monitoring of the surface of a lignite stockpile with a thermovision camera positioned somehow above the stockpile could save more data which could “say” a lot more than what my previous research said and could diagnose correctly the state of the deposit.

References

1. Research project between “Constantin Brancusi” University of Targu-Jiu and National Society of Lignite from Oltenia (N.S.L.O) Targu-Jiu as beneficiary, project no. 302/S/07.05.2010 - The management of N.S.L.O—Targu-Jiu coal deposits in order to improve the quality of coal and the thermographic expertise of coal piles to prevent the self ignition phenomenon, May–Nov 2010, Alina Dinca—member of the team
2. Internal piece of legislation of N.S.L.O Targu-Jiu—Technical instructions for coal depositing
3. Huidu E (1996) Coal storing systems. Ed. Tehnica, Bucuresti
4. Racocceanu C, Popescu L-G, Popescu C (2011) Research on factors affecting auto-Ignition of coal energy. Models Methods Appl Sci, WSEAS, Drobeta, pp 79–84. ISBN:987-1-61804-044-2. <http://www.wseas.us/e-library/conferences/2011/Drobeta/IAASAT/IAASAT-12.pdf>
5. Cruceru M, Diaconu B, Popescu L (2011) Self oxidation of Romanian lignite during storage. Recent Res Energy Environ, WSEAS, Drobeta, pp 335–340. ISBN:978-960-474-274-5. <http://www.wseas.us/e-library/conferences/2011/Cambridge/EE/EE-55.pdf>
6. Dinca A, Cercel C (2008) Environment protection through detection of hot spots using thermography in coal deposits before self ignition. Annals of the “Constantin Brâncuși”

University of Târgu Jiu, ECOTREND 2008, Economy Series, Issue 3/2009, pp 163–172.
ISSN:1844-7007

7. http://www.flir.com/uploadedFiles/CS_EMEA/Application_Stories/Media/Downloads/Nastup_EN.pdf
8. Flir Systems T200—User's Manual

Chapter 101

A Framework Intelligent Mobile for Diagnosis Contact Lenses by Applying Case Based Reasoning

Eljilani Mohammed

Abstract This paper is lead to the development of the system that allows detection of which suitable type of lenses thus will minimize the margin of destruction. The proposed system serves as advisory system to assist in predicting type of lenses in the early stage, through analysis of similar historical events of several aspects of paddy diseases cases and features which are stored in time-series form (temporal information), restructured and re-designed into case based format. The AI technique, case based reasoning (CBR) supports the process of finding the similarity.

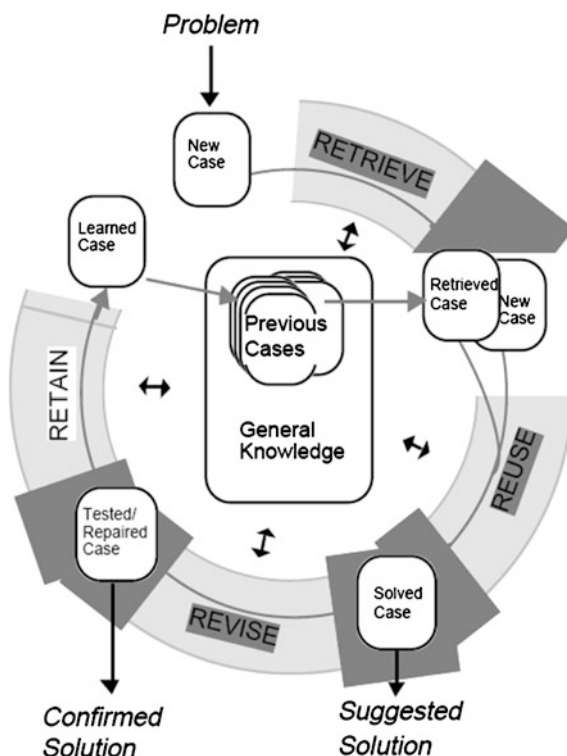
101.1 Introduction

The ability to learn is one of the most defining characteristics of intelligent behavior, including the process of learning many things, acquisition of new knowledge and develops the skills of being aware of through the practical application and the discovery of a new skill through observation and experience. Computers can't be considered intelligent only if they have the ability to learn, including a possible ability to do new things and adapt to new situations rather than to implement all the work ordered it without the benefit [1–3].

Learning algorithms uses several of technique to solve problems called case based reasoning is a general paradigm for reasoning from experience. It assumes a memory model for representing, indexing, organizing past cases and a process model for retrieving and modifying old cases and assimilating new ones.

E. Mohammed (✉)

Department of computer Science, Sebha University, Sebha, Libya
e-mail: jilani@sebhau.edu.ly

Fig. 101.1 The CBR cycle

101.2 The Case Based Reasoning Cycle

At the highest level of generality, a general CBR cycle may be described by the following four processes:

1. Retrieve: the most similar case or cases.
2. Reuse: the information and knowledge in that case to solve the problem.
3. Revise: the proposed solution.
4. Retain: the parts of this experience likely to be useful for future problem solving.

A new problem is solved by retrieving one or more previously experienced cases, reusing the case in one way or another, revising the solution based on reusing a previous case, and retaining the new experience by incorporating it into the existing knowledge-base (case-base). The four processes each involve a number of more specific steps, which will be described in the task model. In Fig. 101.1, this cycle is illustrated.

