Hanning Yuan
Jing Geng
Fuling Bian (Eds.)

Communications in Computer and Information Science 699

Geo-Spatial Knowledge and Intelligence

4th International Conference
on Geo-Informatics in Resource Management
and Sustainable Ecosystem, GRMSE 2016
Hong Kong, China, November 18–20, 2016
Revised Selected Papers, Part II

Part 2

Springer

Communications in Computer and Information Science 699

Commenced Publication in 2007
Founding and Former Series Editors:
Alfredo Cuzzocrea, Dominik Ślęzak, and Xiaokang Yang

Editorial Board

More information about this series at http://www.springer.com/series/7899

Hanning Yuan · Jing Geng
Fuling Bian (Eds.)

Geo-Spatial Knowledge and Intelligence

4th International Conference
on Geo-Informatics in Resource Management
and Sustainable Ecosystem, GRMSE 2016
Hong Kong, China, November 18–20, 2016
Revised Selected Papers, Part II

Springer

Editors
Hanning Yuan
Beijing Institute of Technology
Beijing
China

Fuling Bian
Wuhan University
Wuhan
China

Jing Geng
Beijing Institute of Technology
Beijing
China

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-981-10-3968-3 ISBN 978-981-10-3969-0 (eBook)
DOI 10.1007/978-981-10-3969-0

Library of Congress Control Number: 2017932437

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

The 4th Annual 2016 International Conference on Geo-Informatics in Resource Management and Sustainable Ecosystem (GRMSE 2016) was held in Hong Kong, China, during November 18–20, 2016. It aims to bring researchers, engineers, and students to the areas of geo-spatial information science, engineering, and systems in socioeconomic development, resource management, and sustainable ecosystem. GRMSE 2016 features unique mixed topics of spatial data mining, geographical information science, photogrammetry and remote sensing, data science, data engineering, cloud computing, deep learning, and recent applications in the context of building a smarter planet, healthier life, more enjoyable ecology and more sustainable resources.

We received a total of 311 submissions from various parts of the world. The international Program Committee worked very hard to have all papers peer-peer reviewed before the review deadline. The final program consisted of 118 papers. There were four key note speeches and five invited sessions. All the keynote speakers are internationally recognized leading experts in their research fields, who have demonstrated outstanding proficiency and have achieved distinction in their profession. The proceedings are published as a volume in Springer's *Communications in Computer and Information Science* (CCIS) series. Some excellent papers were selected and recommended to the special issue of *Journal of Environmental Science and Pollution*, a Science Citation Index Expanded journal. We would like to mention that, due to the limitation of the conference venue capacity, we were not able to include many fine papers in the program. Our apology goes to those authors.

We would like to express our sincere gratitude to all the members of international Program Committee and organizers for their enthusiasm, time, and expertise. Our deep thanks also go to the many volunteers and staff members for the long hours and hard work they have generously given to GRMSE 2016. We are very grateful to Professor Fuling Bian, Professor Hui Lin and Professor Yichun Xie for their support in making GRMSE 2016 possible. The generous support from Beijing Institute of Technology is greatly appreciated. Finally, we would like to thank all the authors, speakers, and participants of this conference for their contributions to GRMSE 2016.

January 2017 General Chair

Organization

The Advisory Committee

Hui Lin	Institute of Space and Earth Information Science (ISEIS), The Chinese University of Hong Kong, Hong Kong
Qingquan Li	Shenzhen University, Shenzhen, China

Honorary General Chair

Fuling Bian	Wuhan University, China

General Co-chairs

Shuliang Wang	Beijing Institute of Technology, China
Yong Xia	Northwestern Polytechnical University, China
Hongzhi Wang	Harbin Institute of Technology, China

International Program Committee Co-chairs

George Christakos	San Diego State University, USA
Yangge Tian	Wuhan University, China
Qingwen Xiong	Wuhan University, China
Quan Zou	Tianjin University, China

International Editorial Committee Co-chairs

Fuling Bian	Wuhan University, Wuhan, China
Hanning Yuan	Beijing Institute of Technology, Beijing, China
Jing Geng	Beijing Institute of Technology, China

International Program Committee

Tao Chen	Tsinghua University, China
Chau Yuen	Singapore University of Technology and Design (SUTD), Singapore
Maytham Safar	Kuwait University, Kuwait
Alfrendo Satyanaga Nio	Nanyang Technological University, Singapore
Pengfei Zhang	Institute for Infocomm Research (I^2R), Singapore

Mohd Adib Bin Mohammad Razi	Universiti Tun Hussein Onn Malaysia, Malaysia
Hanning Yuan	Beijing Institute of Technology, China
Jing Geng	Beijing Institute of Technology, China
Huijun Yang	Northwest A&F University, China
Hongyi Li	Jiangxi University of Finance and Economics, China
Ismail Rakip Karas	Karabuk University, Turkey
Xianglin Zhan	Civil Aviation University of China, China
Ray-I Chang	National Taiwan University, China
Qunyong Wu	Fuzhou University, China
Qian He	Guilin University of Electronic Technology, China
Ken Chen	Chengdu University of Technology, China
Fuucheng Jiang	Tunghai University, Taiwan
Mohd Haziman Wan Ibriahim	Universiti Tun Hussein Onn Malaysia, Malaysia
Ho Pham Huy Anh	Ho Chi Minh City University of Technology (HUT), Vietnam
Le Sun	Victoria University, Melbourne, Australia
Xia Zhang	Wuhan University, China
Mojtaba Maghrebi	University of New South Wales, Australia
Maciej Zieba	Wroclaw University of Technology, Poland
Jianguo Sun	Harbin Engineering University, China
Ulas Akkucuk	Bogazici University, Turkey
Cheng-Yuan Tang	Huafan University, Taiwan
Mohammed A. Akour	Yarmouk University, Jordan
Chien-Hung Yeh	Feng Chia University, Taiwan
Yi-Kuei Lin	National Taiwan University of Science & Technology (Taiwan Tech), Taiwan
Zongyao Sha	Wuhan University, China
George Christakos	San Diego State University, USA
Ping Fang	Tongji University, China
Kuishuang Feng	University of Maryland, USA
Nanshan Zheng	China University of Mining and Technology, China
Changsheng Cai	Central South University, China
Zhenhong Li	University of Glasgow, UK
Yuqi Bai	Tsinghua University, China
Sabine Baumann	Technische Universität München, Germany
Qinghui Huang	Tongji University, China
David Forrest	University of Glasgow, UK
Arie Croitoru	George Mason University, USA
James Cheng	Manchester Metropolitan University, UK
Paul Torrens	University of Maryland, USA
Stephan Mäs	Technische Universität Dresden, Germany

Gina Cavan	Manchester Metropolitan University, UK
Jan Dempewolf	University of Maryland, USA
Bor-Wen Tsai	National Taiwan University, Taiwan
Yu Liu	Peking University, China
Xiaojun Yang	Florida State University, USA
Yan Liu	The University of Queensland, Australia
Jinling Wang	University of New South Wales, Australia
Xiaolei Li	Wuhan University, China
Pariwate Varnakovida	Prince of Songkla University, Thailand
Manfred F. Buchroithner	Technische Universität Dresden, Germany
Anthony Stefanidis	George Mason University, USA
Chaowei Yang	George Mason University, USA
Xiaoxiang Zhu	Technische Universität München, Germany
Matt Rice	George Mason University, USA
Jianjun Bai	Shaanxi Normal University, China
Yongmei Lu	Texas State University, USA
Alberta Albertella	Technische Universität München, Germany
F. Benjamin Zhan	Texas State University, USA
Huamin Wang	Wuhan University, China
Edwin Chow	Texas State University, USA
Lin Liu	University of Cincinnati, USA
Shuqiang Huang	JiNan University, China
Weihua Dong	Beijing Normal University, China
Mengxue Li	University of Maryland, USA
Wenwen Li	Arizona State University, USA
André Skupin	San Diego State University, USA
Alan Murray	Arizona State University, USA
Mike Worboys	The University of Maine, USA
Amirhossein Sajadi	Case Western Reserve University, USA
Chien-Hung Yeh	Feng Chia University, China
Helmi Zulhaidi Mohd Shafri	Universiti Putra Malaysia, Malaysia
Peng-Sheng Wei	National Sun Yat-Sen University, Taiwan
Maria Hallo	Notre Dame University, Belgium
Jingyu Yang	Shenyang Aerospace University, China
Zulkifli Mohd Rosli	Universiti Teknikal Malaysia Melaka, Malaysia
M. Thang Trung Nguyen	Ton Duc Thang University, Vietnam
Chan King-ming	Hong Kong, SAR China
Huan Yu	Chengdu University of Technology, China
Yong Xia	Northwestern Polytechnical University (NPU), China
Rosmayati Binti Mohemad	Universiti Malaysia Terengganu, Malaysia
Sedat Keleş	Çankırı Karatekin University, Turkey
Yanying Chen	Meteorological Science Institute of Chongqing, China
Xiukai Ruan	Wenzhou University, China

Togay Ozbakkaloglu	The University of Adelaide, Australia
Xicheng Tan	Wuhan University, China
Tomasz Andrysiak	UTP University of Science and Technology, Poland
Ping Zhang	Jilin University, China
Ting Yang	Tianjin University, China
Yo-Sheng Lin	National Chi Nan University, Taiwan
Imran Memon	Zhejiang University, Hangzhou, China
Megat Farez Azril	Universiti Kuala Lumpur, Malaysia
Ximing Fu	Tsinghua University, China
Jiann-Shu Lee	National University of Tainan, Taiwan
Dandan Ma	University of Chinese Academy of Sciences, China
Zhiyu Jiang	University of Chinese Academy of Sciences, China
Huada Daniel Ruan	Beijing Normal University-Hong Kong Baptist University United International College (UIC), Zhuhai, China
Wong Man Sing Charles	Hong Kong Polytechnic University, China
Pensyarah Nursabillilah Binti Mohd Ali	Universiti Teknikal Malaysia Melaka, Malaysia
Aldy Gunawan	Singapore Management University, Singapore
Rana Rahim-Amoud	Lebanese University, Lebanon
Hui Yang	Beijing University of Posts and Telecommunications, China
Zuraidi Saad	Universiti of Teknologi MARA, Malaysia
Lixin Wang	Paine College, USA
Weimo Liu	George Washington University, USA
Jianping Chen	China University of Geosciences, China
Indranil SenGupta	North Dakota State University, USA
Muhammad Tauhidur Rahman	King Fahd University of Petroleum & Minerals (KFUPM), Saudi Arabia
Delia B. Senoro	Mapua Institute of Technology Manila, Philippines
Zengxiang Li	Institute of High Performance Computing, Singapore
Chee-Ming Chan	Universiti Tun Hussein Onn Malaysia, Malaysia
Agnieszka Cydzik-Kwiatkowska	University of Warmia and Mazury in Olsztyn, Poland
Yi-You Hou	Southern Taiwan University of Science and Technology, Taiwan
Maguid H.M. Hassan	The British University in Egypt (BUE), Egypt
Peng-Yeng Yin	National Chi Nan University, Taiwan
Shian-Chang Huang	National Changhua University of Education, Taiwan
Nor Amani Filzah Bt. Mohd Kamil	University Tun Husseion Onn Malaysia, Malaysia
Artur Krawczyk	AGH University of Science and Technology, Poland

Guoqing Li	Institute of Soil and Water Conservation, CAS & MWR, China
Jinghu Pan	Northwest Normal University, China
Guodong Wang	South Dakota School of Mines and Technology, USA
Hongzhi Wang	Harbin Institute of Technology, China
Bin Liu	Dalian University of Technology, China
Xin Yan	Wuhan University of Technology, China
Ali Karrech	University of Western Australia, Australia
Syed Abdul Rehman Khan	Iqra University and Brasi School of Supply Chain Management, USA
Saouli Hamza	University Khider Mohamed, Algeria
Huey-Ming Lee	Chinese Culture University, Taiwan
Lily Lin	China University of Technology, Taiwan
Jolanta Mizera-Pietraszko	Opole University, Poland
Hanmin Jung	Korea Institute of Science and Technology Information (KISTI), South Korea
Chenfei Gao	AT&T Labs Research
Qiang Gao	Beihang University, Beijing, China
Ben-Shun Yi	Wuhan University, China
Yong Xia	Northwestern Polytechnical University, China
Yun-Xiao Zu	Beijing University of Posts and Telecommunications, China
Jen-Fa Huang	Electrical Engineering, National Cheng Kung University, Taiwan
Jian Wang	Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, China
Tzong-Yi Lee	Yuan Ze University, Taiwan
Wei-Chiang Wu	Da-Yeh University, Taiwan
Wen-Tsai Sung	National Chin-Yi University of Technology, Taiwan
Faizal Mustapha	Universiti Putra Malaysia, Malaysia
Chin-Ling Chen	Chaoyang University of Technology, Taiwan
Nursabillilah Binti Mohd Ali	Universiti Teknikal Malaysia Melaka, Malaysia
Zhen-Dong Wang	Jiangxi University of Science and Technology, China
Sina Vafi	Charles Darwin University, Australia
Trong-Minh Hoang	Posts and Telecommunication Institute of Technology, Vietnam
Deng Chen	Wuhan Institute of Technology, China
Yuan-Long Cao	Jiangxi Normal University, China
Xi-Ming Fu	Tsinghua University, China
Tian-Hua Xu	University College London, UK
Malka N. Halgamuge	University of Melbourne, Australia

Jing-Yu Yang	Shenyang Aerospace University, China
Fang-Jun Huang	Sun Yat-sen University, China
Ying-Ji Zhong	Ohio State University, USA
Jian-Guo Sun	Harbin Engineering University, China
Yi-Fei Wei	Beijing University of Posts and Telecommunications, China
Chi-Wai Kan	Hong Kong Polytechnic University, SAR China
Shih-Chuan Yeh	De Lin Institute of Technology, Taiwan
Muh-Tian Shiue	National Central University, Jhongli, China
Sarmad Sohaib	University of Engineering and Technology, Taxila, Pakistan
Yasin Kabalci	Nigde University, Turkey
Tomasz Andrysiak	University of Science and Technology, Poland
Marcin Kowalczyk	Warsaw University of Technology, Poland
I-Shyan Hwang	Yuan Ze University, Chung-Li, China
Cheng-Yuan Tang	New Taipei City, Taiwan
Yu-Chen Hu	Providence University, Taiwan
Megat Farez Azril	Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Malaysia
Chang-Yu Liu	South China Agricultural University, China
Prosanta Gope	Singapore University of Technology and Design, Singapore
Ming-Jian Li	University of Wisconsin Madison, USA
Choi Jaeho	Chonbuk National University, South Korea
Muhammed Enes Bayrakdar	Duzce University, Turkey
Rkia Aouinatou	Mohamed V Agdal Rabat, Rabat, Morocco
Najeeb Ullah Khan	CECOS University, KPK, Pakistan
Sadaqat Jan	University of Engineering & Technology, Peshawar, Pakistan
Yee-Jin Cheon	University of Science and Technology, Daejon, South Korea
K. Balakrishnan	Karpaga Vinayaga College of Engineering and Technology, Chennai, India
Imran Memon	Zhejiang University, China
Bongani Ngwenya	Solusi University, Zimbabwe
Alexey Nekrasov	Southern Federal University, Taganrog, Russia
Dmitry Popov	Moscow State University of Printing Arts, Russia
Qing-zheng Xu	Xi'an Communications Institute, China
Hsing-Chung Chen	Asia University, Taiwan
Muhammad Zeeshan	National University of Sciences & Technology, Pakistan
Chi-Wai Chow	National Chiao Tung University, Taiwan
Yair Wiseman	Holon Institute of Technology, Israel

Panayotis Nastou	University of the Aegean, Samos, Greece
Arianna D'Ulizia	University of Rome La Sapienza, Rome
D.M. D'Addona	University of Naples Federico II, Italy
Gihwan Cho	Chonbuk National University, South Korea
M. Arunachalam	K.L.N College of Information Technology, India
Parvaneh Mansouri	Azad University, Iran
José Manuel Machado	University of Minho, Portugal
Bartłomiej Płaczek	University of Silesia, Poland
Ittipong Khemapech	University of the Thai Chamber of Commerce, Thailand
Yang Yue	Juniper Networks, USA
Abul Bashar	Prince Mohammad Bin Fahd University, Kingdom of Saudi Arabia
Abderrahmen Mtibaa	Texas A&M University, Qatar
Michael S. Okundamiya	Ambrose Alli University, Nigeria
Hanmin Jung	Korea Institute of Science and Technology Information, South Korea
Wen-Jie Zhang	Minnan Normal University, China
MdArafatur Rahman	University of Naples Federico II, Italy
Hung-Chun Chien	Jinwen University of Science and Technology, Taiwan
Hari Mohan Rai	Krishna College of Engineering, Ghaziabad, India
Yogendra Kumar Jain	Samrat Ashok Technological Institute, India
Rahul Dutta	Oracle India Pvt. Ltd., India
Anqi He	Queen Mary University of London, UK
Arnulfo Luévanos Rojas	Autonomous University of Coahuila, México
Di Zhang	Waseda University, Japan
Janusz Wielki	University of Warsaw, Poland
Ben Wu	Princeton University, USA
Yue Cao	University of Surrey, UK
Qiang Qu	Innopolis University, Russia
Piotr Kulczycki	Polish Academy of Sciences, Poland
Hyunsung Kim	Kyungil University, Korea
Hassene Seddik	ENSIT Tunisia, Tunisia
Liang Zhao	Georgia Gwinnett College, USA
Ivo Stachiv	National Taiwan University, Taiwan
Phongsak Phakamach	Royal Thai Army, Thailand
Ashok Kumar Kulkarni	Malla Reddy Institute of Medical Sciences, Thailand
Gurjot Singh Gaba	Lovely Professional University, Jalandhar, Punjab, India
Dimitris Kanellopoulos	University of Patras, Greece
Ljiljana Trajkovic	Simon Fraser University, Canada
Chenfei Gao	AT&T Labs, USA
Elsayed Esam M. Khaled	Assiut University, Egypt

Rukhsana Ruby	Shenzhen University, China
Basel Ali Mahafzah	The University of Jordan, Jordan
Alexandru Vulpe	University Politehnica of Bucharest, Romania
Luis Gomez Deniz	University of Las Palmas de Gran Canaria, Spain
Guodong Wang	Chinese Academy of Sciences, China
Luca Reggiani	Politecnico di Milano, Italy
Jianzhou Zhao	Cadence Design System, China
R. Raja	Alagappa University, India
Basile Christaras	Aristotle University of Thessaloniki, Greece
Mirko Barbuto	Roma Tre University, Italy
Roberto Nardone	University of Naples Federico II, Italy
Kamran Arshad	Ajman University of Science and Technology, UAE
Janusz Klink	Wroclaw University of Technology, Poland
Apostolos Gkamas	University Ecclesiastical Academy of Vella, Greece
Shadi G. Alawneh	Oakland University, USA
Alexandra Bousia	University of Thessaly, Greece
Houda Mzoughi	National Engineering School of Sfax, Tunisia
Emna Ben Slimane	National Engineering School of Tunis, Tunisia
Arun Agarwal	Siksha 'O' Anusandhan University, India
Klimis Ntalianis	Athens University of Applied Sciences, Greece
Imran Shafique Ansari	Texas A&M University at Qatar, Qatar
Paul Loh Ruen Chze	Nanyang Polytechnic, Singapore
Ismail Erturk	Kocaeli University, Turkey
Jiahu Qin	University of Science and Technology of China, Hefei, China
Min-Shiang Hwang	Asia University, Taiwan
Fangyong Hou	National University of Defense Technology, Changsha, China
Cheng-Yuan	Huafan University, Taiwan
Fangjun Huang	Sun Yat-sen University, China
Meng-Chou Chang	National Changhua University of Education, Taiwan
Liangxiao Jiang	China University of Geosciences, China
Wanan Xiong	University of Electronic Science and Technology of China, China
Tianhua Xu	University College London, London, UK
Andrzej Glowacz	AGH University of Science and Technology, Kraków, Poland
Rozaida Ghazali	Universiti Tun Husssein Onn Malaysia, Malaysia
Hongli Chen	ZheJiang Sci-Tech University, China
Mohamad Al Ladan	Haigazian University, Lebanon
Wanchen Huang	Wu Feng University, Minxiong, Taiwan
Tao-Ming Wang	Tunghai University, Taiwan

Rong-Jong Wai	National Taiwan University of Science and Technology, Taiwan
Xiuyan Ma	Dalian University of Technology, China
Lamei Zhang	Harbin Institute of Technology, China
Jyh-Cheng Chen	National Yang-Ming University, Taiwan
Yupeng Hu	Hunan University, China
Ying-Chun Chuang	Kun Shan University, Taiwan
Ahmet H. Ertas	Karabuk University, Turkey
Jianxun Zhang	Chongqing University of Technology, China
Aleksandra Mileva	Goce Delchev University, Macedonia
Hui-Mi Hsu	National Ilan University, Taiwan
Hamidah Ibrahim	Universiti Putra Malaysia, Kuala Lumpur, Malaysia
Yingji Zhong	Ohio State University, USA
Yun Lin	Harbin Engineering University, China
Guoming Lai	Guangdong Polytechnic of Science and Technology, China
Yinghua Zhou	Chongqing University of Posts and Telecommunications, China
Guojun Mao	Central University of Finance and Economics, China
Kurban Ubul	Xinjiang University, China
Ruipeng Ning	East China Normal University, China
Duanduan Chen	Beijing Institute of Technology, China
Zhiting Lin	Anhui University, China
Weiyu Yu	South China University of Technology, China
Hongjun Li	Beijing Forestry University, China
Liping Yang	Huazhong Agricultural University, China
Farn Wang	National Taiwan University, Taiwan
Lain-Chyr Hwang	I-Shou University, Taiwan
Mahmood K. Ibrahem Al Ubaidy	Al-Nahrain University, Iraq
Juin-Ling Tseng	Minghsin University of Science and Technology, Taiwan
Biju T. Sayed Mohammed	Dhofar University, Oman
Tran Cao Quyen	University of Engineering and Technology, Pakistan
Bappaditya Mandal	Institute for Infocomm Research, Singapore
Simon K.S. Cheung	The Open University of Hong Kong, Hong Kong, SAR China
Megat Farez Azril	System and Networking Section Universiti Kuala Lumpur, Malaysia
Massila Kamalrudin	Universiti Teknikal Malaysia Melaka, Malaysia
Lee Beng Yong	Universiti Teknologi MARA Sarawak, Malaysia
Andy Shui-Yu Lai	Technological and Higher Education Institute of Hong Kong, SAR China
Carlos Humberto Salgado	Universidad nacional de San Luis, Argentina