Fig. 101.2 Sample of server log file data

| Classes | tear production rate | astigmatic | spectacle prescription | age | ID |
|---------|----------------------|------------|------------------------|-----|----|
| 3 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 1 | 1 | 1 | 2 |
| 3 | 1 | 2 | 1 | 1 | 3 |
| 1 | 2 | 2 | 1 | 1 | 4 |
| 3 | 1 | 1 | 2 | 1 | 5 |
| 2 | 2 | 1 | 2 | 1 | 6 |
| 3 | 1 | 2 | 2 | 1 | 7 |
| 1 | 2 | 2 | 2 | 1 | 8 |
| 3 | 1 | 1 | 1 | 2 | 9 |
| 2 | 2 | 1 | 1 | 2 | 10 |

101.2.1
Experimental Work

101.2.1.1
General Knowledge (Previous Cases)

Data Acquisition

- (a) Donor: Benoit Julien (Julien@ce.cmu.edu)
- (b) <http://archive.ics.uci.edu/ml/datasets/Lenses>

Data Description

The data has been acquired from UCI Machine Learning, This data set includes descriptions of Database for fitting contact lenses includes four attributes and one class attribute. The total number of instances is 24.

Attribute Information:

1. Age of the patient: (1) young, (2) pre-presbyopic, (3) presbyopic.
2. Spectacle prescription: (1) myope, (2) hypermetrope.
3. Astigmatic: (1) no, (2) yes.
4. Tear production rate: (1) reduced, (2) normal.

Three Classes

- 1 :the patient should be fitted with hard contact lenses,
- 2 :the patient should be fitted with soft contact lenses,
- 3 :the patient should not be fitted with contact lenses.

Sample of Data

Sample of raw data is shown in Fig. 101.2

101.2.1.2
Calculating Similarity Between Cases

At the heart of a CBR system is the computation of similarity between a new case—the user’s input—and previous cases stored in a case base. Cases are

Fig. 101.3 Local similarity’ pseudo code

Begin

Input: the cases in case base with the new case,

Output: the result of the calculation local similarity of each case,

For I← 0 to all Case’ features Do

Sim i (a, b) = |a - b| / range

Sim i (a, b) = min (1, Sim i (a, b))

End

Return result

End

associated with qualitative and quantitative parameters called features or attributes, which represent the important facts about each case. The CBR algorithm calculates the similarity between cases based on feature value pairs of the new and each historical case. A similarity measure has the following attributes:

- Reflective: a case always is similar to itself
- Symmetric: If A is similar to B, then B is similar to A

(a) Local Similarity

Similarity between two cases is based on the similarity between the two cases’ feature. The Local similarity calculation depends on the type of the feature. Similarity can be calculated for numeric and non-numeric value. For this study, all the cases’ features are numeric. Therefore, the following local similarity formulation is used (Fig. 101.3):

$$\text{Sim (a, b)} = |a - b|/\text{range}$$

- where:
- a** is the value of new case’ feature.
 - b** is the value of old case’ feature.
 - Range** is the absolute value of difference between the upper and lower boundary of the set.

(b) Global Similarity

Once a set of local similarities has been calculated for each feature, the CBR system calculates the *global similarity* of the candidate cases. No single similarity measure is perfectly appropriate for all domains, and CBR systems use different global similarity measures to provide acceptable case-matching behavior. One simple approach to measure the similarity between two cases A and B with p features is weighted-block-city:

$$\text{Global} = [1 - (\text{case distance}/\text{sum(w)}) * 100]$$


```

Begin
Input: all calculated local similarity,
Output: the result of the calculation Global similarity,
For I← 0 to all Case' features Do
     $\text{sim } i = \sum \text{sim}(a, b) * \omega(i)$ 
     $\text{Global sim} = (1 - (\text{Local sim } i / \sum \omega(i))) * 100$ 
End
Return the highest similarity
End

```

Fig. 101.4 Global similarity' pseudo code

where:

Case distance: is the local similarity calculated.

Sum (w): is the weight of the attribute.

Case distance = localism * sum (w).

The general flow of the process (Fig. 101.4):

101.3 Evaluation

The evaluation was performed to determine the accuracy of the system. However, In order to determine the accuracy of the algorithm used for this application, the accuracy of the algorithm was tested by taking one of the cases in case base as a test case. If the result returns 100 % similarity then it would suggest that the algorithm is sound.

$$\text{Accuracy} = (\text{Totalof correctness} / \text{Total cases}) * 100$$

101.4 Discussion

The system starts with screen for user to input the value of the new cases feature in the specified field. User must input four cases observed data to run the diagnosis with three actors, Firstly the user who interacts with the system. The second is the client Pocket PC. The third entity is the server. The system calculates the range of each attribute in the case base. The value of new case was compared with every value of the same attribute in the case base to find local similarity for that particular attribute. The same process was repeated with the other attribute in the case base. Finally, the system calculates the global similarity for the new case compared to each case in the case base. The result of the diagnosis then will be

presented into a message box to the user. The top three highest similarities are presented.

101.5 Conclusions

This paper discussed the framework of mobile intelligent application process based on CBR for diagnosis of the type of contact lenses for the patient according to previously undiagnosed cases. That can aid significantly in improving the decision making of the physicians. These systems help physicians and doctors to check, analyze and repair their solutions. The physician's inputs a description of the domain situation and their solutions and the system can recalls cases with similar solutions and presents their outcomes to the physician.

References

1. Holt A, Benwell GL (1999) Applying case-based reasoning techniques in GIS. *Int J Geogr Inf Sci* 13(1):9–25
2. Abubaker E (2007) Intelligent mobile paddy diseases using fuzzy temporal information
3. Salem A-BM (2007) Case based reasoning technology for medical diagnosis. *World Acad Sci Eng Technol* 31:9–13