Adam Glowacz	AGH University of Science and Technology, Poland
Nur Sukinah Aziz	TATI University College, Malaysia
Krzysztof Gdawiec	University of Silesia, Poland
Chien-Hung Yeh	Feng Chia University, Taichung, Taiwan
Bai Li	Zhejiang University, Zhejiang, China
Ming Ming Wong	Sarawak Campus, Malaysia
Kai Tao	Nanyang Technological University, Singapore
Jun Ye	Sichuan University of Science & Engineering, China
Quanyi Liu	Tsinghua University, China
Zhendong Wang	Jiangxi University of Science and Technology, Ganzhou, China
Zhu Tang	National University of Defense Technology, China
Najam ul Hasan	Dhofar University, Oman
Chengyu Liu	Shandong University, Jinan, China
Sanjeevikumar Padmanaban	University of Johannesburg, South Africa
Fengqi Tan	University of Chinese Academy of Sciences, China
Bing Wen	Xinjiang Institute of Ecology and Chinese Academy of Science, China
Qiang Ye	Nanjing Institute of Physical Education and Sports, China
Shuai Liu	Inner Mongolia University, China
Yuhua Wang	Wuhan University of Science and Technology, China
Fei Huang	Ocean University of China, China
Sen Bai	Chongqing Communication Institute, China
Fali Cao	Xi'an Jiaotong University, China
Binyi Liu	Tongji University, China
Bo Cheng	Earth Observation & Digital Earth Chinese Academy of Sciences, China
Chun Shi	Hainan Normal University, China
Weichun Pan	Zhejiang Gongshang University, China
Sathaporn Monprapussorn	Srinakharinwirot University, Thailand
Seethalakshmi Rajashankar	SASTRA University, India
Partha Pratim Ray	Sikkim University, India
Wenchen Hu	University of North Dakota, USA
K.M. Suceendran	Tata Consultancy Services, India
Siwei Chen	National University of Defense Technology, China
Wei Chen	China University of Mining and Technology, China
Chuanfei Xu	Concordia University, Canada
Ti Peng	Southwest Jiaotong University, China
Jianjiao Chen	Georgia Institute of Technology, USA
Jinzhu Gao	University of the Pacific, USA
Lifeng Wei	Beijing University of Civil Engineering and Architecture, China
Rui Sun	Beijing Normal University, China

Anhua He	China Earthquake Administration, China
Ning Zhang	Beijing Union University, China
Imran Memon	Zhejiang University, Pakistan
Qian Tang	Xidian University, China
Xiaofei Zhang	Nanjing University of Aeronautics and Astronautics, China
Lianru Gao	Chinese Academy of Sciences, China
Liang Yang	Guangdong University of Technology, China
Zhenjiang Dong	Nanjing University of Science and Technology, China
Shuo Liu	Institute of Remote Sensing and Digital Earth Chinese Academy of Sciences, China
Qingke Wen	Institute of Remote Sensing and Digital Earth Chinese Academy of Sciences, China
Fan Ning	Beijing University of Posts and Telecommunications, China
Bo Cheng	Beijing University of Posts and Telecommunications, China
Tianhong Li	Peking University, China
Xiaofeng Wang	Chang'an University, China
Shuqing Hao	China University of Mining and Technology, China
Xianchuan Yu	Beijing Normal University, China
Zhaoyang Li	Jilin University, China
Shengcheng Cui	Chinese Academy of Sciences, China
Baiqiu Zhang	Jilin University, China
Yongzhi Wang	Jilin University, China
Ying Li	Dalian Maritime University, China
Chaokui Li	Hunan University of Science and Technology, China
Behshad Jodeiri Shokri	Hamedan University of Technology, Iran
Anand Nayyar	KCL Institute of Management and Technology, India
Hongjun Cao	Ocean University of China, China
Hong Fan	Institute of Remote Sensing and Digital Earth Chinese Academy of Sciences, China
Hyunsung Kim	Kyungil University, South Korea
B. Shanmugapriya	Sri Ramakrishna College of Arts and Science for Women, India
Erfeng Ren	Qinghai University, China
Qianli Ma	University of California, USA
Elena Simona Lohan	Tampere University of Technology, Finland
Laura Mónica Vargas	National University of Córdoba, Argentina
Dionisio Machado Leite	Federal University of Mato Grosso do Sul, Brazil
Edwin Lughofer	Johannes Kepler University Linz, Germany
Alberto Cano	Virginia Commonwealth University, USA
Andrew Kusiak	The University of Iowa, USA
Wilfried Uhring	University of Strasbourg, France

Khor Shing Fhan	Universiti Malaysia Perlis (UniMAP), Malaysia
Jeonghwan Gwak	Gwangju Institute of Science and Technology, South Korea
Ashok Prajapati	IEEE Computer Society South-East Michigan, USA
Leszek Borzemski	Wroclaw University of Technology, Poland
Ramesh K. Agarwal	Washington University, USA
Oscar Esparza	Universitat Politècnica de Catalunya, Spain
Meng Xianyong	Zhuhai College of Jilin University, China
Shian-Chang Huang	National Changhua University of Education, Taiwan
Kuniaki Uehara	Kobe University, Japan
Anjali Awasthi	Concordia University, Canada
Guo-Shiang Lin	Da-Yeh University, Taiwan
Zhenguo Gao	Harbin Engineering University, China
Chunjiang Duanmu	Zhejiang Normal University, China
Iyad Al Khatib	Politecnico di Milano, Italy
Fengxiang Qiao	Texas Southern University, USA
Mehdi Ammi	University of Paris-Sud, France
Daniel Thalmann	Nanyang technological University, Singapore
Roberto Llorente	Universitat Politècnica de València, Spain
Lulu Wang	Hefei University of Technology, China
Cuicui Zhang	Tianjin University, China
Abdallah Makhoul	University of Bourgogne Franche-Comté, France
Alain Lambert	University of Paris-Sud, France
Tchangani Ayeley	University of Toulouse III, France
Bahareh Asadi	Islamic Azad university of Tabriz, Iran

International Steering Committee

Hui Lin	Institute of Space and Earth Information Science (ISEIS), The Chinese University of Hong Kong, SAR China
Qingquan Li	Shenzhen University, China
Zongyao Sha	Wuhan University, China
Xicheng Tan	Wuhan University, China
Pengfei Zhang	Institute for Infocomm Research (I^2R), Singapore
Wenzhong Shi	The Hong Kong Polytechnic University, Hong Kong, SAR China
Ismail Rakip Karas	Karabuk University, Turkey
Yonghui Zhang	Central South University, China
Lin-gun Liu	ATL, China
Chung-Neng Huang	National University of Tainan, Taiwan

International Editorial Committee

Fuling Bian	Wuhan University, China
Jing Geng	Beijing Institute of Technology, China
Srikanta Patnaik	SOA University, India
Bo Cheng	Beijing University of Posts and Telecommunications, China
Fangjun Huang	Sun Yat-sen University, China
Rui Sun	Beijing Normal University, China
Ning Zhang	Beijing Union University, China
Jinzhu Gao	University of the Pacific, USA
Wenchen Hu	University of North Dakota, USA

Abstracts of Keynote Speeches

Abstracts of Keynote Speeches

Name: Prof. Hui Lin
The Chinese University of Hong Kong, Hong Kong, China

Position held:
Chen Shupeng Professor of GeoInformation Science, Department of Geography and Resource Management
Director, Institute of Space and Earth Information Science

Research Interests:
Microwave Remote Sensing Image Processing and Analysis
Virtual Geographic Environments (VGE) Spatial Database and Data Mining
Spatially Integrated Humanities and Social Science

Keynote Speech Title:
InSAR Remote Sensing for Urban Infrastructure Health Diagnosis

Abstract. The metropolitan area of Hong Kong is characterized by large reclamations with high density skyscrapers and infrastructure. Any inevitable movement of the infrastructure and built environment may pose a threat to infrastructure health and public safety. The development of InSAR remote sensing technology has shown its potential for the diagnosis of the infrastructure health.

Name: Prof. Shuliang Wang

Beijing Institute of Technology, Beijing, China

Shuliang Wang, Ph.D., a scientist in data science and software engineering, is a professor in Beijing Institute of Technology in China. His research interests include spatial data mining, and software engineering. For his innovatory study of spatial data mining, he was awarded the Fifth Annual Info Sci-Journals Excellence in Research Awards of IGI Global, IEEE Outstanding Contribution Award for Granular Computing, and one of China's National Excellent Doctoral Thesis Prizes.

Guest Editor:
International Journal of Systems Science
International Journal of Data Warehousing and Mining
Lecture Notes in Artificial Intelligence

Keynote Speech Title:
Spatial Data Mining Under Big Data

Abstract. It offers a systematic and practical overview of spatial data mining, which combines computer science and geo-spatial information science, allowing each field to profit from the knowledge and techniques of the other. To address the spatiotemporal specialties of spatial data, the authors introduce the key concepts and algorithms of the data field, cloud model, mining view, and Deren Li methods. The data field method captures the interactions between spatial objects by diffusing the data contribution from a universe of samples to a universe of population, thereby bridging the gap between the data model and the recognition model. The cloud model is a qualitative method that utilizes quantitative numerical characters to bridge the gap between pure data and linguistic concepts. The mining view method discriminates the different requirements by using scale, hierarchy, and granularity in order to uncover the anisotropy of spatial data mining. The Deren Li method performs data preprocessing to prepare it for further knowledge discovery by selecting a weight for iteration in order to clean the observed spatial data as much as possible. In addition to the essential algorithms and techniques, the book provides application examples of spatial data mining in geographic information science and remote sensing. The practical projects include spatiotemporal video data mining for protecting public security, serial image mining on nighttime lights for assessing the severity of the Syrian Crisis, and the applications in the government project 'the Belt and Road Initiatives'.

Name: Prof. Yong Wang

University of Electronic Science and Technology of China, Chengdu, China
East Carolina University, Greenville, USA

Current research activities

- Investigation of scale and scale effect on SAR application to urban target Evaluation of water level variations in reservoirs using In SAR technique Thin cloud removal for Landsat 8 imagery
- Submerged aquatic vegetation (SAV) assessment
- Flooding mapping using geo-spatial datasets in rural area

Keynote Speech Title:

Issues in Applying Geoinformatics and Big-Data as Additional Assessment Tools for Macro-Socioeconomic Development

Abstract. Annual socioeconomic datasets released by governmental agencies at the local, state, and national levels portrait socioeconomic statuses within different levels of political boundaries. The data collection costs labor, time, and money. The collected data may consist of errors. Remote sensors provide constant Earth observation. Remotely sensed datasets are multi-temporal and freely available mostly. The datasets are widely used to assess landuse and land cover (LULC) types changes through time, and the changes intuitively reflect the socioeconomic status and development. Thus, the development of additional assessment tools through analyses of remote sensed data is of great interest. Unfortunately, analyzing both types of datasets, one constantly faces analytical and/or statistical challenges. No matter what an approach is applied, following issues must be considered. Otherwise, one will undoubtly concern the results and decisions/actions made based on the outcomes. The issues include data selection, distributions of selected datasets, data transformation, missingness of data, single or multiple independent variables, sensitivity of results to sample sizes, and finally alternative. In this study, we use socioeconomic development of Chengdu City, China between 1978 and 2014 as an example to address above issues. In particular, areas of the impervious surface and agricultural land are derived using spaceborne multi-temporal Landsat data. The domestic gross productivity (GDP) per person released by the statistic department of the municipal government of Chengdu is selected. Between 1978 and 2014, the area of the impervious surfaces and GDP per person increase approximately exponentially. The area of agricultural decreased. Proper transformation is individually applied so that each dataset varies linearly with time. Due to pervasive cloud cover in Chengdu, areas of the impervious surfaces and agricultural lands cannot be derived annually. The multiple imputation method based on the Monte Carlo Markov chain (MCMC) approach is used. Then, GDP per person as the function of the impervious surface area, and as the function of the impervious surface area and agricultural area are statistically established and assessed. The result is satisfactory in regression analysis and crosstab evaluation. It should be noted that the minimum number of required sample size increase rapidly as the number of independent variables increases. Therefore, the use of one or two LULC types as independent variables is recommended.

Name: Prof. Huada Daniel Ruan

Beijing Normal University, Beijing, China
Hong Kong Baptist University, Hong Kong, China
United International College (UIC), Zhuhai, China

Research interest:

- Synthesis, activation, modification and characterization of nanomaterials, their applications as sorbents, catalysts, medications, pigments, additives in environment, agriculture, chemistry and medicine, and their commercialization
- Applications of modified mineral-waste and organic-waste materials for the removals of heavy metals and toxic organic compounds in relation to environmental remediation
- The characteristics of environmental pollutants relating to human health Environmental auditing and assessment relating to environmental management and evaluation of climate change
- Interactions of soil minerals, heavy metals and microbes in contaminated soil materials and bioremediation of contaminated soils
- Environmental chemistry including water quality; air, water and soil pollution; plant nutrition; sediment chemistry; non-point pollution; eutrophication and heavy metal transport, accumulation and contamination
- Renewable energy with emphasis on bio-fuel and solar energy

Keynote Speech Title:
The Application of Environmental GIS

Abstract. Geographic Information System (GIS) generally fulfils the following applications: mapping, monitoring, modelling, measurement and management for a number of fields including political science, education, health care, real estate, business, urban planning and environmental science. The application of a GIS in environmental science can be drawn in environmental monitoring; risk assessment; watershed, floodplain, wetland and aquifer management; groundwater modelling and contamination tracking; hazardous or toxic facility siting; pollutant distribution and remediation; and simulation of process in urban and natural environment. Fundamental investigation of environmental pollution with case studies related to the application of GIS is addressed, and the development of GIS for environmental research and education is discussed in this study.

Name: Prof. Qiang Gao
Beihang University, Beijing, China

Position held:
Professor in School of Electronic and Information Engineering, Beihang University, Beijing, China

Research Interests:
Wireless Communication; Wireless Networks

Keynote Speech Title:
Outage Performance Analysis and Comparison of Two-Way Relaying Systems

Abstract. Cooperative communication has been an effective method for improving system reliability by utilizing the spatial diversity to combat wireless impairments. However, one-way relaying leads to lower spectrum efficiency because it consumes more resources than conventional direct transmission. Recently, two-way relaying (TWR) has drawn much attention since it can provide spectrally efficient transmission with high reliability.

This talk first compares the outage performance differences between amplify-and-forward (AF) and decode-and-forward (DF) in two-way relaying. It is well known that outage performance differences between AF and DF in one-way relaying are apparently related to the average signal-to-noise ratio (SNR). We reveal that it is the target spectral efficiency rather than SNR that determines the superiority in outage performance of different relaying schemes, i.e. DF outperforms AF in the low target spectral efficiency region and the other way around in the high target spectral efficiency region.

Then we investigate the outage performance of two-way amplify-and-forward relaying over block fading channels. Previous research on TWR has been mainly based on the assumption that the channel quality remains constant for one round of data exchange. However, this assumption does not realistically reflect the actual environment as channel conditions fluctuate over time. Our results show that the outage performance of the TWR-AF system deteriorates over block fading channels compared with that over constant-quality channels. Under block fading channels, the TWR system exhibits the outage floor phenomenon, which is not the case for constant-quality channels.

Name: Prof. Tao Gong

Donghua University, Shanghai, China

Prof. Tao Gong received the MS degree in Pattern Recognition and Intelligent Systems and Ph.D. degree in Computer Science from the Central South University respectively in 2003 and 2007. He is an associate professor of immune computation at Donghua University, China, and he was a visiting scholar at Department of Computer Science and CERIAS, Purdue University, USA. He is the General Editors-in-Chief of the first leading journal Immune Computation in its field, and an editorial board member of some international journals. He is a Life Member of Sigma Xi, The Scientific Research Society, a Vice-Chair of IEEE Computer Society Task Force on Artificial Immune Systems, and Chen Guang Scholar of Shanghai. His research has been supported by National Natural Science Foundation of China, Shanghai Natural Science Foundation, Shanghai Educational Development Foundation and Shanghai Education Committee etc. He has published over 100 papers in referred journals and international conferences, and over 20 books such as Artificial Immune System Based on Normal Model and Its Applications, and Advanced Expert Systems: Principles, Design and Applications etc. His current research interests include computational immunology and immune computation. He is also a committee member of intelligent robots committee and natural computing committee in the Association of Artificial Intelligence of China.

Keynote Speech Title:
Cooperative Immune Computation Against Collaborative Attacks in Cyberspace

Abstract. A security problem of cooperative immunization against collaborative attacks such as Blackhole attacks and wormhole attacks, in the mobile ad hoc networks such as the Worldwide Interoperability for Microwave Access (WiMAX) networks, was discussed. Because of the vulnerabilities of the protocol suites, collaborative attacks in the mobile ad hoc networks can cause more damages than individual attacks. In human immune system, nonselfs (i.e., viruses, bacteria and cancers etc.) can attack human body in a collaborative way and cause diseases in the human body. With the inspiration from the human immune system, a tri-tier cooperative immune model was built to detect and eliminate the collaborative attacks (i.e., nonselfs) in the mobile ad hoc networks. ARM-based Network Simulator (NS2) tests and probability analysis were utilized in the prototype for immune model to analyze and detect the attacks. Experimental results demonstrate the validation and effectiveness of the model proposed by minimizing the collaborative attacks and immunizing the mobile ad hoc networks.

Name: Prof. Ji Zhang

University of Southern Queensland, Toowoomba, Queensland

Research Interest:
Prof. Ji Zhang is currently working for the University of Southern Queensland (USQ), Australia. He is an Australian Endeavour Fellow, Queensland Fellow and Izaak Walton Killam Fellow (Canada). He received his degree of Ph.D. from the Faculty of Computer Science at Dalhousie University, Canada. Prof. Zhang's research interests in the area of Computer Science include knowledge discovery and data mining (KDD), Big Data analytics, bioinformatics, information privacy and security, and health informatics. He has published over 90 papers, some appearing in top-tier international journals including IEEE Transactions on Dependable and Secure Computing (TDSC), Information Sciences, WWW Journal, Bioinformatics, Knowledge and Information Systems (KAIS), Soft Computing, Journal of Database Management and Journal of Intelligent Information Systems (JIIS) and international conferences such as VLDB, ACM CIKM, ACM SIGKDD, IEEE ICDE, IEEE ICDM, WWW, DASFAA, DEXA and DaWak. Prof. Zhang is the recipient of a number of prestigious grants and awards including International Science Linkages Grants by Australian Academy of Science (2012 & 2010), Australian Endeavor Award (2011), USQ Research Excellence Award (2011), Head of Department Research Award (2011), Queensland International Fellowship (2010), Izaak Walton Killam Scholarship, Killam Trust, Canada (2007–2008) and IEEE ICDM Student Travel Award by Microsoft and IBM, USA (2006). He was the visiting professor of Michigan State University, USA in 2010 and Nanyang Technological University (NTU), Singapore in 2011.

Keynote Speech Title:
A Parallelized Graph Mining Approach for Efficient Fraudulent Phone Call Detection

Abstract. In recent years, fraud is becoming more rampant internationally with the development of modern technology and global communication. Due to the rapid growth in the volume of call logs, the task of fraudulent phone call detection is confronted with Big Data issues in real-world implementations. In this talk, I will present a highly-efficient parallelized graph-mining-based fraudulent phone call detection framework, namely PFrauDetector, which is able to automatically label fraudulent phone numbers with a "fraud" tag, a crucial prerequisite for distinguishing fraudulent phone call numbers from the normal ones. PFrauDetector generates smaller, more manageable sub-networks from the original graph and performs a parallelized weighted HITS algorithm for significant speed acceleration in the graph learning module. It adopts a novel aggregation approach to generate the trust (or experience) value for each phone number (or user) based on their respective local values. We conduct a comprehensive experimental study based on a real dataset collected through an anti-fraud mobile application, Whoscall. The results demonstrate a significantly improved efficiency of our approach compared to FrauDetector and superior performance against other major classifier-based methods.

Name: Prof. Quan Zou
Tianjin University, Tianjin, China

Editorial Board Member of Scientific Report, PLOS ONE
Special issue guest editor for Neurocomputing, Current Proteomics
Organizing Committee Chair of BIIP2015
Special Session Organizer of IJCNN2016

Program Committee member of the CCIB2011 (Special Session on Computational Collective Intelligence in Bioinformatics, during the 3rd International Conference on Computational Collective Intelligence, ICCCI2011 Gdynia, Poland September 21–23, 2011); WAIM2014,2015,2016 (International conference on Web-Age Information Management); FSDK2014(The 11th International Conference on Fuzzy Systems and Knowledge Discovery); APWeb2016
Outstanding Reviewers for Computers in Biology and Medicine (Elsevier, top 10th percentile in terms of the number of reviews completed within two years, 2015.2)
Reviewer of Bioinformatics, Briefings in Bioinformatics, IEEE/ACM Transactions on Computational Biology and Bioinformatics, IEEE Journal of Biomedical and Health Informatics, Scientific Reports, BMC Bioinformatics, PLOS One, Amino Acids, Gene, Neural Networks, Journal of Theoretical Biology, Computers in Biology and Medicine, Computational Biology and Chemistry, Molecular Biology Reports, BioMed Research International, Current Bioinformatics, Protein & Peptide Letters, Computational and Mathematical Methods in Medicine, Frontiers of Computer Science, etc.

Keynote Speech Title:
Computational Prediction of miRNA and miRNA-Disease Relationship

Abstract. MicroRNA is a kind of "star" molecular, and serves as a "director" since it can regulate the expression of protein. In 2006, related works on gene silence won Nobel price, which made miRNA be the hot topic in molecular genetics and bioinformatics. Mining miRNA and targets prediction are two classic topics in computational miRNAnomics. In this talk, we focus on the miRNA mining problems from machine learning views. We point out that the negative data is the key problem for decreasing the False Positive rather than exploring better features. miRNA-disease relationship prediction is another hot topic in recent years. We introduce some novel network methods on calculating miRNA-miRNA similarity, which is the key issue for miRNA-disease relationship prediction.

Name: Dr. Arun Kumar Saraf
Department of Earth Sciences, Indian Institute of Technology Roorkee, India
Research specialization: Geographic Information System (GIS), Remote Sensing & Digital Image Processing

Honours and Awards:

a. INSA – Royal Society, UK Fellowship – 2002
b. INSA – Chinese Academy of Sciences Bilateral Fellowship - 2011
c. National Remote Sensing Award-2001
d. GIS Professional of the Year-2001
e. National Scholarship for Study Abroad 1986, Govt. of India
f. Indo-US S&T Fellowship, 1994–1995
g. Khosla Research Award 1996
h. Khosla Research Prize 1996
i. Khosla Research Prize 1997
j. Excellent Performance Recognition by IITR for the years 2001–2002
k. Excellent Performance Recognition by IITR for the years 2002–2003
l. Excellent Performance Recognition by IITR for the years 2003–2004
m. Excellent Performance Recognition by IITR for the years 2004–2005
n. Best Paper Award in Map Asia 2004 (Beijing, China)
o. Nominated as Scientific Board Member of the International Geoscience Programme (IGCP) Scientific Board of UNESCO and IUGS

Keynote Speech Title:
Geoinformatics in Mapping of Fog-Affected Areas over Northern India and Development of Ion Based Fog Dispersion Technique

Abstract. Fog is a phenomenon that affects the Indo-Gangetic Plains every year during winter season (December – January). This fog is sometimes in the form of radiation fog and other also occurs as a mixture with other gases, known as smog (smoke + fog). There are various factors contributing to the formation of fog, that may be either meteorological, topographical or resulting from pollution. Fog has been mapped for the winter seasons of the years 2002–2016. In these winter seasons, fog affected areas were found to be changing significantly. The net cover of fog during a season varies in space, time intensity and frequency of occurrence. Presently, it is now possible to map and to predict fog formation to some extent. However, so far it has not been possible to disperse fog, though theoretically it has been discussed in literature. In the current work, experiments were conducted to find out the possibility and effectiveness of a negative air ionizer for fog dispersion. The experiments were carried out with fog, dhoop smoke and a mixture of both to generate smog. Two different glass chambers of different sizes were used in a closed room and the impact of air ionizer on dispersion was studied by testing the time taken for dispersion with or without the ionizer. The results show a significant performance with air ionizer indicating the effectiveness of the ion generator, which reduced the time taken for dispersion (in comparison to without ionizer) by about half.

Abstracts of Invited Talks

Abstracts of Invited Talks

Name: Dr. Ismail Rakip Karas

Karabuk University, karabük, Turkey

Research Interests:
GeoInformatics, Geographic Information Systems, GIS, Three Dimensional

Geographic Information Systems (3DGIS), Network Analyses, Software Development for GIS, Web based GIS, Geo-Databases, Spatial Data Structures, Computer Graphics, Computational Geometry, Image Processing, Graph Theory, Location Based Services

Speech Title:
3D Network Analyses Based on Smart Evacuation System for Indoor

Abstract. The number of buildings, which are very tall, complex and located on wider areas, has been increasing in today's modern cities. Having dozens of floors, hundreds of corridors, and rooms, and passages, these buildings are almost like a city in terms of their complexity and number of people accommodated. Due to size and complexity of buildings, there are many new problems to be addressed. Evacuation of the buildings quickly and seamlessly is the leading problem in case of emergency. Fire, power outage, terrorism (explosions, bomb threat, hostage-taking incidents), chemical spills, earthquake, flood, etc., are some of the extraordinary occasions that may be encountered or affect indoors. In such kind of cases, formation of panic, crowd, congestion, crush, unable to reach exit, etc. are frequently encountered.
In this talk, 3D Network Analyses and Interactive Human Navigation System for indoor which consists of three components will be presented. The first component is used to extract the geometrical and 3D topological vector data automatically from architectural raster floor plans. The second component is used for network analysis and simulations. It generates and presents the optimum path in a 3D modeled building, and provides 3D visualization and simulation. And the third component is used to carry out the generation of the guiding expressions and it also provides that information for the mobile devices such as PDA's, laptops etc via Internet.
In addition, an Intelligent Evacuation Model for Smart Buildings will be introduced in this presentation. The model dynamically takes into account environmental (smoke, fire, etc.) and human-induced (age, disability, etc.) factors and generates personalized evacuation route by performing network analysis interactively and in real-time. Intelligent Control Techniques (Feed-Forward Artificial Neural Networks) has been used in the design of the model.

Name: Dr. Huan Yu

Chinese Academy of Sciences, Beijing, China

Research Area:
Intelligent Simulation of Landscape Changes; Remote Sensing Application

Education Backgrounds:

2013 - Working as Associate Professor at Chengdu University of Technology;
2012–2014 Working as post-doctoral scientist at Chengdu University of Technology;
2010–2013 Working as lecturer at Chengdu University of Technology;

Speech Title:
The Distribution Characteristics of Halogen Elements in Soil Based on RS and GIS Methods

Abstract. Soil chemical elements are important parameters for soil origin diagnosis, and are sensitive indicators of human disturbance process. The present study attempts to evaluate the influence from human activities on halogen elements (fluoride and iodine). This study also attempts to seek a route to explore the spatial relationships between human disturbances and halogen elements according to geospatial theories and methods. Moreover, the spatial correlations between element anomalies and human disturbed landscapes are calculated to explore the influence from human activities on halogen elements, thereby determining the specific response mechanism. The study results indicate that landscapes influence halogen elements in diverse ways and that element iodine is closely related with road and mine landscapes. Furthermore, strong relationships exist between fluoride and road landscapes, which suggest that this element is affected by road landscapes significantly. Fluoride and iodine are unrelated with city landscapes, and fluoride is unrelated with mine landscapes. These provide a reference for the research on the interaction mechanism between halogen and environment. Therefore, it can be concluded that a response mechanism exploration of soil element aggregation and human disturbance is practicable according to geospatial theories and methods, which provides a new idea for studying the soil element migration.

Name: Prof. Chong-yi Yuan

Peking University, Beijing, China

Graduated from Department of mathematics, Nanjing University, 1964
Graduated from institute of Mathematics, Chinese Academy of Sciences, 1968
Switch from mathematics to the study of computer software, 1975
2 more years in Canada as a visiting scholar, Toronto University and Waterloo University, 1977–1979
3+ years in Germany as a visiting scholar to learn Petri Nets from Prof. Carl Adam Petri, 4 times in the 80s last century
Left Institute of Mathematics and started teaching in Peking University, Dec. 1992, Department of Computer Science at that time, School of Electronics Engineering and Computer Science now
Two master courses were taught: Petri Nets and Parallel Program Design from 1993 to 2009
Retired 2005
Professor and Ph.D. supervisor, named Chong-yi Yuan, born 1941
4 books: Petri Nets (1989), Petri Net Principles (1998), Principles and Applications of Petri Nets (2005), Petri Net Applications (2013)

Speech Title:
OESPA: Semantic Oriented Theory of Programming

Abstract. Testing is now a necessary step before a program is put to use. Formal semantics, including operational semantics, functional semantics etc., do not help in this regard. OESPA is a new theory that combines syntax and semantics together to allow program verification instead of testing. It consists of 3 parts: OE, operation expression, for programming, SP, semantic predicates, for precise semantics description, a semantic axioms. To compute semantics from OE. Examples are included for illustration.

Contents – Part II

Advanced Geospatial Model and Analysis for Understanding Ecological and Environmental Process

Contents – Part I

Smart City in Resource Management and Sustainable Ecosystem

Ecological and Environmental Data Processing and Management

Advanced Geospatial Model and Analysis for Understanding Ecological and Environmental Process

Spatial-Temporal Evolution Pattern and Future Scenario Analysis of Water Resources Carrying Capacity of Ningbo City

Yanjuan Wu[1,2(✉)], Zhiming Feng[1], and Yanzhao Yang[1]

[1] Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China
wuyj.lls@igsnrr.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. Contradictions between population distribution and socio-economic development has been taken as a major topic regarding to the realization of coordinated growths of resource, society, economy and ecosystem, especially in city region. Water resources carrying capacity (WRCC) can reflect the quantitative relationship between the total regional population, economic scale and the volume of water resources, which can objectively demonstrate the regional water consumption level under a certain scale of population and economy. Indices of WRCC (WCC and WCCI) were applied to analyze the WRCC of Ningbo city, based on the relationship between population size and water resources of Ningbo city. WCC and WCCI can effectively solve the uncertainty among the total regional population, economic scale and the volume of water resources. Because of different trends in counties, there are different WRCC in these counties. Time-series and multi-scales of WRCC in Ningbo city were assessed using geographic information system (GIS) method. Also, the future WRCC and its regional variation of Ningbo city in 2020 and 2030 were analyzed based on the practical situations using scenario analysis method. Results showed that in the presence of water resources volume of Ningbo city fluctuated, affected by water resource endowment in the past 15 years, the entirety of Ningbo city of WRCC was from abundance to surplus, then surplus with relatively high, and when the flow year was from high flow years to normal flow years, then to low flow years. In terms of the differences of the WRCC among counties, the various WRCC grades of prefecture-level counties should need different kinds of strategies during different flow years to reach their stable states because that in average flow years, water resources existed surpluses beyond balance in Cixi (county-level) city and Ningbo city municipal district. In low flow years, critical overload happened in both Cixi city and Ningbo city municipal district. In the context of the year of 2020 and 2030, the overall WRCC of Ningbo city would stay in abundant levels, whereas the contradictive relationship between population and water resources would be fairly remarkable.

Keywords: Water resources · Endowment · Carrying capacity · WCC · WCCI · Ningbo city

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 3–12, 2017.
DOI: 10.1007/978-981-10-3969-0_1

1 Introduction

Resources and environmental carrying capacity assessment is a rapidly growing field where measures of carrying capacity are used within an assessment framework to evaluate and compare, WRCC is one of the major focuses in the research of resources and environmental carrying capacity (Hester and Little 2013; Fang and Liu 2010). There is a close connection between the regional water utilization, the population growth and economic development, so, it is important to critically evaluate how water resources support the population with the economic development (Zhu et al. 2010). The key point of WRCC evaluation is to achieve conclusions by considering multiple factors based on the supply-demand analysis with guarantee level, etc. (Song et al. 2011). Generally, water resources supply-demand relationship mainly contains two aspects research: economy and population, and water resources carrying capacity for population is the direct index for assessing the carrying capacity (Li et al. 2000; Yao et al. 2002).

WRCC had already turned into a fundamental topic in the research of sustainable development and water resources security strategies, which had attracted great attention from academia (Sagoff 1995; Daily and Ehrlich 1996; Yan et al. 2013). Various WRCC evaluation models and approaches had appeared, the categories of WRCC evaluation can be divided into two major, one of the categories built mathematical models to simulate the evolution of each factor, based on the interactive relationships of each factor in the WRCC system include the fuzzy comprehensive evaluation and the principal components analysis (Gong and Jin 2009; Fang and Liu 2010; Zhu et al. 2010; Dou et al. 2015), the other category of methods established evaluation approaches and criteria from the perspective of phenomenon of the WRCC system, e.g., the conventional tendency method and the multi-objective comprehensive analysis method (Xu 1999; Song et al. 2011; Li et al. 2016).

In this paper, WRCC of Ningbo was studied in the context of the WRCC Measures framework, conducted multi-scale and time series analysis, started from systematic evaluation on water resources endowment in Ningbo. Specifically, research content of water resources endowment characteristics included precipitation, surface and ground water resources, and total volume of water resources. Then, WCC and WCCI model were used based on the population-water relationship, which were widely applied for WRCC calculation. The overall WRCC levels of Ningbo city during 2000–2014 were systematically reviewed and discussed. Also, the future WRCC of Ningbo city was predicted using scenario analysis method by setting the comprehensive water consumption criteria per capita, for the years of 2020 and 2030. Also, multi-scale analysis of Ningbo city WRCC and its regional differences, with different level of hydrological flow years, was conducted (Song et al. 2011; Yulianto et al. 2014). The results of this paper would provide holistic view and scientific basis to help policymakers to optimally utilize water resources and socioeconomic sustainable development.

2 Data and Methods

Ningbo is located at the coast of the East China Sea and southeast corner of the Yangtze River Delta, with a typical north subtropical monsoon climate. There are eleven counties in Ningbo city, considering district area and data constrain, we combine the six municipal districts as one, so there are six counties in Ningbo city in this article including Yuyao, Cixi, Fenghua, Ninghai, Xiangshan and Ningbo city district. The city's multi-year average precipitation is 1,517 mm, with the features of unequal precipitation distribution over the year, large inter-annual variation, and alternate high- and low-flow years. At the same time, the terrain with high altitude in the southwest and low altitude in the northeast enhances the spatial-temporal uneven distribution of water resources.

This paper primarily involved water-resource, socio-economic and basic geographic data of Ningbo city during 2000–2014 in two spatial scales, namely, county-level and city-level. Information of water resources endowment was from *the water resources bulletins* of Ningbo city. The permanent residential population data of each country and district in 2000 and 2010 were from the Fifth and Sixth *Population Census data bulletins of Ningbo*. The population projection data of 2020 and 2030 were from the research results of land planning of Ningbo. The basic geographic data were obtained by scanning, registering, proofreading, and revising the standard state of 1:250,000 digital administrative area maps at both provincial- and city-level along with the administrative area map of Zhejiang Province in 2010.

The WRCC reflects the relationship between regional population and water resources, which can be expressed by the population scale that regional water resources could constantly support under the comprehensive water consumption per capita (The national population and family planning commission development planning and information department, 2009), as shown in the following formulas,

$$WCC_i = W_i / Wpc_i \tag{1}$$

$$WCCI_i = Pa_i / WCC_i \tag{2}$$

Where WCC represents water resources carrying capacity (persons or persons/km^2), W represents water resources volume (m^3), W_{pc} represents comprehensive water consumption per capita (m^3/person), i represents one of the counties of the study area. $WCCi$ represents WRCC of county i, W_i represents the water resources volume of county i (m^3), and $WCCI_i$ represents the water resources carrying index of county i.

According to the regional WRCC evaluation criteria, the WRCC of different regions can be divided into three categories, water resources surplus (WCCI < 0.67), population-water balance (0.67 $\leq$ WCCI < 1.33), and water resources overloading (WCCI $\geq$ 1.33). The three categories are further graded into eight levels (Table 1).

This paper used the comprehensive WRCC indexes including WCC and WCCI to conduct evaluation, in order to fully characterize the regional difference and temporal variation of water resources carrying capacities. We chose the years of 2011 (low-flow year), 2012 (high-flow year), and 2014 (normal-flow year) as typical years. We calculated the WRCC of different hydrological years based on the comprehensive water

Table 1. Ningbo WRCC evaluation criteria

Category	Level	WCCI
Water resources surplus	High abundance	<0.33
	Abundance	0.33–0.50
	Surplus	0.50–0.67
Population-water balance	Slight surplus	0.67–1.00
	Critical overloading	1.00–1.33
Population overloading	Light overloading	1.33–2.00
	Moderate overloading	2.00–5.00
	Severe overloading	>5.00

resources volume per capita in the set "current" year, 450 m^3. Besides, this paper took 470 and 550 m^3 per capita as Ningbo's future conditions in 2020 and 2030 to assess the WRCC of Ningbo city, based on the water resources utilization status quo and variation tendency of Ningbo city.

We provided context for measures in the WRCC Measures section showed in Fig. 1. There were two main stages for the WRCC Measure, (1) analyzed on the water resources endowment which included (a) Precipitation (b) surface and ground water resources (c) total volume of water resources. (2) calculated WRCC (Fig. 1), which included (a) Levels of comprehensive water consumption per capita (b) calculation combined with population. To be specific, WRCC could be at multiple scales, included spatial and temporal patterns of WRCC, and spatial patterns indicated the region and all kinds of sections which consisted of the region, while temporal patterns indicated history, presence, and future.

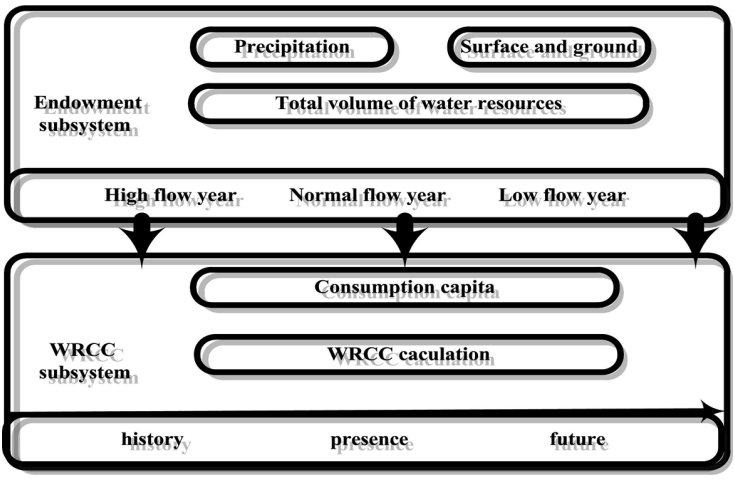


Fig. 1. Schematic of the research process

3 Results

3.1 Analyses of Water Resources Endowment in Ningbo

Precipitation. Figure 2 displayed obvious fluctuations of precipitation during 2000–2014. According to the data from the *Ningbo Water Resources Investigation and Assessment*, the perennial average precipitation of Ningbo was 1,517 mm, average precipitation of the year of 2014 was 1620 mm, which was about 6.8% greater than the perennial average value. The maximum precipitation occurred in 2012 (2,104 mm), which was defined as a high-flow year, and the minimum precipitation happened in 2003 (1,015 mm), making it a low-flow year.

Figure 3 showed the regional distribution of precipitations at Ningbo city in 2014. Precipitations were higher, with relatively large volumes in Yuyao, Ninghai, and

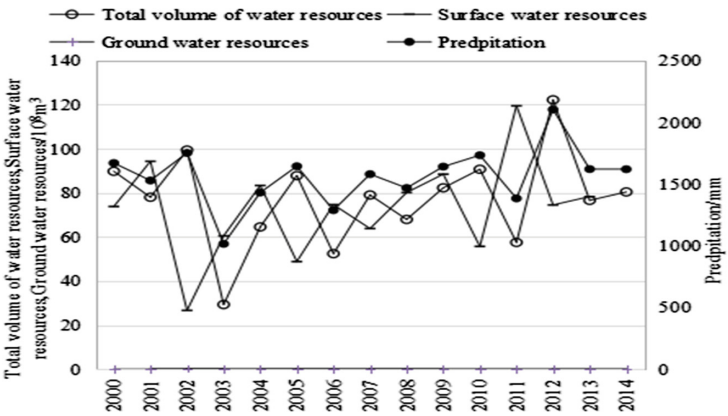


Fig. 2. Water resources endowment of Ningbo city during 2000–2014

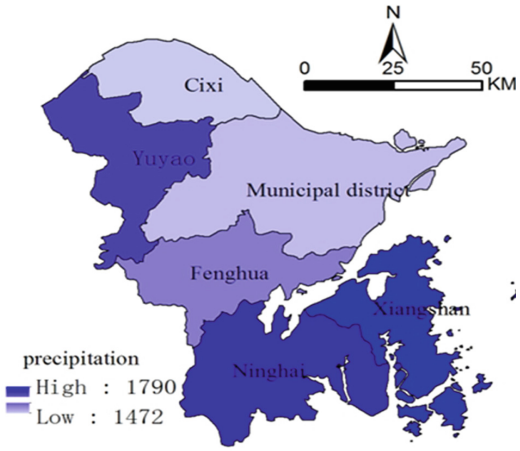


Fig. 3. The spatial pattern of precipitation of Ningbo city in 2014

Xiangshan, than the overall value of Ningbo city, and the precipitation were 1684.6 mm, 1745.6 mm, and 1790.3 mm, respectively. The precipitation was moderate in Fenghua, which was 1583.2 mm, whereas Cixi and the Ningbo municipal district had relatively low values, which were 1472 mm, and 1492.2 mm.

Surface and Ground Water Resources. As was shown on Fig. 2, the fluctuation of surface water volume of Ningbo was almost similar with precipitation during 2000–2014, because that volume of the surface water was affected by precipitation. The volume of surface water was 7.831 billion m^3 in 2014, which was 6.7% higher than the volume of perennial average surface water resources (7.336 billion m^3). The maximum and minimum values of surface water resources occurred in 2012 and 2003, respectively. The maximum, second largest and minimum surface water resources volumes possess were Ningbo city municipal district, Ninghai, and Cixi, respectively.

Figure 2 showed that the ground water volume declined with fluctuations during the period of 2000–2014. In 2014, the ground water resources volume of Ningbo was 2.03 billion m^3, or 0.21 billion m^3 after a deduction of the part of surface water (about 1.82 billion m^3). To be specific, after deducting the overlapping part of surface water in calculation, the ground water resources volume of Ningbo was 0.422 billion m^3 in 2001, then declining to 0.266 billion m^3 in 2003, after the above declining, it saw a certain rebound rising to 0.446 billion m^3 in 2005, yet the volume declined from 0.437 billion m^3 (2007) to 0.179 billion m^3 (2011), then slightly increased to 0.273 billion m^3 (2012), and finally fluctuated to 0.21 billion m^3 (2014).

Total Volume of Water Resources. As was shown in Figs. 2 and 4, in the past 15 years, the total volume of water resources fluctuated within a large range, which was influenced by the volume of surface water resources. To be specific, the years of 2012 and 2003 experienced the most and least water resources with a volume of 12.22 and 2.931 billion m^3, respectively. In the year of 2014, the total water resources volume of Ningbo city was 8.041 billion m^3, 6.8% higher than the perennial average (7.531 billion m^3). Among the total amount of water resources volume of counties from maximum to minimum, the order were Ningbo municipal district (1.81 billion m^3), Ninghai (1.32 billion m^3); Yuyao (1.264 billion m^3), Xiangshan (1.149 billion m^3), and Fenghua (1.73 billion m^3), and Cixi (0.768 billion m^3).

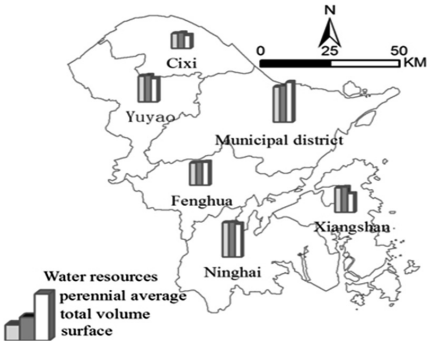


Fig. 4. The spatial pattern of water resources endowment of Ningbo city in 2014

3.2 Evaluation of Water Resources Carrying Capacity of Ningbo City

Previous and Present Evaluation of WRCC. As was shown in Fig. 5, the WCCI was within the range of 0.28–0.97 during various hydrological years, and WRCC was in surplus state overall. In normal-flow year, WCC was about 17.87 million, water resources carrying index was 0.44. In high-flow year, WRCC was in the state of abundance, WCCI was 0.28, WCC was about 27.16 million, substantially higher than that of normal-flow years, or the current actual population of Ningbo city. In low-flow year, WCC of Ningbo city was approximately 27.16 million, significantly lower than that of high-flow years, with WCCI of 0.60, but the WRCC was still in the state of surplus.

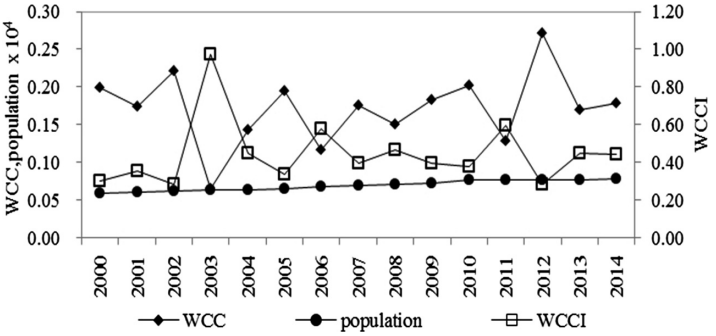


Fig. 5. WRCC of Ningbo city during 2000–2014

WRCC of various types of hydrological years for counties in Ningbo city showed great differences from Table 2. In normal-flow year, WRCC of Cixi and Ningbo city municipal district were relatively low, at the state of slight surplus. WRCC of Yuyao was in the moderate with the level of abundant. WRCC of Fenghua, Xiangshan, and Ninghai **were** relatively high with high abundant level. In high-flow year, the WRCC of each county in Ningbo city was slight surplus overall. In low-flow year, critical overloading occurred in Cixi and Ningbo city municipal district.

Table 2. WRCC of Ningbo city counties of flow years

Flow years	Yuyao	Cixi	Fenghua	Xiangshan	Ninghai	Municipal district
Low-flow	0.50	1.29	0.22	0.30	0.25	1.09
High-flow	0.24	0.52	0.13	0.13	0.12	0.48
Normal-flow	0.34	0.86	0.19	0.18	0.17	0.87

Prediction Results of WRCC in 2020. WRCC would show significant differences during different hydrological year in 2020. If it was normal-flow year, WCCI would be 0.45, WCC would be 17.11 million. If it was high-flow year, WCC would be 26 million, substantially higher than that of normal-flow year, and WCCI would be 0.30, which suggested a state of abundance. If it was low-flow year, WCC and WCCI would be 12.23 million and 0.63, respectively, which indicated the state of surplus.

Table 3 showed differences of WRCC for counties during different hydrological year in 2020. If it was a normal-flow year, critical overloading would occur in Cixi and Ningbo city municipal district. If it was a high-flow year, six counties would be in surplus. If it was a low-flow year, numbers of surplus counties, balance counties, and overloading counties would be 4, 1, and 1, respectively.

Table 3. Prediction of WRCC at Ningbo city counties during flow years in 2020

Flow years	Ningbo city	Yuyao	Cixi	Fenghua	Xiangshan	Ninghai	Municipal district
Low-flow	0.63	0.61	1.58	0.27	0.37	0.31	1.33
High-flow	0.30	0.29	0.64	0.16	0.16	0.15	0.59
Normal-flow	0.45	0.42	1.05	0.24	0.22	0.21	1.06

With the variation of hydrological year type from high flow to normal flow and then to low flow, the number of WRCC surplus counties in Ningbo city would decline from 6 to 3, number of overloading counties would increase to 1, and number of balance counties would increase to 2. Among all these counties, Cixi would be the main area of overloading, while Fenghua, Xiangshan, and Ninghai would be areas with water resources relatively abundant.

Prediction Results of WRCC in 2030. WRCC would be different during different hydrological year at Ningbo city in 2030. If it was normal-flow year, WCC would be 14.62 million and WCCI would be 0.52. If it was a high-flow year, WCC would be 22.22 million, significantly larger than that of normal-flow year, and WCCI would be 0.34, suggesting a state of abundance. If it was low-flow year, WCC and WCCI would be 10.45 million and 0.73, respectively, showing that the population-water relationship would still be slight surplus.

WRCC would be different during hydrological year at Ningbo city in 2030. As was shown in Table 4, if it was normal-flow year, Cixi and Ningbo city municipal district would experience critical overloading of water utilization. If it was high-flow year, the number of counties in surplus would be six. If it was low-flow year, numbers of surplus, balance, and overloading counties would become 3, 2, and 1, respectively.

Table 4. Prediction of WRCC at Ningbo city counties during flow years in 2030

Flow years	Ningbo city	Yuyao	Cixi	Fenghua	Xiangshan	Ninghai	Municipal district
Low-flow	0.73	0.72	1.86	0.32	0.44	0.36	1.58
High-flow	0.34	0.34	0.76	0.19	0.19	0.18	0.70
Normal-flow	0.52	0.50	1.24	0.28	0.26	0.24	1.25

With the variation of hydrological year type from high flow to normal flow and to low flow, the number of WRCC surplus counties in Ningbo city would decline from 6 to 3, number of overloading counties would increase to 2. Cixi and Ningbo city municipal district would be the main overloading areas.

4 Discussion and Conclusions

This paper explicitly expressed the interaction between population and water resources and indirectly discussed the interactive relationship between water resources and socioeconomic development by analyzing water utilization prospects using WCC and WCCI. The regional difference and spatial pattern of WRCC were revealed by quantitatively calculating WCC and WCCI during different horological years. The future population distribution-based WRCC in 2020 and 2030 were systematically predicted using WCC.

It could be seen that the total volume of water resources of Ningbo city fluctuated in a relatively wide range, affected by precipitation and surface water resources volume. And the WRCC study showed that WRCC of Ningbo city had been at the state of water resources surplus primarily during 2000–2014, with obvious differences among various hydrological year types. Yet, the WRCC would gradually declined and water resources utility problems appeared with the variation of hydrological year from high flow to normal flow then to low flow. Also, the WRCC decreased from high abundance to surplus, and even WRCC overloading problems emerged in several counties, results for the differences of precipitation, surface water resources distribution which lead to the water resources distribution discrimination of counties.

From the whole region in Ningbo city, water resource carrying capacity can support 12 million population, even in low flow years, but from the county level, Cixi and Ningbo city Municipal district will be overloading, so the main problem in Ningbo city was the uneven distribution of water resources. So, it is suggested that Ningbo city should coordinate the balance between water resources shortage and socioeconomic development in the future resulted for kinds of situation. To establish and implement comprehensive water resources management program and arrange water resources among regions in Ningbo city would help avoid the WRCC overloading regions problems, such as water diversion in Ningbo city or from other regions. And with the help of decision-making above for Ningbo city, it can also promote water resources sustainable utilization and accelerate the socioeconomic development at city regions.

Sustainable water resources management has become a critical issue for the socioeconomic development of cities that suffer the shortage of water resources. WRCC is common and typical measure to evaluate the state of sustainable water resources, of which, WCC and WCCI model are flexible and explicitly represents time and space. Water use is typically measured as volume, and can include water consumed and water polluted. Considering spatial terms, water use includes internal and virtual uses. Therefore, the assessment of the WRCC needs to focus on water resource-environment-socioeconomic system, which can lead to the forcing and promoting effects of water resources on population and socioeconomically clearly examined (Li et al. 2016; Forokoro and Xie 2011). At the same time, it will also be a detailed research of the mechanisms of environmental self-purification, which can better quantitatively reflect the interactions between population, water resources, and economic development and further answer the scientific problem of how to realize the pursuit of development without degrading the water environment quality (Yao et al. 2002).

References

Hester, E.T., Little, J.C.: Measuring environmental sustainability of water in watersheds. Environ. Sci. Technol. **47**, 8083–8090 (2013)

Fang, C.L., Liu, X.L.: Comprehensive measurement for carrying capacity of resources and environment of city clusters in central China. Chin. Geogra. Sci. **20**(3), 281–288 (2010)

Zhu, Y.H., Drake, S., Lu, H.S., Xia, J.: Analysis of temporal and spatial differences in eco-environmental carrying capacity related to water in the Haihe River Basins, China. Water Resour. Manag. **24**, 1089–1105 (2010)

Song, X.M., Kong, F.Z., Zhan, C.S.: Assessment of water resources carrying capacity in Tianjin city of China. Water Resour. Manag. **25**, 857–873 (2011)

Li, L.J., Guo, H.H., Chen, B., Sun, H.L.: Water resource supporting capacity of Chaidamu Basin. Environ. Sci. **2**, 20–23 (2000)

Yao, Z.J., Wang, J.H., Jiang, D., Chen, C.Y.: Advances in study on regional water resources carrying capacity and research on its theory. Adv. Water Sci. **13**(1), 111–115 (2002)

Daily, G.C., Ehrlich, P.R.: Socioeconomic equity, sustainability, and earth carrying capacity. Ecol. Appl. **6**(4), 991–1001 (1996)

Sagoff, M.: Carrying capacity and ecological economics. Bioscience **45**(9), 610–619 (1995)

Yan, S., Dong, S.C., Li, Z.H., et al.: Carrying capacity of water resources for three-north shelterbelt construction in China. J. Resour. Ecol. **4**(1), 050–055 (2013)

Gong, L., Jin, C.L.: Fuzzy comprehensive evaluation for carrying capacity of regional water resources. Water Resour. Manag. **23**, 2505–2513 (2009)

Dou, M., Ma, J.X., Li, G.Q., Zuo, Q.T.: Measurement and assessment of water resources carrying capacity in Henan Province, China. Water Sci. Eng. **8**(2), 102–113 (2015)

Xu, Z.M.: A scenario based framework for multi-criteria decision analysis in water carrying capacity. J. Glaciol. Geocryol. **21**(2), 99–106 (1999)

Li, N., Yang, H., Wang, L.C., Huang, X.J., Zeng, C.F., Wu, H.: Optimization of industry structure based on water environmental carrying capacity under uncertainty of the Huai Basin within Shandong Province, China. J. Clean. Prod. **112**, 4594–4604 (2016)

Yulianto, S.J.P., Widyawati, N., Kristoko, D.H., Hasiholan, B.S.: Geographic information system for detecting spatial connectivity brown planthopper endemic areas using a combination of triple exponential smoothing-Getis Ord. Comput. Inf. Sci. **7**(4), 21–29 (2014)

Forokoro, K., Xie, Z.: WebGIS to managing natural resource: case of flooded pasture in Lake Débo and Walado Débo. Comput. Inf. Sci. **4**(3), 131–137 (2011)

The national population and family planning commission development planning and information department. Functional partition of population development research, Beijing. World Affairs Press (2009)

Predict Port Throughput Based on Probabilistic Forecast Model

Yihan Chen[1], Zhonghua Jin[2], and Xuejun Liu[1(✉)]

[1] School of Urban Design, Wuhan University,
Donghu Nanlu 8#, Wuhan City, Hubei Province, China
lxj5598@163.com
[2] Department of Urban Planning and Environmental Policy,
Texas Southern University, 3100 Burne St., Houston, TX 77004, USA

Abstract. When the service region of ports overlap, consignors' selecting behaviors for shipping ports become homogeneous to commuters' choosing behaviors on trips. The commuters' travel behaviors can be described through a probabilistic model in transportation planning. In this study, we adopt the transportation probabilistic forecast model to forecast port throughput. First, we amend the model with a port attraction coefficient to forecast port throughput distributions between different ports. Then, forecast for each port throughput is obtained by reallocation of regional total port throughput to each nearby port. We use the port of Fuyang as an empirical research in this paper to validate the methodology. Results compared between this method and traditional regression model indicate that this method provides more persuasive reasoning.

Keywords: Port throughput · Probabilistic forecast model · Port attraction coefficient · Cargo distribution

1 Introduction

The inland waterways freight transport is an economic and environmental friendly transport mode. Development of in land waterway transport, especially for freight transport, not only promotes economic development, but also controls environmental pollution [1]. Port throughput forecast is an important part of shipping development planning. In recent years, with the accelerated process of urbanization in China, the distance between cities along the inland waterways is gradually reduced. Therefore, the distance between inland ports is as well decreased. Overlapping phenomenon appears more and more often between adjacent ports. Cargo shippers in these overlapping areas have more choices. Consignors' selection between ports becomes similar to commuters' choice of route for trips. In transportation research, probability model is usually used to predict commuters' choice for potential travel routes.

In this study, by both drawing from the travel route choice probability model in traffic planning and introducing the port attraction coefficient, we build a shipper-to-port selection model. The throughput of each port is obtained by reallocating the regional total port through to each nearby port.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 13–21, 2017.
DOI: 10.1007/978-981-10-3969-0_2

2 Literature Review

Numerous scholars performed multi-angle studies on the forecast of port cargo throughput using traditional mathematical model, intelligent algorithms, and some other methods [2]. Traditional models perform statistical analysis to resolve the relations between port throughput and a variety of affecting factors by using conventional mathematical methods. In forecasting the Northern Guangxi Port logistics demand, Wang et al. [3] utilized a cubic exponential smoothing method. Chou et al. [4] adopted a modified regression model to forecast the amount of containers imported from Taiwan. de Gooijer and Klein [5] forecasted the incoming steel traffic counts at the port of Antwerp using multivariate time series model.

Due to the extensive use of intelligent algorithms in transportation research, as well as the great enhancement in computing capacities, a variety of intelligent algorithms have been adopted to the forecast of port throughput. Wei et al. [6] used artificial neural network to forecast the number of containers at the port of Kaohsiung. Based on LSSVR, Xie et al. [7] applied mixed model for port throughput forecast. Xiao et al. [8] forecasted port container throughput using Particle Swarm Algorithm. Xu and Wang [9] forecasted the cargo throughput for the port of Qingdao based on TEI@I methodology. Huang et al. [10] as well forecast container throughput for the port of Qingdao using mixed model. Reside on the theory of Markov and Gray forecast model, Zang et al. [11] forecasted waterway freight volume in Chongqing. Linsheng et al. [12] used a combination of multiple linear regression method and the BP neural network to study the Fangcheng port throughput.

Combination of different forecasting methods provides good forecast model for port throughput, but only a few of them involves competition between ports that are geographically adjacent or located in the same region. With the acceleration of urbanization process in China, the scale of urban land is growing. Towns are getting closer, and the distance between ports is decreasing. The rapid development of land transportation, consignors are able to select from more distant ports to ship goods. Thus, competition arouses between geographically adjacent ports. When forecast port throughput, the consignors' selective mind must be taken into account in order to reassign the port cargo throughput effectively. For such situation, Yuan and Xie [13] introduced the selection probability theory to construct a negative exponential model, which was used to forecast port cargo throughput. Liu and Chen [14] further modified the travel mode choice model based on the accessibility theory to improve the forecast accuracy. In our study, we establish a throughput allocation model based on the probability theory to obtain regional total port throughput, and then reallocate it onto each nearby port.

3 Probability Distribution Model

In transportation research, travel time or travel distance is often being utilized as impedance. Usually, commuters choose the shortest route when facing with multiple choices. However, due to the constantly changing traffic conditions, commuters have limited information on traffic. It always results in rather longer route choices. In fact,

there is a higher probability that commuters choose shorter routes. The probability of choosing each possible route can be calculated using LOGIT route choice model. Based on such knowledge, probabilistic route choice model has been constructed for transportation research purposes as follow [15]:

$$P(r,s,k) = \exp[-\theta \cdot t(k)/\bar{t}] \Big/ \sum_{i=1}^{m} \exp[-\theta \cdot t(i)/\bar{t}] \tag{1}$$

Where:

P (r, s, k)	represents the share of transportation mode k from area r to area s;
t(k)	represents the impedance of route k;
$\bar{t}$	represents average impedance of each route;
θ	represents the undetermined coefficient;
m	represents the numbers of valid travel routes.

The shipping cost of goods can be divided into three segments, cost of delivery from the origin point to departure port, cost of shipping from departure port to destination port, and cost of delivery from destination port to final destination. When there are several choices at the point of origin where ports locate relatively close to each other, the cost from departure port to destination port and the cost from destination port to final destination are relative less distinctive. The major distinctive cost of choosing different ports is relaying on the cost of delivery from origin to the departure port. This cost can be treated as traffic impedance for goods transportation to the port. According to function (1), we build port selection model as follow:

$$p_{kj} = \frac{e^{-\theta t(k)/\bar{t}}}{\sum_{i=1}^{n} e^{-\theta t(i)/\bar{t}}} \tag{2}$$

Where:

P_{kj}	represents the probability of consignor in area k choose port j;
t(k)	represents the impedance function of goods transportation;

The impedance function comprises two parameters, the generalized to-port cost (E_{ij}) and the port attractiveness coefficient (A_j) which reflects the port characteristics:

$$t(k) = E_{ij}/A_j \tag{3}$$

E_{ij} represents the generalized to-port cost;

Generalized to-port cost is the cost of transport from the origin point i to port j. It includes cargo transport cost, transfer cost, cost of time, and other indirect costs.

A_j represents the port attractiveness coefficient;

The port attractiveness coefficient reflects the attraction of the port to consignors. It is mainly affected by the accessibility, shipping prices, service levels, and etc. Among all of those factors, accessibility plays a decisive role. Port accessibility can be

measured using average distance of cargo transport to this port, which is calculated as a rate of port turnover over port throughput.

$\bar{t}$ represents average impedance to each port;
θ represents the distribution parameter; In practical application, the average value is between 3.0 and 3.5. We select the value of 3.3 representatively [15];
n represents the number of ports.

4 Case Study

Fuyang City is located in the northwest of Anhui Province. Based on the Fuyang City Master Plan (2012–2030), the city will build two major ports, the Fuyang port and the Yingshang port, and four 4 regular ports, the Taihe port, the Jieshou port, the Linquan port, and the Funan port. From 2011 to 2015, the city's total port throughput is 49,124,400 tons. Comparing to the period between 2006 and 2010, the total throughput grew by 184%. From 2016 to 2020, 2.65 billion yuan has been budgeted to invest in port construction hoping to reach 10 million tonnage increase [16].

Due to the shortage of statistics for Funan port, this study will only study the ports of Fuyang, Taihe, Yingshang, Jieshou, and Linquan. The layout of each port is shown in Fig. 1. All five ports are located relatively close to each other within the administrative divisions of Fuyang. They have a large overlapping hinterlands area. The cargo shipper may choose any port. This is a typical situation where ports compete within one region.

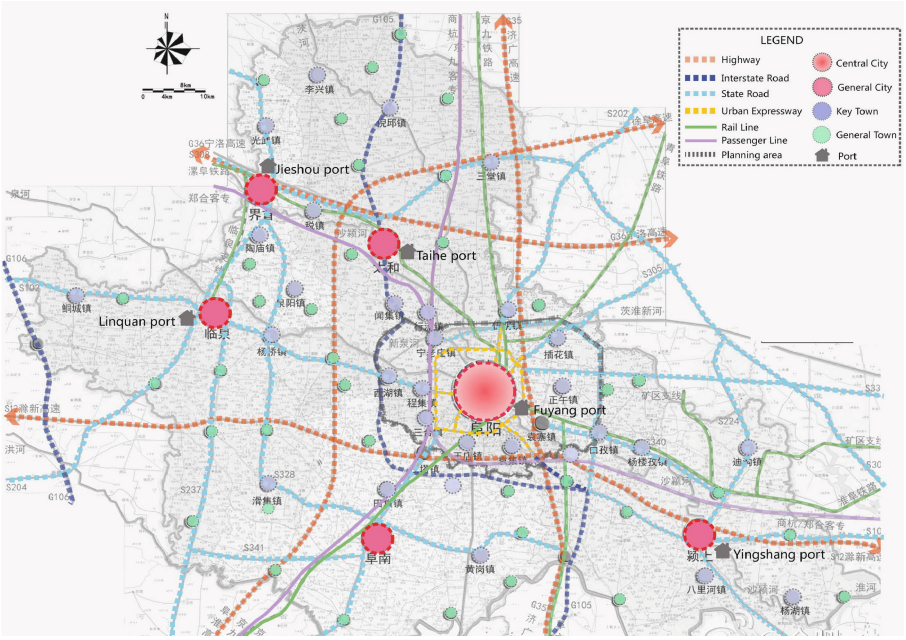


Fig. 1. Layout of Fuyang ports

The cargo throughput in recent years in ports of Fuyang is shown in Table 1.

Table 1. Fuyang City Port cargo throughput statistics (2006–2014) (unit: 10,000 tons) (Source: "Statistical Yearbook of Fuyang City", "China Port Yearbook 2012 Edition", www.soshoo.com)

Year	Total	Fuyang port	Linquan port	Taihe port	Yingshang port	Jieshou port
2006	189	49	17	49	48	26
2007	290	43	20	96	77	55
2008	453	105	18	103	111	114
2009	366	83	23	128	86	46
2010	430	102	33	147	98	50
2011	509	108	47	105	201	48
2012	681	121	61	62	356	46
2013	1020	229	118	151	434	88
2014	1311	417	120	157	522	96

Using the trend analysis, according to the data of Table 1, the relationship between the cargo throughput and time is obtained as follows:

$$y = 122.04x - 21. \tag{4}$$

where

y: cargo throughput, unit: 10,000 tons,
x: year.

By using formula regression analysis above, correlation coefficient is R2 equals to 0.84, indicating that the correlation is very well. Based on this formula, the total throughput of Fuyang port will be 18.09 million tons in 2020. This result is comparable with the 18 million tons prediction in the literature [17] (Fig. 2).

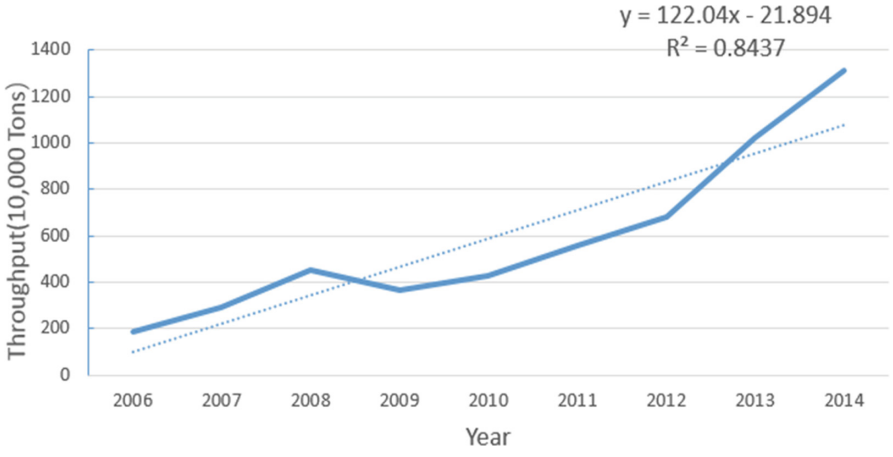


Fig. 2. Port throughput for Fuyang prediction regression analysis chart

The hinterland of Fuyang is relatively small. Most cargos transported by land transportation to the ports. Usually, there is no transfer cost within this area. The generalized to-port cost can be treated as the cost from origin to departure port. In the study region, shipping prices between ports are comparable. The port attractiveness coefficient is obtained from the ratio of the port turnover.

$$Aj = F/T \tag{5}$$

F represents the average freight turnover volume over past years,
T represents the average cargo throughput over past of years.

Table 2 shows the freight turnover volume of each port in Fuyang in the recent years.

Table 2. Statistics of the freight turnover volume for Fuyang port (2006–2014) (unit: 10,000 tons). (Source: "Statistical Yearbook of Fuyang City", www.soshoo.com)

Year	Total	Fuyang port	Linquan port	Taihe port	Yingshang port	Jieshou port
2006	95370	16793	7385	27973	30030	13189
2007	156494	15152	11085	53077	45627	31553
2008	129841	9586	9883	16999	58501	34872
2009	73881	9873	8308	9491	38568	7641
2010	87466	11688	9836	11236	45660	9046
2011	87372.5	9069	9915.5	6093	55613	6682
2012	87279	6450	9995	950	65566	4318
2013	86549	8006	12219	3090	61974	1260
2014	107085	18874	12452	741	74623	395

The attractiveness coefficient of each port is calculated according to Eq. (3) using data in Tables 1 and 2. Table 3 shows the results.

The port cargo throughput in 2020 is calculated through GIS according to function (1). And the throughput of each port is predicted through the traditional linear regression mode. The traditional linear regression of each port are in Table 4. The results of two models are shown in Table 5.

Table 3. Results of the average cargo throughput, freight turnover volume, and attractiveness coefficient of each port. (Unit: 10,000 Tons)

Name	Average throughput	Average freight turnover	Port attractiveness coefficient
Fuyang port	140.10	11721.22	83.66
Linquan port	50.77	10119.83	199.34
Taihe port	110.90	14405.56	129.90
Yingshang port	217.72	52906.89	243.00
Jieshou port	63.22	12106.22	191.48

Table 4. The linear regression of each port

Name	Linear regression function	R^2
Fuyang port	y = 34.829x − 34.042	0.6653
Linquan port	y = 13.585x − 17.156	0.8117
Taihe port	y = 8.1537x + 70.131	0.3367
Yingshang port	y = 59.951x − 82.035	0.8621
Jieshou port	y = 4.0838x + 42.804	0.1489

Table 5. The port cargo throughput in 2020 based on two model (Unit: 10,000 Tons)

Name	Fuyang port	Linquan port	Taihe port	Yingshang port	Jieshou port	Total
Throughput by Eq. (1)	520	198	220	750	181	1869
Throughput by Linear regression function	488	187	190	817	104	1786
Error value	0.06	0.06	0.14	-0.09	0.43	

5 Discussion

(1) There are many factors that affect the attraction of the ports, and the accessibility plays a major role. The port of Fuyang and the port of Yingshang are located in the central area. They are well connected with the industrial land use and storage land use. The two ports are therefore more attractive to cargos.

(2) The turnover of the port reflects its shipping scope. The greater the turnover, the greater the attraction. The port of Yingshang has the largest turnover mileage. Although its location is less favorable than the port of Fuyang, it is the most attractive port.

(3) There are differences between the results predicted by the probability distribution model and the results forecasted by traditional linear regression model. Overall, the smaller the R2 value of the linear regression model, the greater the difference. It indirectly proved that the probability distribution model is more reliable.

(4) By analyzing the port throughput at different locations, the port attraction coefficient and port accessibility are found to be the most influential factors to the throughput. This finding is in accordance with the basic laws of transportation. Also, these two factors can be used as important basis for the port site selection and construction planning.

(5) In this study, the determination of the impedance coefficient is relatively simple. It is only based on road length as the basis for calculation instead of using cost of travel time in land transportation. The determination of the port attraction ignores the evaluation criteria of the port service and the port infrastructure. The two parameters can be further analyzed in follow-up studies.

6 Conclusion

In a certain area, the behavior of cargo shippers' selection of a port is similar to that of the commuters' choice of a trip route. In this study, we adopted the probability choice model from transportation planning to construct a distribution model for port throughput. We then applied this model to forecast the throughput of Fuyang port. Results are compared with the traditional regression model. It indicates that our model provides more explanatory logistics and more reasonable results. This study can as well provide reference for ongoing and future ports planning and construction.

Acknowledgement. This research is funded by the Natural Science Foundation of Hubei [grant number 2014CFB709] and the National Natural Science Foundation of China [grant number 51579182].

References

1. Chou, M.T., Lee, H.S., Lin, K.: A study of forecasting the volume of trans and the harbor operation for port of Kaohsiung. J. Marit. Sci. **12**(2), 235–250 (2003)
2. Baumont, C., Ertur, C., Gallo, J.: Spatial analysis of employment and population density: the case of the agglomeration of Dijon 1999. Geogr. Anal. **36**(2), 146–176 (2004)
3. Wang, J.-M., Zhu, F.-Y., Sui, B.-W., Jiang, Z.-J.: Research on demand forecasting and development pattern of port logistics for Guangxi beibu gulf port. Logist. Sci-Tech **12**, 26–28 (2010)
4. Chou, C.-C., Chu, C.-W., Liang, G.-S.: A modified regression model for forecasting the volumes of Taiwan's import containers. Math. Comput. Model. **47**(9), 797–807 (2008)
5. de Gooijer, J.G., Klein, A.: Forecasting the Antwerp maritime steel traffic flow: a case study. J. Forecast. **8**(4), 381–398 (1989)
6. Wei, C.H., Yang, Y.C.: A study on transit containers forecast in Kaohsiung port: applying artificial neural networks to evaluating input variables. J. Chin. Inst. Transp. **11**(3), 1–20 (1999)
7. Xie, G., Wang, S., Zhao, Y., Lai, K.K.: Hybrid approaches based on LSSVR model for container throughput forecasting: a comparative study. Appl. Soft Comput. **13**(5), 2232–2241 (2013)
8. Xiao, J., Xiao, Y., Fu, J., Lai, K.K.: A transfer forecasting model for container throughput guided by discrete PSO. J. Syst. Sci. Complex. **27**(1), 181–192 (2014)
9. Xu, L., Wang, S.: Analysis and forecasting of port logistics based on TEI@I methodology. J. Transp. Syst. Eng. Inf. Technol. **12**(1), 173–179 (2012)
10. Huang, A., Lai, K., Li, Y., Wang, S.: Forecasting container throughput of Qingdao port with a hybrid model. J. Syst. Sci. Complex. **28**(1), 105–121 (2015)
11. Zang, W.Y., et al.: Freight volume prediction for Chongqing water transport based on gray Markov. Port & Waterw. Eng. **462**(1), 30–33 (2012)
12. Linsheng, F., Jianxin, D., Hao, M., Qi, Z.: Throughput forecasting of fangcheng port based on TEI@I methodology. Logist. Technol. **34**(10), 75–79 (2015)
13. Yuan, H., Xie, Y.: On port's throughput probabmty forecasting model. Port & Waterw. Eng. **16**(4), 28–30 (2007)
14. Liu, X., Chen, Y.: A modified probabilistic forecast model of port throughput based on accessibility. Electron. J. Geotech. Eng. **07**(21), 4845–4854 (2016)

15. Wei, W., Jiqian, X., Tao, Y.: Urban Traffic Planning and It's Application. Southeast University Press, Nanjing (1998)
16. Paper, Fuyang Evening: The 13th Five-Year Plan of Fuyang is Formulated. In ed. (2016)
17. Shang, J.: Research of Demand Analysis and Forecast for Comprehensive Transportation Hub. Southwest Jiaotong University, Chengdu (2014)

Principal Component Analysis of Building Cluster Factors

Hua Ai[1], Qiang Liu[2(✉)], Zhen Wang[2], Zezhong Zheng[2], Yaosen Huang[2], and Zhiqin Huang[3]

[1] Neijiang Normal University, No. 705, Dongtong Road, Neijiang 641100, Sichuan, People's Republic of China
[2] School of Resources and Environment, University of Electronic Science and Technology of China, No. 2006, Xiyuan Avenue, Wust Hi-Tech Zone, Chengdu 611731, Sichuan, People's Republic of China
liuqiang_em@sina.com
[3] Department of Land and Resources of Sichuan Province, No. 4, Baihui Road, Chengdu 610072, Sichuan, People's Republic of China

Abstract. Building properties on a map can be represented by multiple building characterization factors. In this paper, using principal component analysis method, we analyzed multiple factors characterizing buildings. Also, through dimensionality reduction transformation into a small amount of comprehensive factors, this paper proposed simplified expression of building properties, to better meet the need of map generalization for buildings.

Keywords: Principal component analysis · Map generalization · Building cluster factor

1 Introduction

Research on map generalization methods for buildings, the core element on large scale base maps, is the focus and key point in the research field of map generalization. Based on the multi-constraint building group clustering method, aiming to hierarchical clustering, Qianhu [1] presented global and local building cluster constraints is proposed for global and local constraints on buildings clustering, range, and evaluated priority of orientation, similarity. Using Delaunay triangles and Voronoi maps, Tinghua and Xiang [2] put forward distributed analysis model for building cluster by computing some variables of cluster structure, such as distribution density, topology neighborhood, adjacent distance and adjacent direction. In her intelligent building clustering research, Boyan et al. [3] depicted buildings by integrating a series of relevant parameters, such as centroid, spacing, location relationship between features. Overall, although the above methods used more or less some characteristics factors of buildings, a systematic and comprehensive study on building factors is defective.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 22–26, 2017.
DOI: 10.1007/978-981-10-3969-0_3

2 Definition of Building Factors

According to understanding of building characteristics on the map, this paper presented 9 building characteristics factors, which were applied to extract building characteristics on the map.

(1) building area factor

It refers to the area of a single building polygon.

(2) contraction factor

It refers to the reciprocal of the shape factor of a building.

(3) density factor

It refers to area ratio, which is equivalent to the ratio of a total area of buildings to the circular area in a circle.

(4) fractal dimension factor

It represents complexity of building polygons, and the formula for calculating the fractal dimension factor is shown in formula 1, where l represents the long side of building, and S is the area of building:

$$\text{Fract} = \frac{2.0 \times \log_e^1}{\log_S^e}. \tag{1}$$

(5) factor of ratio of length to width

It refers to the ratio of the short side to the long side of the minimum bounding box of a building, as shown in Fig. 1(a).

(6) minimum bounding box area factor

It is the product of length and width of the minimum bounding box.

(7) distance factor from the adjacent road

It refers to the distance to the nearest road, as shown in Fig. 1(b).

(8) direction factor

It refers to the orientation of building, which is defined as the angle between the minimum axis and the horizontal axis of the minimum bounding box of a building, as shown in Fig. 1(c).

(9) shape factor

It refers to the flat rate of building polygons. The smaller the shape factor of the building is, the more flat the building is, as shown in Fig. 1(d).

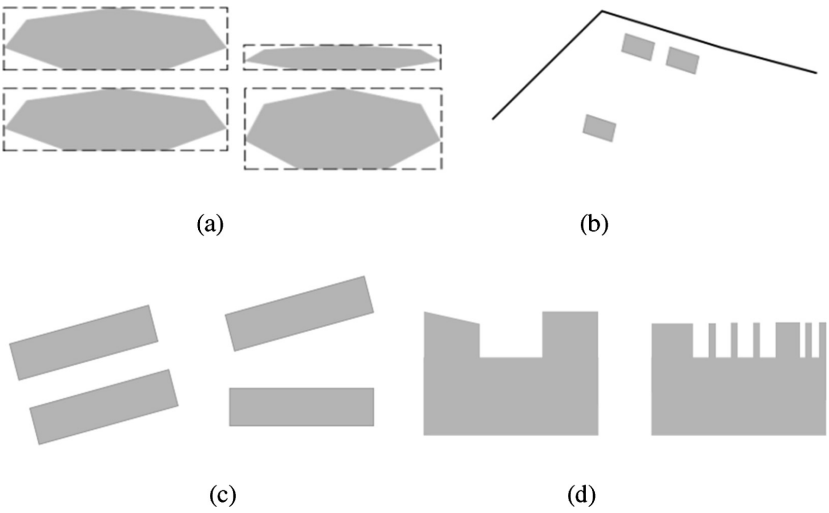


Fig. 1. Some factors. (a) factor of ratio of length to width, (b) distance factor, (c) direction factor, (d) shape factor.

3 Principal Component Analysis

There are not only multiple building clustering factors, but also the coupling relationships among these factors. Using principal component analysis (PCA) can merge multiple factors into several independent comprehensive factors, to reduce the mutual interference among factors [4], and to be able to extract required information quickly from linear combination of these indicators. Principal component analysis in multivariate analysis is to study how to replace the original variable with fewer comprehensive factors. The essence of principal component analysis is to describe a thing with fewer description indicators. Secondly, it can also be used to arrange the weight of multiple indicators [5].

4 Analysis of Building Cluster Comprehensive Factors

Firstly, principal components were determined by solving the correlation coefficient matrix. And on the basis of this, we can establish a relatively complete system of comprehensive factors which can reflect building characteristics.

As shown in Table 1, the correlation coefficient matrix of the data obtained from the 9 factor variables on a certain urban map.

After constructing the correlation coefficient matrix, the initial eigenvalue and the total variance contribution are obtained by matrix computation, as shown in Table 2. The cumulative contribution rate of the first 5 factors can reach more than 89%. By the factor load matrix (component matrix) in Table 3, the contribution rate of the 5 factors to the 9 different building factors can be seen. Based on the results of Table 2, the 4

Table 1. Building factor correlation coefficient matrix.

	area	cont	dens	frac	ltwt	bxar	dist	dire	shap
area	1.00	−0.33	0.63	−0.56	0.08	0.98	0.04	−0.14	0.34
cont	−0.33	1.00	−0.16	0.11	0.50	−0.39	−0.01	0.05	−0.98
dens	0.63	−0.16	1.00	−0.49	0.13	0.61	0.13	−0.13	0.17
frac	−0.56	0.11	−0.49	1.00	−0.43	−0.53	−0.20	0.37	−0.13
ltwt	0.08	0.50	0.13	−0.43	1.00	0.07	0.03	−0.20	−0.44
bxar	0.98	−0.39	0.61	−0.53	0.07	1.00	0.03	−0.13	0.41
dist	0.04	−0.01	0.13	−0.20	0.03	0.03	1.00	−0.12	0.02
dire	−0.14	0.05	−0.13	0.37	−0.20	−0.13	−0.12	1.00	−0.04
shap	0.34	−0.98	0.17	−0.13	−0.44	0.41	0.02	−0.04	1.00

Table 2. Initial eigenvalues and total variance contribution.

Component	Eigenvalue	Variance (%)	Cumulation (%)
a1	3.855	38.554	38.554
a2	2.313	23.131	61.684
a3	1.160	11.603	73.288
a4	0.955	9.547	82.835
a5	0.687	6.865	89.701
a6	0.452	4.517	94.218
a7	0.351	3.510	97.727
a8	0.203	2.033	99.760
a9	0.007	0.073	100.00

Table 3. Factor load matrix

	a1	a2	a3	a4	a5
area	0.826	0.330	0.346	0.053	−0.106
cont	−0.730	0.650	0.124	0.030	−0.037
dens	0.620	0.414	0.249	0.219	−0.169
frac	−0.595	−0.587	0.211	0.064	−0.144
ltwt	−0.074	0.779	−0.164	−0.282	0.453
bxar	0.855	0.266	0.327	0.036	−0.063
dist	0.116	0.165	−0.514	0.817	0.102
dire	−0.252	−0.311	0.658	0.318	0.543
shap	0.735	−0.614	−0.123	−0.031	0.071

factors with the lowest contribution rate are excluded, and then the first five factors with higher overall contribution rate in factor analysis are retained.

However, the resulting factors is not the principal components with all information, but the eigenvalue of the matrix. Variable computation to the 5 factors is still necessary. The values of these factors, namely the values of the characteristic vectors were standardized.

The 5 factors are computed by summing over the 9 building factor weighted by the eigenvector corresponding to the factor, which can reflect 89% of all building shape factor information. They can be used to represent all the morphological characteristics of the buildings, from the above analysis, PCA method of multi factor can overcome the defects of lack of information to the second clustering of single building factor. Composite factor is composed of the principal components, and it has a comprehensive information characteristic.

Acknowledgments. This work was partially supported by Science Research Program of Land and Resources Department of Sichuan Province (No. KJ201613 and No. KJ20159), and The Project Supported by the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources (No. KF-2016-02-007).

References

1. Qianhu, S.: Research on Clustering Method of Building Cluster Based on Multiple Constraints. Central South University, Changsha (2011)
2. Tinghua, A., Xiang, Z.: The aggregation of urban building clusters based on the skeleton partitioning of gap space. In: Fabrikant, S.I., Wachowicz, M. (eds.) The European Information Society. Lecture Notes in Geoinformation and Cartography, pp. 153–170. Springer, Heidelberg (2007)
3. Boyan, C., Qiang, L., Xiaowen, L.: Intelligent building grouping using a self-organizing map. J. Acta Geod. et Cartograph. Sin. **42**(2), 290–294 (2013)
4. Liangjian, W., Wei, L.: Study on driving force of land use change in Wuzhou city. J. Econ. Geogr. **19**(4), 74–79 (1999)
5. Jing, Y., Zhengong, T., Yuzhe, L.: The application of SPSS software to analysis and evaluation of the principal components of drinking water quality. J. Environ. Sci. Technol. **07**, 171–174 (2011)

Progressive Network Transmission Method Research of Vector Data

Shengli Wang[1,2], Zezhong Zheng[2(✉)], Chengjun Pu[2], Mingcang Zhu[3], Yong He[4], Zhiqing Huang[5], Yicong Feng[5], Mengge Tian[2], and Jiang Li[6]

[1] Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources, Shenzhen 518040, Guangdong, People's Republic of China
[2] School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu 611731, Sichuan, People's Republic of China
zezhongzheng@uestc.edu.cn
[3] Land and Resources Department of Sichuan Province, Chengdu 610072, Sichuan, People's Republic of China
[4] Sichuan Institute of Geo-Environment Monitoring, Chengdu 610081, Sichuan, People's Republic of China
[5] Information Center, Land and Resources Department of Sichuan Province, Chengdu 610072, Sichuan, People's Republic of China
[6] Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA 23529, USA

Abstract. Vector data contains a lot of important features. Progressive transmission is a key technology to solve the real-time rendering and network transmission of vector data. By studying the traditional progressive transmission method of vector data and considering the spatial position and geometric features of vector data, we proposed an efficient progressive transmission method. We divided the vector data into blocks based on spatial location, then applied a Visvalingam-Whyatt algorithm to build a multi-scale model. Finally the progressive transmission of vector data was achieved. Our method satisfies the viewer's needs to display data from different rendering scale and has important significance for client users to interact in real time.

Keywords: Vector data · Visvalingam-Whyatt · Progressive transmission

---

The Project Supported by the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources (Grant No. KF-2016-02-032, and Grant No. KF-2016-02-036); Science Research Program of Land and Resources Department of Sichuan Province (Grant No. KJ201613 and Grant No. KJ20159); Open Fund of State Key Laboratory of Water Resources and Hydropower Engineering Science (Grant No. 2014SWG04); The Key Technologies Research and Development Program of Hubei Province (Grant No. 2015BCA290); Open Fund of Key Laboratory of Geoscience Spatial Information Technology, Ministry of Land and Resource of the P.R. China (Grant No. KLGSIT201411); Opening Foundation of Guangxi Key Laboratory for Spatial Information and Geomatics, Guangxi, China (Grant No. 140452413, and Grant No. GKN120711516); State Key Laboratory of Remote Sensing Science (Grant No. OFSLRSS 201318); and 2015 Hongkong and Mainland College Students Exchange Program (Grant No. 2014547).

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 27–36, 2017.
DOI: 10.1007/978-981-10-3969-0_4

1 Introduction

Vector data is mainly managed on a single-machine environment. Cloud platform is the development trend of vector data management. Networks transmission, distributed access and 3D rendering of massive data are supporting technologies, but also a great challenge of cloud platform [1]. Therefore, it is of great significance to study the network transmission, distributed access and 3D real-time rendering of massive vector data. The progressive transmission technology is the key to solve these problems. Traditional vector network geographic information system needs to download the client as well as vector data in one-time, which will take some time under the present conditions of limited bandwidth, and gives a bad user experience. Thus, a new vector data transmission thought of progressive transmission network began to emerge and become one of the mainstream and the focus of the present research [2].

Researchers have proposed a number of progressive transmission methods in recent years, i.e. the US Bertolotto and Egenhofer [3], Buttenfield [4], Weibel and Dutton [5] Domestic Bi-sheng and Bi-jun [6]. But methods of progressive transmission of vector data are still not perfect and need to be improved.

Based on the present progressive transmission methods and considering the characteristics of vector data, we proposed an efficient progressive transmission method. The main idea is that the vector data is abstracted to expressions of spatial distribution and geometric characteristics for the study vector data organization. Vector data organization is vector data placement method to facilitate the vector data management and transmission.

For massive vector data, the basic framework to achieve progressive transmission is shown in Fig. 1, and need to address the following four aspects:

(1) Spatial grid division. The problem that each layer data is too large is solved using grid cell.
(2) The geometry simplification algorithm. Vector data consists of the complex shape, multi-node curve. According to the different scales display requirements, the geometric simplification algorithm is used to summarize and select nodes, and simplify the curve.
(3) Multi scale data organization. Reasonable organization of spatial data can facilitate the transmission of vector data. In this paper, spatial grid cells and different levels of spatial data are organized.
(4) The network transmission. Network transmission is the way organized vector data converted into the client. It involves two aspects: the principle of data acquisition and performance optimization.

2 Experimental Data

Experimental data is land use data in Chengdu in 2000. Planar data size is 22.2 M, and linear data size is 11.3 M. Chengdu lies in between east longitude 102°54′–104°53′ and north latitude 30°05′–31°26′.

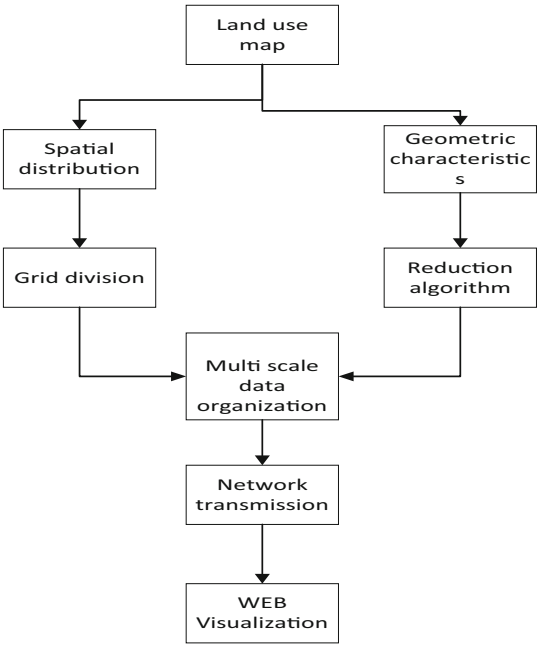


Fig. 1. Massive vector data progressive network transmission frame.

3 Method

In the process of oriented mass vector data transmission, we first select the simplified data to transfer before more detailed data, so that users can get a better experience. Spatial grid, curve simplification algorithm and spatial data organization method are key techniques to solve the transmission of massive vector data.

3.1 Spatial Distribution

From the overall composition of the map, spatial distribution is an important factor in progressive network transmission. Spatial elements contain a large amount of location information, involving a large spatial scope. Therefore, considering the spatial distribution, we segment the map by latitude and longitude drawing on the idea of spatial grid, expressing the elements in grid units.

Grid units are composed by the latitude and longitude lines. The size of line spacing determines the transmission performance of spatial data. The size is 0.3° * 0.3° in this paper. Block size value of planar vector data is averaged 1.29 M, and block size value of linear vector data is averaged 0.58 M. The grid units is shown in Fig. 2.

In this paper, the grid division is carried out according to the latitude and longitude. When the single element is located in a plurality of grids, the element are not divided, and the grid cell to which the feature belongs is determined based on the starting point coordinate value of the feature.

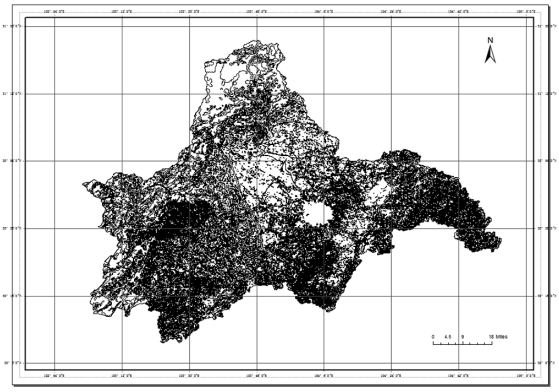


Fig. 2. Map grid of land use data of the Chengdu.

3.2 Geometry Simplification Algorithms

Currently, the popular geometry simplification algorithms are Douglas-Peuker and Visvalingam-Whyatt. The core concept of Douglas-Peuker is to compute the distance from point to line, and the choice of the point depends on the distance and a threshold [7]. Visvalingam-Whyatt algorithm can keep the original shape of the curve, and its core idea is to compute the area near the three nodes constitute the triangle, and the choice of a point depends on the size of the area [8].

In this paper, we use an improved Visvalingam-Whyatt algorithm to simplify land use data, which can be described as.

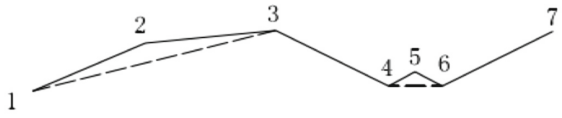
(1) Calculate the curve vertex weights for each point, which is the area size.
(2) Terminate the calculation if the number of vertices on the curve is less than two.
(3) Select the vertex of the minimum weight value.

When the vertex with the smallest weight is smaller than the specified threshold, delete it, return to the second step and continue judge until the weights of the remaining vertices are greater than the threshold. The process is shown in Fig. 3.

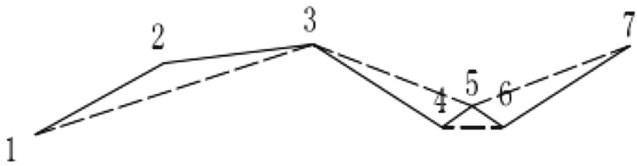
In our approach, land use data is divided into five layers, and vertices of each level is obtained within a predetermined range the area values on the basis of Visvalingam-Whyatt algorithm obtained, in other words it is an incremental value. Improved Visvalingam-Whyatt can be described as:

(1) Calculate the curve vertex weights for each point, which is the area size.
(2) Set area threshold range of each level;
(3) Put vertices into different level according to weight value.

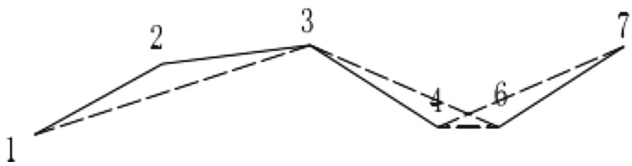
After the actual test, data size of each layer is listed in Tables 1 and 2. In the current environment, network transmission is able to satisfy user's interaction experience.



A. Before treatment



B. Processing



C. After treatment

Fig. 3. Visvalingam-Whyatt algorithm.

Table 1. Data unit size value (planar data).

Level	Minimum value (byte)	Maximum value (byte)
1st-level	97.7	405.0
2nd-level	88.2	423.0
3rd-level	101.3	396.9
4th-level	86.2	401.1
5th-level	83.0	399.5

Table 2. Data unit size value (linear data).

Level	Minimum value (byte)	Maximum value (byte)
1st-level	97.7	405.0
2nd-level	88.2	423.0
3rd-level	101.3	396.9
4th-level	86.2	401.1
5th-level	83.0	399.5

3.3 The Organization of Multi-scale Data

After multi-scale construction of spatial data, it has been able to meet the needs of progressive transmission, but we still need a reasonable spatial data organization method. In this paper, we store space data in the form of files. The basic idea is as follow:

(1) Partition vector data according to spatial grid;
(2) Divide data within a single grid unit into five levels. The first level stores simplified vector data. The second, three, four, five level store incremental vector data.
(3) A single data represents the data of a certain level within a certain grid cells. Data storage format is json. The folder structure is: X:\Data\level number\latitude\longitude-latitude.json.

3.4 Network Transmission

The main flow of progressive network transmission is shown in Fig. 4, including two key methods:

(1) Data acquisition principles. When the user requests data from the client, we need to obtain the current user's visual angle range and perspective height, and determine which level to return within spatial grid cells based on the eye alt. We transport rough data from the server side to the client, and then a more detailed data is transported [9].
(2) Performance optimization. To prevent the delay of loading data on the client, we pre-download some data that is outside the boundary of the view range. When the map is moved outside the current view range, it can be directly loaded locally. We purge the loaded data which is not in the current range and proximity to save memory.

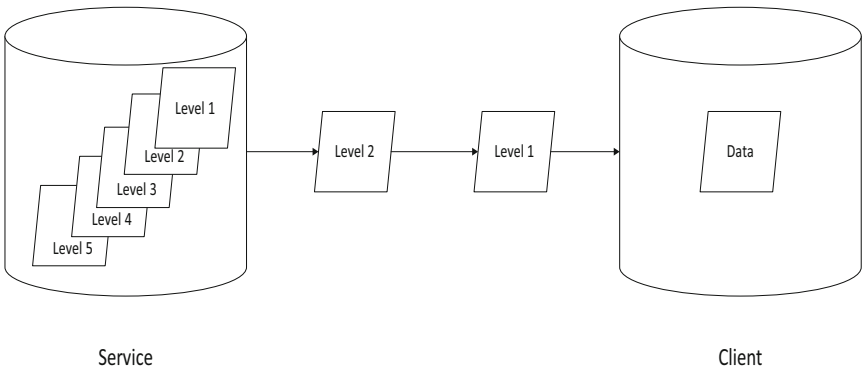


Fig. 4. Land use data progressive network transmission process.

4 Implementation of Progressive Network Transmission Method

Progressive network transport system consists of server-side, client-side and vector data. The key techniques of progressive network transmission were discussed earlier in the paper. We will describe the implementation of network transmission of vector data in the following part.

4.1 System Instance

In this paper, multi-scale vector data is constructed by OGR library, and vector data is reconstructed by World Wind in the client. Client interface is shown in Fig. 5.

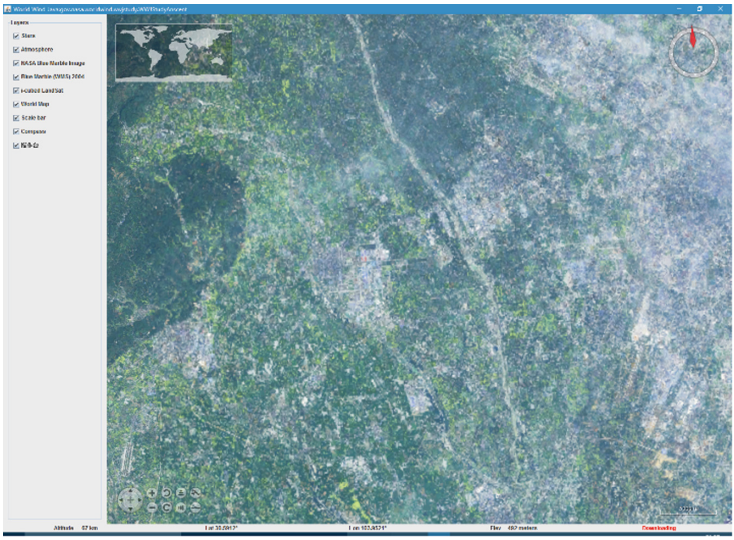


Fig. 5. Client interface.

4.2 Experiment Analysis

(1) Land use data reduction

The data used to be simplified is land use data. The number of elements is 94,700, and contains 1,803,145 points. The vertices reduced to 309,530 after simplification, as shown in Fig. 6.

(2) Vector data transmission progressive effect

In the process of progressive transmission, we select the land use data in Chengdu and it is divided into five levels. Figure 7 shows the display effects of some local regions in the data transmission process in five levels. It can be seen that there are increasing details in the five levels from Fig. 7. When a user requests data, the rough

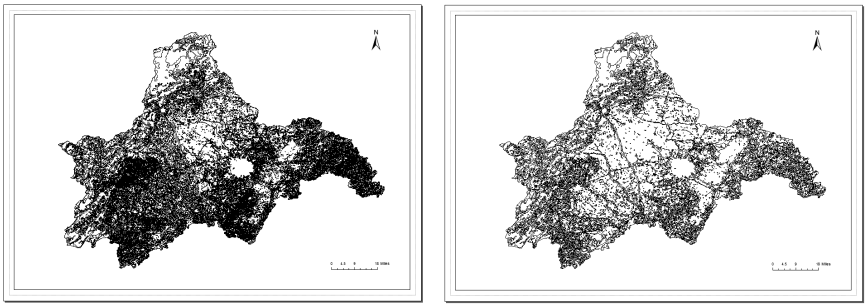


Fig. 6. Land use graph and its simplification effect.

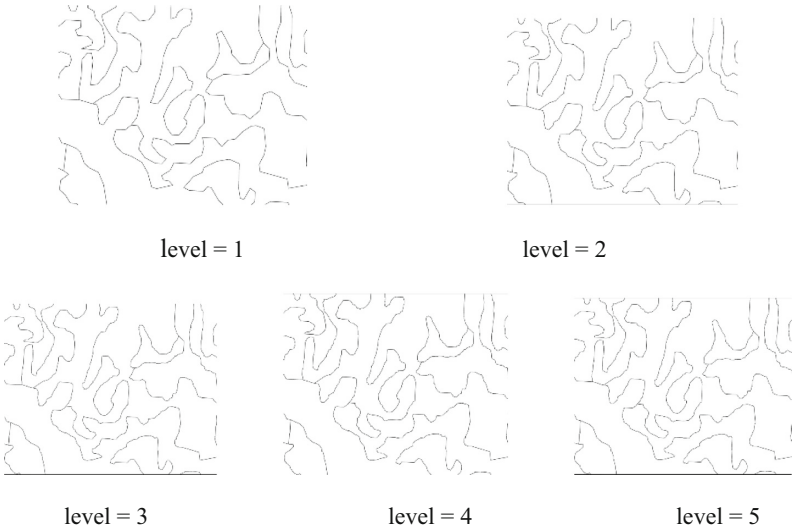


Fig. 7. A display effect of drawing land use data transmission.

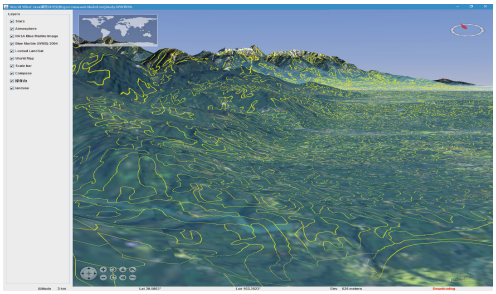


Fig. 8. Displayed renderings of a land use data in 3D scene.

data is transferred from the server-side first, and then more detailed data was transferred. Figure 8 is land use data displayed in 3D scene.

(3) Transmission speed comparison

It can be seen from Table 3 that transfer rates of line and area data have little difference. Both transmissions are fast enough for users to browse quickly. But it has an unresponsive phenomenon in the process of browsing. It remains to be further studied in future research work.

Table 3. Comparison of land use data transmission speed

Land use data	Size (MB)	Average speed (kb/s)
Line	11.3	365.3
Polygon	22.2	361.9

5 Conclusion

With the rapid development of computer technology, network has penetrated into various fields. The bandwidth of the network is an important constraint of its development. It has become a research hotspot in the field of network geographic information that how to improve the transmission speed of vector data under the limited bandwidth conditions. In this paper, we propose a grid-based, hierarchical and progressive network transmission method, which improves the efficiency of network transmission, and the burden of computer memory is loosed through memory optimization. But a more in-depth study on issues for the vector data networks progressive transmission is needed, which can be summarized as two points: algorithms for vector data simplification and storage methods of vector data.

References

1. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J. Mol. Biol. **147**, 195–197 (1981)
2. Huang, B., Jiang, B., Li, H.: An integration of GIS, virtual reality and the internet for visualization, analysis and exploration of spatial data. Int. J. Geogr. Inf. Sci. **15**(5), 439–456 (2001)
3. Bertolotto, M., Egenhofer, M.J.: Progressive transmission of vector map data over the World Wide Web. GeoInformatica **5**(4), 345–373 (2001)
4. Buttenfield, B.P.: Transmitting vector geospatial data across the internet. In: Egenhofer, M.J., Mark, D.M. (eds.) GIScience 2002. LNCS, vol. 2478, pp. 51–64. Springer, Heidelberg (2002). doi:10.1007/3-540-45799-2_4
5. Weibel, R., Dutton, G.: Generalising spatial data and dealing with multiple representations. Geogr. Inf. Syst. **1**, 125–155 (1999)
6. Bi-sheng, Y., Bi-jun, L.: State-of-the-art of the progressive transmission of spatial data over the internet. J. Image Graph. **6**, 006 (2009)

7. Taylor, G.: Line simplification algorithms (2005). Accessed 15 Apr 2005
8. Kolesnikov, A.: Vector maps compression for progressive transmission. In: 2nd International Conference on Digital Information Management, ICDIM 2007, vol. 1, pp. 81–86. IEEE (2007)
9. Lindstrom, P., Pascucci, V.: Terrain simplification simplified: a general framework for view-dependent out-of-core visualization. IEEE Trans. Vis. Comput. Graph. **8**(3), 239–254 (2002)

Comparison of Different Remote Sensing Monitoring Methods for Land-Use Classification in Yunnan Plateau Lake Area

Ce Wang, Shu Gan(✉), Da Yi, and Yang Wu

Kunming University of Science and Technology, Kunming 650093, China
851649146@qq.com

Abstract. Remote sensing image classification is an important technology to get information. At present, different remote sensing monitoring methods has been widely used in region land cover. To improve classification accuracy is the key of remote sensing data processing and application. This paper selects Xingyun Lake that the typical Plateau Lake area of Yunnan province and the surrounding lakeside zone as research area. Based on the 30 TM Landsat remote sensing image of the research area, using supervised classification, BP neural network, and object-oriented classification to compare accuracy of three kinds of classification methods. It was found that development of BP neural network and object-oriented classification training produces more accurate results than supervised training. Object-oriented classification also produced more accurate classification than the BP neural network classification, but did not improve the accuracy significantly. The results will help to promote surface coverage information of remote sensing rapidly extraction and dynamic monitoring in the Yunnan plateau lake, moreover, it has important scientific significance to protect and formulate rationalization.

Keywords: Remote sensing monitoring · Surface coverage · Classification technology · Plateau lakes · Xing Yunhu

1 Introduction

Identifying various features on the land by RS Image Recognition is an important part of RS technology development, and Classification of RS image is extremely important for thematic information extraction, dynamically change monitoring, thematic cartography and construction of Remote Sensing Database. The initial remote sensing classification is through visual interpretation. However, visual interpretation is limited to a single band or a three-band (RGB) color composite, which makes the result of the classification with strong subjective. Therefore, automatic classification of remote sensing is more suitable for mapping land-use in a large area. While land-use and

Thanks to the support of National Natural Science Foundation of China (No. 41561083, No. 41261092) and Natural Science Foundation of Yunnan Province (2015FA016).

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 37–42, 2017.
DOI: 10.1007/978-981-10-3969-0_5

land-cover patterns may be obvious to an image interpreter, automatically mapping them could be difficult because automated classification techniques do not possess the superior pattern recognition capabilities of the human brain, so that it is challenging to achieve an accurate classification. Domestic and foreign scholars have been tirelessly engaged in it, With the generation of a series of new classification algorithms, such as object oriented, BP neural network system, decision tree, support vector machine, the traditional supervised classification system and non-supervised classification system are obviously no longer meet the precision requirement.

2 Study Area

Located at 102°45′ to 102°48′ east longitude, and 24°17′ to 24°23′ north latitude, the study area is one of the typical plateau lakes in Yunnan, which plays an important role in the production and life of local residents. In light of continuous land-use changes in Xingyun Lake and lack of regard to regional environment of the plateau lake, there is a need for creating a current land-use information database (Fig. 1). Based on related research achievements to choose more quickly and efficiently classification methods have become inevitable.

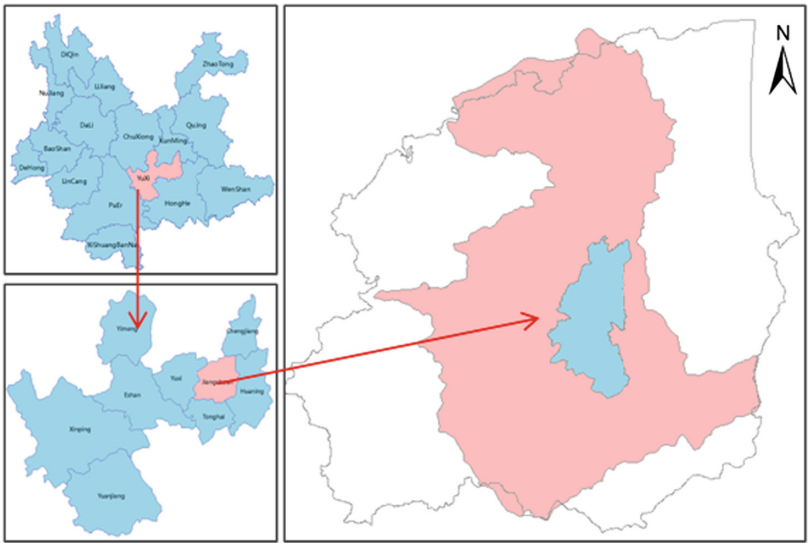


Fig. 1. Location of the study area

3 Methodology

The main research work of this paper is to analyze and compare the accuracy of the different remote sensing classification monitoring technology on multi-scheme experimental (Fig. 2). Several established methods for surface cover classification from RS

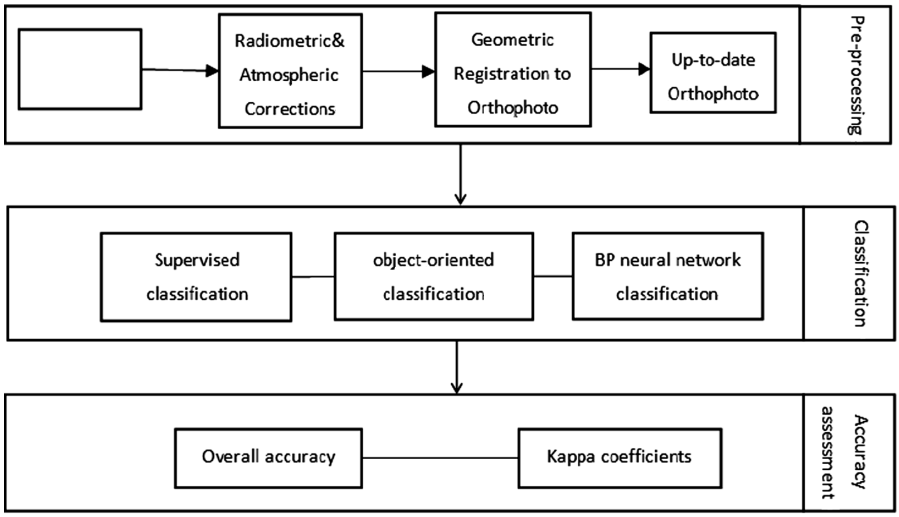


Fig. 2. Research flowchart

data were compared, it include traditional supervised classification, development of BP neural network classification and object-oriented classification. The surface cover classification technique presented in this work can be used to produce information pertaining changes in land-uses, dynamic monitoring of water resources and land exploitation and protection. The information could be further used to study the relations between rapid extraction and dynamic monitoring of land cover information by remote sensing and resource and environmental protection management of the lake region in Yunnan.

Specific objectives of the current study are: 1. After image pre-processing, to compare supervised, object oriented and BP neural network surface cover classification techniques; 2. For each map, a confusion matrix was created and accuracy measures were calculated. Compared it and choose the more feasible and reliable method apply to the remote sensing monitoring in Yunnan plateau lake area.

4 Results and Discussion

The primary source for surface cover classification is a Landsat-5 TM image acquired on 5-February-2014. The selected area appears cloud free. Pre-processing of the image included one-step radiometric and atmospheric corrections using the dark-object subtraction method and the latest radiometric calibration coefficients published. For this study. Level 1 of the Anderson classification system was used. The different land-covers included Urban or built-up land, Agricultural fields, Rangeland, Forest, Water bodies and Barren land in this study. A Landsat TM image was classified and post-classification processed using three methods: supervised, BP neural network and object-oriented classification methods. The different results are presented in Fig. 3.

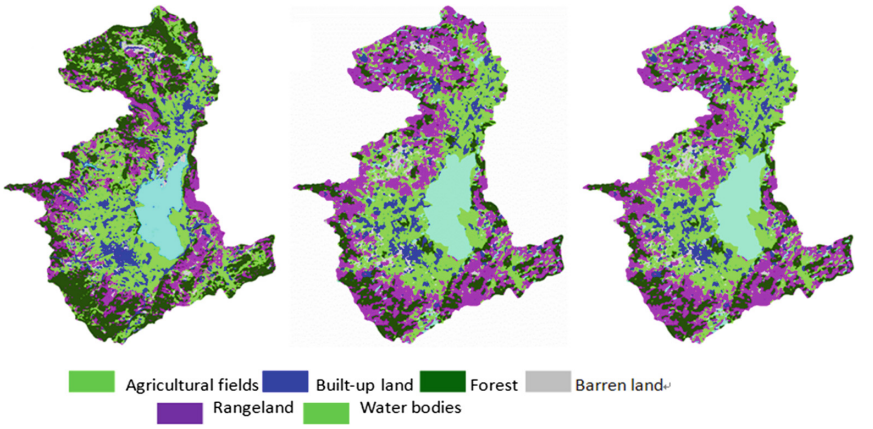


Fig. 3. The results of different classification methods

Following this, the classification products' accuracy was assessed. A comparison of the products' accuracy was conducted to find out if the accuracy differences are statistically significant (Table 1). It was found that development of BP neural network and object-oriented classification training produces more accurate results than supervised training. Object-oriented classification also produced more accurate classification than the BP neural network classification, but did not improve the accuracy significantly. On the whole, judging by the overall accuracy and overall kappa statistics, it is apparent that supervised classification's overall accuracy is 78.03%, Kappa statistic is 0.73; BP neural network classification's overall accuracy is 88.7043%, Kappa statistic is 0.8572; object-oriented classification's overall accuracy is 92.7551%, Kappa statistic is 0.9060.

Table 1. Accuracy comparison of classification methods

Classification	Overall accuracy	Kappa coefficients
Supervised	78.03%	0.73
object-oriented	92.76%	0.91
BP neural network	88.70%	0.86

In order to further reflect the differences of the three kinds of classification methods in remote sensing image, selected a part of study area which is possess of comprehensive land-cover type, and enlarged it (Fig. 4). It was found that there are too many fragment patches and low accuracy in the image of supervised and BP neural network classification, however object-oriented classification has relative regular patches. Especially after supervised, there are obvious errors and leakage classified in the remote sensing image (Red box section).

Fig. 4. Processing results partial discharge (Color figure online)

(1) Traditional classification method like maximum likelihood method in supervised classification needs fully understand land cover of the study area, for that kind of method has presented limitation at the large area with complex and diversity surface, which requires a combination of other classification method to process hybrid classification for better accuracy. And there are large error like same spectrum with different objects in supervised classification in the interpretive charts, which makes supervised classification hardly to use.
(2) Network training was processed since the initial period of BP neural network classification, which makes a better accuracy than traditional classification method like maximum likelihood method.
(3) In constant, unlike the merging mechanism of bottom-up and recognition mode of pixel based in traditional classification method, the object-oriented classification evade individual error of training sample to a great extent. In particular, the classification accuracy can be obviously improved by using object-oriented classification method, while in the area with great distinct in specter signature, the improvement was not so much obviously.

The accuracy of Object-oriented classification and BP neural network classification is similar in the study of surface coverage monitoring by remote sensing technology in central Yunnan, for the specter signature is of very distinct except same spectrum with different objects between grassland, forest land and dry land, which makes easier in distinguish surface features and obtain better accuracy. From the above, the selection of classification method in different study area should be considered according to the characteristics of different area.

Acknowledgments. Our sincere thanks to Nature Science Foundation of China (NSFC) (Nos. 41561083, 41261092) and Natural Science Fund of Yunnan Province (No. 2015FA016) for providing funding to carry put the research at Kunming University of Science and Technology, China. The authors would like to thank two anonymous reviewers for their constructive comments which were helpful to bring the manuscript into its current form.

References

1. Jia, K., Li, Q., Tian, Y., Wu, B.: A review of classification methods of remote sensing imagery. Spectrosc. Spectral Anal. **10**, 2618–2623 (2011)
2. Yingshi, Z.: The Principle and Method of Analysis of Remote Sensing Application. Science Press, Beijing (2003)
3. Foody, G.M.: Int. J. Remote Sens. **17**, 1317 (1996)
4. Bolstad, P.V., Gessler, P., Lillesand, T.M.: Positional uncertainty in manually digitized map data. Int. J. Geog. Inf. Syst. **4**, 399 (1990)
5. Chengcai, Z., Xiaonan, D., Nan, Z., Ying, Z.: Comparative study of the remote sensing image classification method based on water area estimation. Meteorol. Environ. Sci. **03**, 24–28 (2008)
6. Yin, Z., Du, P.: Study on object-oriented image classification for hyper spectral remote sensing. Remote Sens. Inf. **04**, 29–32 (2007)
7. Mai, G., Tong, X.: Study of BP neural network in rocky desertification remote sensing image classification method. J. Guangxi Teach. Educ. Univ.: Nat. Sci. Ed. **03**, 70–77 (2013)
8. Karnieli, A.: Comparison of methods for land-use classification incorporating remote sensing and GIS inputs. Appl. Geogr. **31**(2), 533–544 (2011)
9. Jing, Y., Yongfeng, C.: Accuracy evaluation of the RadarSat-2 full polari metric data for land cover classification. Comput. Telecommun. (1), 18–20 (2015)
10. Yinhui, Z., Gengxing, Z.: Classification methods of land use cover based on remote sensing technologies. J. China Agric. Resour. Reg. Plan. **03**, 24–28 (2002)
11. Zhong, C.: Research on High Resolution Remote Sensing Image Classification Technology. Chinese Academy of Sciences, Beijing (2006)
12. Cao, B., Qin, Q., Ma, H., Qiu, Y.: Research on spatial variability of water quality parameter and their adequate sampling amount in meiliang bayou of taihu lake. Geogr. Geo-Inf. Sci. **02**, 46–49+54 (2006)
13. Wang, C., Wu, W., Zhang, J.: Classification for remote sensing image based on BP neural network. J. Liaoning Tech. Univ. (Nat. Sci.) **01**, 32–35 (2009)
14. Lu, L., Zhang, Q., Li, G.: Image classification of remote sensing based on BP neural networks. Sci. Surv. Mapp. **06**, 140–143 (2012)

Application of Different Composite Index Methods in the Evaluation of Soil Heavy Metal Pollution

Yingchao Niu[1,2], Zhongfa Zhou[1,2(✉)], Denghong Huang[1,2], and Xu Yuan[3]

[1] School of Karst Science, Guizhou Normal University, Guiyang 550001, Guizhou, China
dfnycnxr@163.com, fa6897@163.com
[2] The State Key Laboratory Incubation Base for Karst Mountain Ecology Environment of Guizhou Province, Guiyang 550001, Guizhou, China
[3] Guizhou Provincial Supervision and Testing Center for Agricultural Product Quality Supervision, Guiyang 550001, Guizhou, China

Abstract. In this paper, evaluation of southeastern guizhou province some important tea-producing county soil heavy metal pollution adopts Nemerow Pollution Index, Yao Index and Mixed Weighted Model of three different composite index methods, and compares the suitability of three methods. Nemerow Pollution Index adopts arithmetic mean of subindex to improve the contribution rate of the most polluted elements and undermines the contribution of each partial effect on comprehensive pollution. Yao Index uses the ratio of index of the maximum and arithmetic mean value as weight, and diminishes the status of the serverest pollution index. But it ignored the partial contribution to the effect of comprehensive pollution. Mixed Weighted Model has good sensitivity, and it can better differentiate the degree of soil heavy metal pollution and reflect the real situation of soil environmental quality. According to the calculation results of three different comprehensive index methods, the analysis results of soil heavy metals comprehensive pollution index are obtained by Geographic Information Systems and IDW. Results show that high concentrations of heavy metals in the east and northeast of the study area.

Keywords: Comprehensive index method · The evaluation of soil heavy metal pollution · GIS · Comparison and analysis · Beidou

1 Introduction

As we all know, it would polluted the soil if the content of heavy metal were more than its natural content. And what's more, if these pollution accumulated for a long time, then it would entered our body through the food chain, caused serious damage to human being health. That is why we said it is very significative to assess the heavy metal pollution to soil [1]. Nowadays there are plenty methods of research about assessment of soil heavy metal pollution. Such as Simple exponential method: Single factor index method, Cumulative index method, Ecological risk coefficient method, etc. And

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 43–50, 2017.
DOI: 10.1007/978-981-10-3969-0_6

synthetical index method: Nemerow Index method, Yao comprehensive index method, mixed weighted, Ecological risk comprehensive coefficient, etc. These soil heavy metal pollution evaluation model has been widely used in the research community [2–5].

Different evaluation methods have different application characteristics. the evaluation method is mainly to study the overall pollution condition, but few scholars take a systemic summary and comparison of the evaluation results, and even if ever, those are only a theoretical introduction, lacking of quantitative discusses various methods of the evaluation results difference [6, 7]. Therefore, this study tries to show the advantages & disadvantages of Mixed Weighted in assessing the soil heavy metal pollution, by using GIS spatial analysis technology based on an important tea-producing county in southeastern Guizhou province, using Nemerow pollution index method, Yao comprehensive index method and mixing weighted pattern, evaluating the applicability of these methods.

2 Evaluation Methods of Soil Heavy Metal Pollution

2.1 Single Factor Index Method

Single factor index method is a kind of relative dimensionless index method, which can fully reflect the pollution degree of different pollution. At present, it has been widely applied in the evaluation of soil and crop pollution or soil environmental quality grade. It can evaluate soil pollution or soil environmental quality with single pollution index, according to the national secondary standard (GB15618-1995), the formula is:

$$P_i = C_i / S_i \tag{1}$$

In this formula, P_i is the environmental quality index of pollutants I; C_i is the measured values (mg/kg); S_i is the evaluation standards of pollutants I(mg/kg). $P_i \leq 1$ means no pollution, means pollution. And the greater of P_i value means the more serious pollution.

2.2 Assessment of Compositive Index Method

Nemerow Pollution Index Method. The formula of Mello pollution index method is:

$$P_n = \sqrt{\frac{\left(\frac{1}{n}\sum_{i=1}^{n} P_i\right)^2 + P_{i(\max)}^2}{2}} \tag{2}$$

In this formula, P_n is Nemerow pollution index; n is the number of monitoring the pollution index; $P_i(\max)$ is the maximum of every index for the pollution index [8, 9].

Yao Comprehensive Index Method. Yao comprehensive index method determines the main pollutant of partial firstly, and then draws the average of all pollutans partial, finally worksout the geometric average of the two partial. Index formula is defined as:

$$P_n = \sqrt{\max(P_i) * (\frac{1}{n}\sum_{i=1}^{n} P_i)} \tag{3}$$

In this formula, P_i is a pollution index; n is the number of monitoring pollution; max(P_i) is the maximum of every index of the pollution index [10].

The Pattern of Mixing Weighted. The formula of mixing weighted is [11]:

$$P = \sum\nolimits_1 W_{i1} I_i + \sum\nolimits_2 W_{i2} I_i \tag{4}$$

In this formula, I_i is the single contaminate index of heavy metal; $\sum_1$ is the sum of all the single pollution index when I_i is greater than 1; $\sum_2$ is the sum of all the I_i single pollution index. When $I_i > 1$, $W_{i1} = \frac{I_i}{\sum_1 \cdot I_i}$; and as for all I_i, $W_{i2} = \frac{I_i}{\sum_2 \cdot I_i}$.

3 Instance Analysis

3.1 Collection and Disposal

Soil sampling point from some important tea-producing country in southeastern Guizhou province. There are 146 sampling point according to the characteristics of tea garden

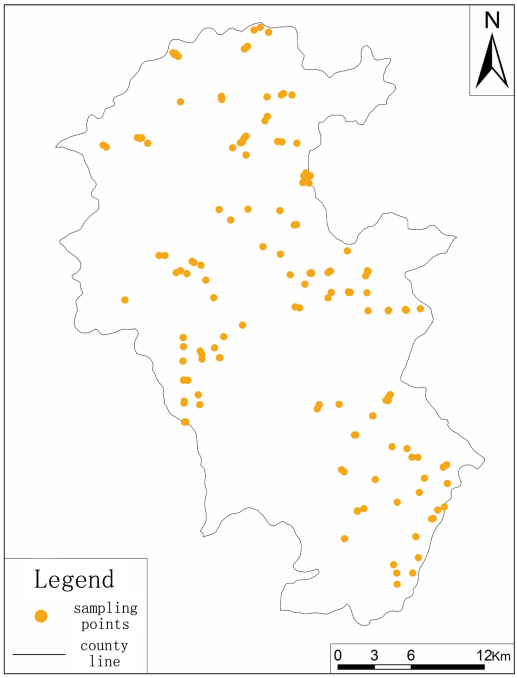


Fig. 1. Distribution map of soil sampling points in the study area

distribution, and we collect by using serpentine method in soil of 50 m * 50 m, getting each sampling point's coordinate of latitude and longitude by using Beidou. What's more, we also record conditions of environment and the surrounding land for corresponding sample. The sampling point and the study area distribution is shown in Fig. 1.

3.2 Analysis and Comparison

Because the number of sampling points is 146, which is more, we choose five representative soil samples in the study area and evaluation within five single contaminate index of heavy metals were calculated using Nemerow index method, Yao the comprehensive index method and the mixed weighted method. The results are shown in Table 1.

Table 1. Comparison of assessment results

Serial number	Hg	As	Cd	Pd	Cr	Nemerow index	Yao index	Mixed weighted
1	0.410	0.272	1.120	0.287	0.458	0.378	0.285	1.822
2	1.580	0.532	1.257	0.156	0.493	0.786	0.635	2.589
3	0.327	1.054	0.393	0.265	0.194	0.328	0.235	1.717
4	1.327	0.730	0.713	0.158	0.489	0.557	0.453	2.224
5	0.323	0.197	0.333	0.122	0.523	0.091	0.078	0.362

Experiments show that Nemerow comprehensive index method weakens the contribution rate of each subindex influence on integrated pollution by using partial arithmetic mean and strengthening the contribution rate of the most polluted element. Yao comprehensive index method is the improvement based on Nemerow index method, which chooses the index of maximum value and the ratio of arithmetic average value as weights, abates the biggest pollution index in the evaluation of position, but also ignores the contribution of each subindex on comprehensive pollution. So it is difficult to show the soil environmental quality differences using Nemerow comprehensive index method and Yao comprehensive index method. As for mixed weighted method, it compares and analyzes the single pollution index first, and is divided into two part, which is single pollution index exceeding official standards and single pollution index qualified, and both of them have different weights. So the result is one to one correspondence. In others words, if each heavy metal pollutans are not overproof of the monitoring sampling points, the comprehensive pollution index is not overproof, and on the contrary, the comprehensive will exceed official standards. So this method avoids the influence that the elements always be changed with the change of pollution elements's weights in Nemerow comprehensive index method and Yao comprehensive index method.

3.3 Assessment of Soil Heavy Metal Pollution

Based on single pollution index to evalutate the siol heavy metal pollution, we get each monitoring point's comprehensive pollution index of Nemerow comprehensive index method, Yao comprehensive index method and mixed weighted method. Then using

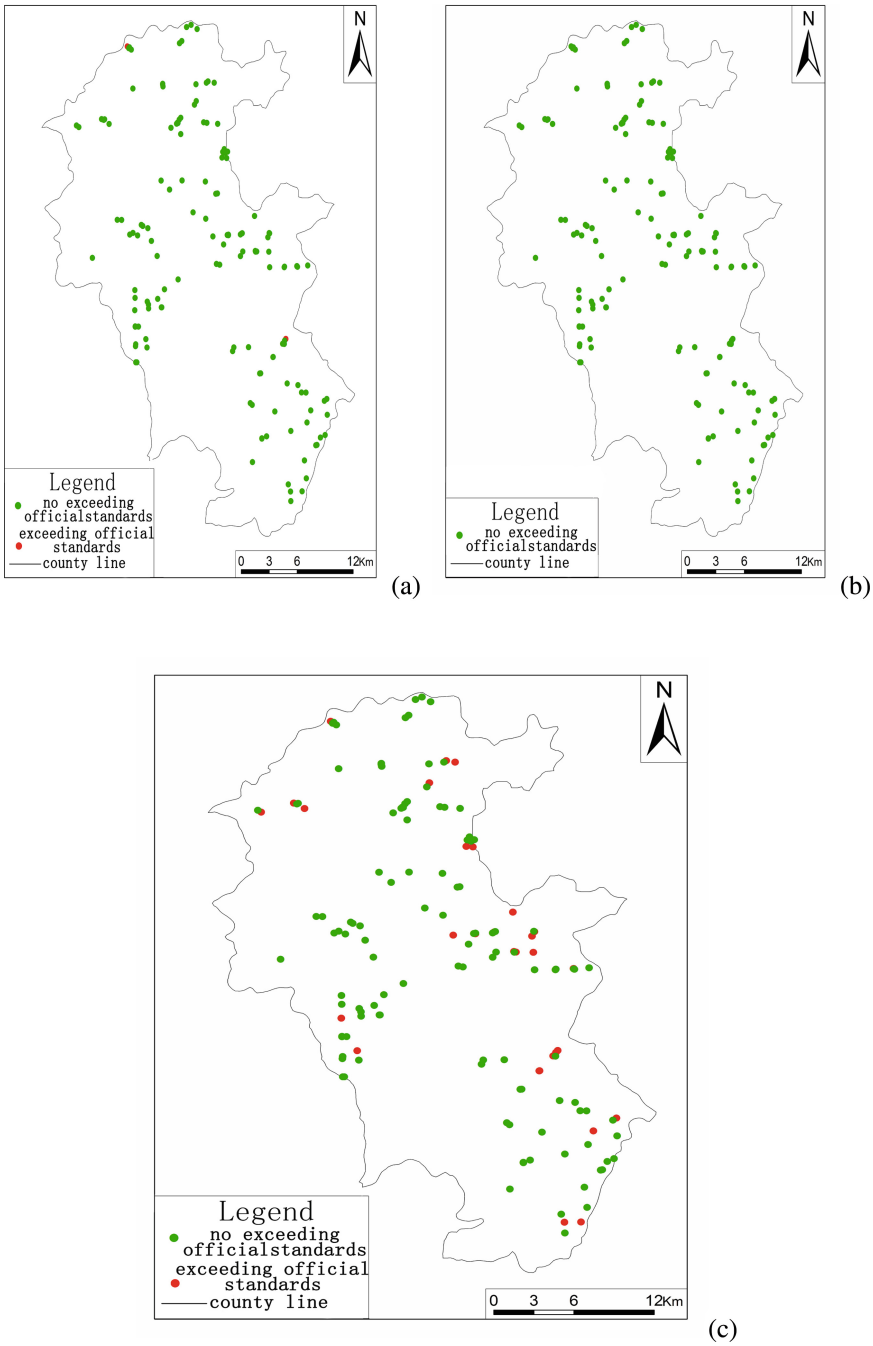


Fig. 2. Map of comprehensive pollution indexes exceeding official standards, the map of (a) exceeding official standards by Nemerow index, the map of (b) exceeding official standards by Yao index, the map of (c) exceeding official standards by mixed weighted

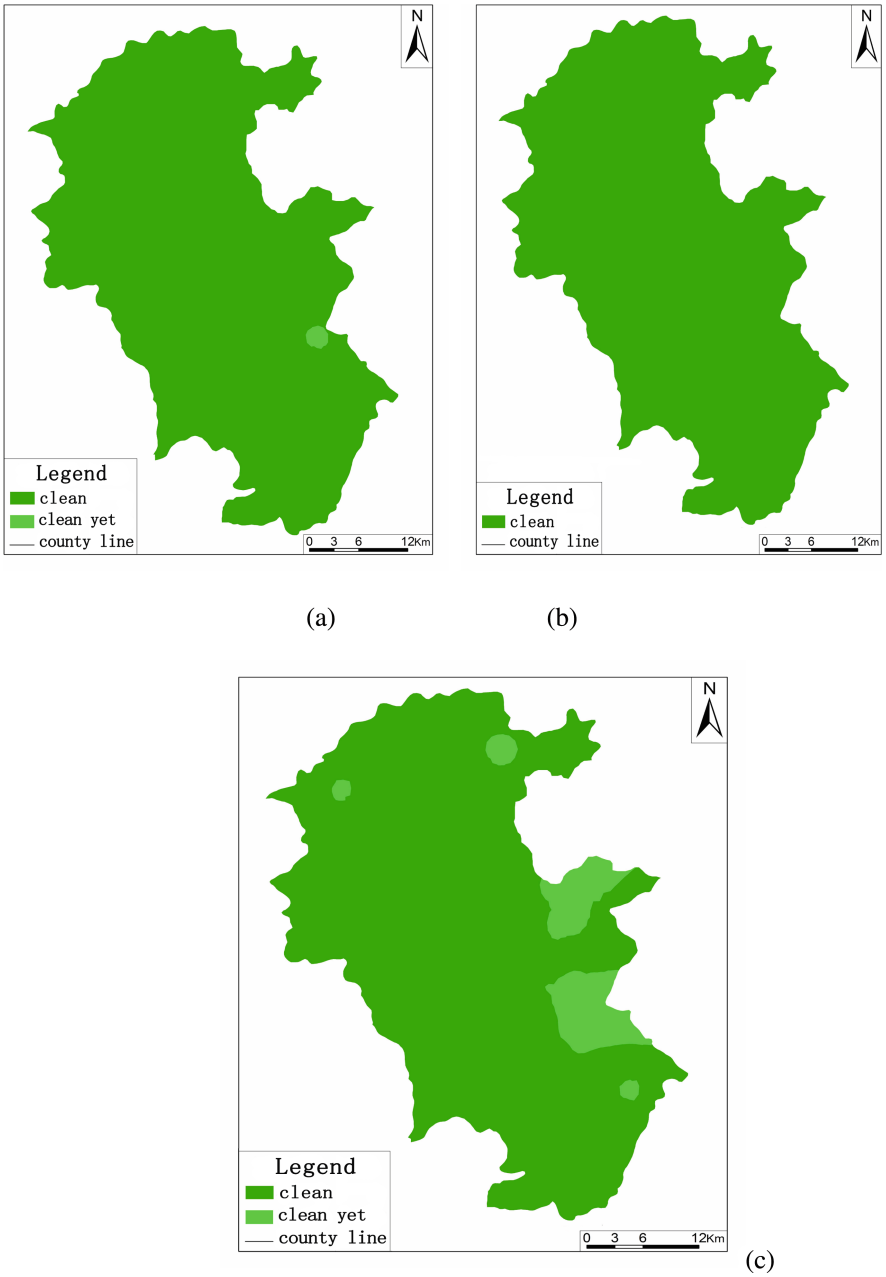


Fig. 3. Map of assessment on the comprehensive pollution indexes: the map of (a) assessment on the comprehensive pollution indexes by Nemerow index, the map of (b) assessment on the comprehensive pollution indexes by Yao index, the map of (c) assessment on the comprehensive pollution indexes by mixed weighted

software Arcgis10.1space analysis, we draw a conclution of comprehensive pollution indexes exceeding official standars, as shown in Fig. 2.

From the Fig. 2, we can know that the numember of exceeding official standards is Yao index<Nemerow index<Mixed weighted. But the trend of three methods of evaluation results is roughly the same.

The monitoring sampling points in the study area is 146, and single element exceeds official standards total 31. From the results of Nemerow index method, Yao index method and Mixed weighted method, we can learn about that it calculates the excess points and the number of single element exceeding official standards being the same, and statistical result is the same as the condition of assessment figure by using mixed weighted method.

3.4 Analysis of Result

We can know that Nemerow index method and Yao index method can only distinguish whether environment is clean or have been contaminated through comparing and analyzing both methods. But if we evaluate single factor pollution index that is greater than 1, and the results may not be exceeding official standards, which is not consistent with actual situation. However, mixed weighted is a method that it compares and analyzes single pollution index and weighed sum, so its results will be more accurate, and it can reflect the real condition of the soil environmental quality. That is why we said this method is better than others in evaluating soil quality.

4 Different Comprehensive Index Evaluation Based on GIS

With the support of ArcGIS10.1 system, we do spatial interpolation using IDW. And it generates forecast surface based on the monitoring points according to the evaluation unit. So that it can shows the area of spatial distribution of soil environmental quality situation. As shown in Fig. 3, in the east and northeast of the study area, there are high concentration of heavy metal, which is roughly the same as sampling points in Fig. 2.

5 Conclusion

(1) Compared with the different results of evaluation result about soil heavy metal from Nemerow pollution index method, Yao comprehensive index method and mixed weighted method, we can know that mixed weighted method can overcome the shortcoming of Nemerow and Yao, and can better reflect the situation of soil quality. And it is more sensitive than the others.

(2) The spatial analysis function of GIS can make analysis and evaluation of the study area, which is an effective tool to analyze the spatial dimension of soil heavy metal pollution. And evaluation results show that the concentration of heavy metals of the east and northeast regions is high, which is the same as the exceeded official standard sampling points.

(3) Although mixed weighted can make up the shortfall of Nemerow index method and Yao index method, yet mixed weighted method has not had corresponding pollution grade division standard. We can only distinguish whether it is clear according to standard, but can not distinguish the pollution degree and level, and calculation is very complex, so we need further research in practice.

Acknowledgments. In this paper, the research was sponsored by Science and technology plan of Guizhou province Mountain high efficiency agriculture industrial park development and application of intelligent management system based on Beidou satellite (GY[2015] No. 3001); Major science and technology projects in Guizhou Province Study and Application on the quality safety evaluation, detection and traceable key technologies of tea and vegetables in Guizhou province ([2013] No. 6024); Major applied basic research project of Guizhou Province Study on the ecological restoration and optimal control of eco-economic system of Karst rocky desertification (JZ[2014] No. 20020-1); Science and technology plan of Guizhou province The research and countermeasures of heavy metal content and the tea garden soil heavy metals in Guizhou province (SY[2011] No. 3092).

References

1. Wang, B., Chen, L.: Review on methods of soil quality evaluation. Soil Water Conserv. Sci. China **4**(2), 120–126 (2006)
2. Zhong, X., Zhou, S., Zhao, Q.: Spatial characteristics and potential ecological risk of soil heavy metals contamination in the Yangtze river delta a case study of Taicang city, Jiangsu province. Geoscience **27**(3), 395–400 (2007)
3. Chunhua, H., Jiang, J., Zhou, W.: Risk evaluation and sources analysis of heavy metals in vegetable field soil of rural area around Poyang lake. Geoscience **32**(6), 771–776 (2012)
4. Krishna, A.K., Mohan, K.R., Murthy, N.N., et al.: Assessment of heavy metal contamination in soils around chromite mining areas, Nuggihalli, Karnataka, India. Environ. Earth Sci. **70** (2), 699–708 (2013)
5. Okedeyi, O.O., Dube, S., Awofolu, O.R., et al.: Assessing the enrichment of heavy metals in surface soil and plant (digitaria eriantha) around coal-fired power plants in South Africa. Environ. Sci. Pollut. Res. **21**(6), 4686–4696 (2014)
6. Fan, S., Gan, Z., Li, M.: Progress of assessment methods of heavy metal pollution in soil. Chin. Agric. Sci. Bull. **26**(17), 310–315 (2010)
7. Ge, W., Jianping, W., Mei, X.: GIS of comprehensively environmental evaluation in Pudong new area. Remote Sens. Technol. Appl. **15**(3), 189–193 (2000)
8. Li, S., Cao, G., Shi, P., et al.: Status quo and evaluation of the spatial distribution of heavy metals in urban soil of Qingdao city. J. Ecol. Rural Environ. **1**, 112–117 (2015)
9. Youning, X., Ke, H., Zhao, A., et al.: Assessment of heavy metals contamination of farmland soils in some gold mining area of Xiao Qinling. J. Soil Sci. **38**(4), 732–736 (2007)
10. Yang, Y., Shi, X., Zhang, C.: Spatial distribution an evaluation of heavy metal pollution of reclaiming village based on nemerow integrated pollution index method. Res. Soil Water Conserv. **4**, 059 (2016)
11. Liu, Y., Zhang, L., Han, X., et al.: Spatial variability and evaluation of soil heavy metal contamination in the urban-transect of shanghai. Environ. Sci. **33**(2), 599–605 (2012)

Hyperspectral Image Denoising Based on Subspace Low Rank Representation

Mengdi Wang[1(✉)], Jing Yu[2], Lijuan Niu[3], and Weidong Sun[1]

[1] State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory of Information and Science, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
wangmdl2@mails.tsinghua.edu.cn, wdsun@tsinghua.edu.cn
[2] College of Computer Science and Technology, Beijing University of Technology, Beijing 100124, China
jing.yu@bjut.edu.cn
[3] Cancer Hospital of Chinese Academy of Medical Sciences, Beijing 100021, China
niulijuan8197@126.com

Abstract. Hyperspectral images (HSIs) are often degraded by different kinds of noises. Low rank (LR)-based methods have achieved great performance in HSI denoising problem. However, the LR-based methods only consider the rank of the whole spectral space, conducting no constraints on the intrinsic structure within the LR space. In fact, the spectral vectors can be classified into different categories based on the land-covers. As a result, the spectral space can be modelled as a union of multiple LR subspaces. Regarding this structure, we introduce the framework of subspace low rank (SLR) representation into HSI denoising problem and propose a novel SLR-based denoising method for HSIs. Experiments conducted on both simulated and real data show that our method achieves great improvement over the state-of-art methods qualitatively and quantitatively.

Keywords: Hyperspectral image · Denoising · Low rank representation · Subspace low rank

1 Introduction

Hyperspectral images (HSIs) can provide abundant spectral information with hundreds of spectral bands, thus is widely are widely used in various fields, such as agriculture, environment monitoring and mineralogy. However, the environmental and sensor noises always degrade the quality of HSIs during the acquisition process, which are harmful to the subsequent tasks, including segmentation, classification, spectral unmixing, and also target detection.

There are lots of outstanding works conducted on HSI denoising. Traditional methods for the denoising of natural images, such principle component analysis (PCA) [1], can be directly applied to HSIs band-by-band, ignoring the spectral information, which is an important characteristic of HSIs. No convincing results were

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 51–59, 2017.
DOI: 10.1007/978-981-10-3969-0_7

achieved by those methods. Combination of the spatial and spectral information in HSIs is essential in HSI processing. Othman and Qian [2] proposed a hybrid spatial–spectral derivative-domain wavelet shrinkage method for HSI denoising and achieved quite good results. Zhang [3] extended the traditional total variation (TV) regularization to a 3-dimensional (3D) one, termed as cubic total variation (CTV), for HSI denoising. In further, Yuan *et al.* [4] improved CTV to be a spectral-spatial adaptive one. Additionally, the conception of principle component of PCA is extended to 3D using tensor analysis. For example, a low rank-(K_1, K_2, K_3) tensor approximation method is proposed in [5] for dimensionality reduction and joint denoising. Low rank (LR)-based method is firstly introduced by Zhang *et al.* [6] for HSI denoising, based on the LR property of the spectral space in HSI. Our recent work [7] combined the nonlocal spatial correlation together within the LR framework. However the LR-based methods have no constraint on the intrinsic structure of the LR space. According to the mixed nature of HSI, the spectral space of HSI is constituted by a union of several LR subspaces $\cup_{i=1}^{K} S_i$ (where S_i represents the subspace). Samples drawn from the union $\cup_{i=1}^{K} S_i$ will be treated as if they are sampled from a single LR space defined by a sum $\sum_{i=1}^{K} S_i$ in LR-based methods [8]. As the sum space contains and is much larger than the union space, the recovery using LR-based methods may be inaccurate.

Subspace low rank (SLR) representation is proposed by Liu *et al.* [9] for subspace segmentation and now is widely used in multiple fields, such as subspace clustering [9, 10], saliency detection [11, 12] and subspace number estimation [8]. SLR representation can structurally represent the data that are derived from a union of multiple LR subspaces. Therefore, here we introduce the framework of SLR representation into HSI denoising. Moreover, different from traditional SLR representation framework that uses the observed corrupted data as the dictionary, our method uses the latent clean HSI as the dictionary to achieve better representation, and the latent data is estimated during the iteration solution procedure.

The rest of this paper is organized as follows. Section 2 introduces the SLR-based denoising method. Experimental results and discussion are shown in Sect. 3, and Sect. 4 draws the conclusion.

2 The Proposed Method

2.1 Noise Model in HSI

HSIs are often corrupted by noises during acquisition. The noises can be divided into sparse noise and Gaussian noise, where sparse noise includes the salt & pepper noise and stripe noise. Therefore an observed HSI $\mathbf{X} \in \mathbb{R}^{B \times P}$ (which is reorganized from the HSI datacube $X \in \mathbb{R}^{M \times N \times B}$ with $P = MN$), where B is the band number and P is the pixel number, can be modelled as

$$\mathbf{X} = \mathbf{X}_0 + \mathbf{S} + \mathbf{N} \tag{1}$$

where $\mathbf{X}_0$ is the latent clean HSI to be recovered, $\mathbf{S}$ is the sparse noise and $\mathbf{N}$ is the Gaussian noise.

2.2 LR-Based Denoising Method

LR-based methods [6, 7, 13] have been widely used in the recovery of $\mathbf{X}_0$ in Eq. (1) for HSI, as the spectral space of HSI is highly LR, which can be derived from LMM [14]. According to LMM, all the spectral vectors in HSI can be represented by linear combination of the spectra of the underlying materials. As a result, the limited number of the materials promises the LR property of the spectral space of HSI. Therefore the reconstruction of $\mathbf{X}_0$ is actually to restore the LR structure from $\mathbf{X}$, for which robust principal component analysis (RPCA) method is widely used. The formulation of RPCA is as follows:

$$\min_{\mathbf{X}_0,\mathbf{S}} \operatorname{rank}(\mathbf{X}_0) + \lambda\|\mathbf{S}\|_0 + \frac{\gamma}{2}\|\mathbf{X} - \mathbf{X}_0 - \mathbf{S}\|_{\mathrm{F}}^2 \tag{2}$$

where λ and γ are balance parameters among different regularization items.

However, Eq. (2) only conducts constraint on the rank of the whole spectral space, without any regularization on the intrinsic structure within $\mathbf{X}_0$. Based on the class property of the landcovers, the spectral space of HSI can be divided into multiple subspaces, termed as $\{S_i\}_{i=1}^K$. The spectral vectors within each class are highly correlated, thus they should lie in a low-dimensional manifold, which means that S_i is LR. Therefore the whole spectral space can be seen as the union of such LR subspaces $\cup_{i=1}^K S_i$. But the LR-based methods using Eq. (2) will regard the samples drawn from the union space as if they are sampled from a single LR space defined by the sum $\sum_{i=1}^K S_i$ because $\sum_{i=1}^K S_i$ is much bigger that $\bigcup_{i=1}^K S_i$, modelling the spectral space of HSI using $\sum_{i=1}^K S_i$ may lead to inaccuracy in denoising.

2.3 SLR-Based Denoising Method

As the spectral space of HSI is a union of multiple LR subspaces, we introduce the framework of SLR representation to structurally restore the spectral space from the noise corruption, within which the observed data X is modeled as

$$\mathbf{X} = \mathbf{AZ} + \mathbf{S} + \mathbf{N} \tag{3}$$

where $\mathbf{A}$ is a dictionary that can span each subspaces in $\mathbf{X}_0$, and $\mathbf{Z}$ should be LR. When $\mathbf{A} = \mathbf{I}$, Eq. (3) is equivalent to Eq. (1). So SLR-based method is actually a generalization of the LR-based method. Equation (3) is first proposed by Liu *et al.* [9], which can structurally represent the union space. Here we introduce the framework into HSI denoising issue. Combining the constraint in Eq. (1), the cost function using SLR representation is

$$\min_{\mathbf{X}_0,\mathbf{Z},\mathbf{S}} \operatorname{rank}(\mathbf{Z}) + \lambda\|\mathbf{S}\|_0 + \frac{\gamma}{2}\left\{\|\mathbf{X} - \mathbf{AZ} - \mathbf{S}\|_{\mathrm{F}}^2 + \|\mathbf{X} - \mathbf{X}_0 - \mathbf{S}\|_{\mathrm{F}}^2\right\}, \tag{4}$$

where $\operatorname{rank}(\mathbf{Z})$ models the low-rankness of the coefficient matrix $\mathbf{Z}$, $\|\mathbf{S}\|_0$ regulates the sparsity of $\mathbf{S}$ using ℓ_0-norm, $\|\mathbf{X} - \mathbf{AZ} - \mathbf{S}\|_{\mathrm{F}}^2$ and $\|\mathbf{X} - \mathbf{X}_0 - \mathbf{S}\|_{\mathrm{F}}^2$ are both fidelity

items, regulating the energy level of the Gaussian noise $\mathbf{N}$. λ and γ are balance parameters. Equation (4) is nonconvex, so we replace the rank minimization and ℓ_0-norm with the nuclear norm and ℓ_1-norm, respectively,

$$\min_{\mathbf{X}_0,\mathbf{Z},\mathbf{S}} \|\mathbf{Z}\|_* + \lambda\|\mathbf{S}\|_1 + \frac{\gamma}{2}\left\{\|\mathbf{X} - \mathbf{AZ} - \mathbf{S}\|_\mathrm{F}^2 + \|\mathbf{X} - \mathbf{X}_0 - \mathbf{S}\|_\mathrm{F}^2\right\}, \tag{5}$$

It is demonstrated in [9] the recovery of $\mathbf{X}_0$ can be guaranteed if $\mathbf{A}$ can well space each of the subspaces. [9] recommended to use $\mathbf{X}$ itself as the dictionary. However in HSI denoising issue, $\mathbf{X}$ is seriously corrupted, the representation by $\mathbf{X}$ may not be accurate. According to the theoretically demonstration of the effectiveness in [9] for the recovery of $\mathbf{X}_0$ using Eq. (5), a choice of the clean data itself can guarantee the recovery of its column space. Therefore, in this paper we use the latent HSI $\mathbf{X}_0$ as the dictionary and it can be estimated simultaneously during the iterative optimal procedure. By setting $\mathbf{A} = \mathbf{X}_0$ and adding an extra matrix $\mathbf{J}$, the optimization is converted to

$$\begin{aligned}\min_{\mathbf{X}_0,\mathbf{Z},\mathbf{J},\mathbf{S}} \|\mathbf{J}\|_* + \lambda\|\mathbf{S}\|_1 + \frac{\gamma}{2}\left\{\|\mathbf{X} - \mathbf{X}_0\mathbf{X} - \mathbf{S}\|_\mathrm{F}^2 + \|\mathbf{X} - \mathbf{X}_0 - \mathbf{S}\|_\mathrm{F}^2\right\} \\ \text{s.t.} \quad \mathbf{Z} = \mathbf{J}.\end{aligned} \tag{6}$$

Equation (6) can be solved using the Inexact Augmented Lagrange Multiplier (IALM) method [15]. According to IALM, the constraint $\mathbf{Z} = \mathbf{J}$ can be removed by introduction of a Lagrange multiplier $\mathbf{Y}$,

$$\begin{aligned}\mathbf{L}(\mathbf{X}_0,\mathbf{Z},\mathbf{S},\mathbf{J},\mathbf{Y}) = {} & \|\mathbf{J}\|_* + \lambda\|\mathbf{S}\|_1 \\ & + \frac{\gamma}{2}\left\{\|\mathbf{X} - \mathbf{X}_0\mathbf{Z} - \mathbf{S}\|_\mathrm{F}^2 + \|\mathbf{X} - \mathbf{X}_0 - \mathbf{S}\|_\mathrm{F}^2\right\}, \\ & + \mathrm{Tr}\left(\mathbf{Y}^\mathrm{T}(\mathbf{Z} - \mathbf{J})\right) + \frac{\mu}{2}\|\mathbf{Z} - \mathbf{J}\|_\mathrm{F}^2\end{aligned} \tag{7}$$

where μ is a positive scalar, $\mathrm{Tr}(\cdot)$ means matrix trace. The variables $\mathbf{J}$, $\mathbf{Z}$, $\mathbf{X}_0$, $\mathbf{S}$ and $\mathbf{Y}$ can be iteratively updated by fixing the others constant. The denoised HSI data can be reconstructed using the optimal solution $\mathbf{X}_0^*$ of Eq. (7).

A detailed algorithm is given in Algorithm 1. For $\mathbf{J}$, when remaining the other variables constant, Eq. (7) can be written as,

$$\arg\min_{\mathbf{J}} \frac{1}{\mu}\|\mathbf{J}\|_* + \frac{1}{2}\|\mathbf{J} - (\mathbf{Z} + \mathbf{Y}/\mu)\|_\mathrm{F}^2 \tag{8}$$

which can be solved using the singular value thresholding (SVT) operator [16]. When the variables that are independent with $\mathbf{Z}$ are ignored, the minimization function will be,

$$\arg\min_{\mathbf{Z}} \frac{\gamma}{2}\|\mathbf{X} - \mathbf{X}_0\mathbf{Z} - \mathbf{S}\|_\mathrm{F}^2 + \mathrm{Tr}\left(\mathbf{Y}^\mathrm{T}(\mathbf{Z} - \mathbf{J})\right) + \frac{\mu}{2}\|\mathbf{Z} - \mathbf{J}\|_\mathrm{F}^2, \tag{9}$$

$\mathbf{Z}$ can be solved by setting the derivative with respect to $\mathbf{Z}$ to be zero, as shown in the equation in the step 2 in Algorithm 1. The update procedure for $\mathbf{X}_0$ is similar to $\mathbf{Z}$. The optimization of $\mathbf{S}$ is shown in the step 2, which can be solved using soft thresholding.

Algorithm 1 IALM-based method for solving problem Eq.(7).

Input Corrupted spectral matrix $\mathbf{X}$, balance parameters λ, γ

Initialize $k = 0$, $\mathbf{Z} = \mathbf{J} = \mathbf{S} = \mathbf{0}$, $\mathbf{X}_0 = \mathbf{X}$, $\mathbf{Y} = \mathbf{0}$, $\mu = 10^{-3}$, $\rho = 1.1$, $\mu_{\max} = 10^6$, $\varepsilon = 10^{-8}$

Output $\mathbf{X}_0^*$, $\mathbf{Z}^*$, $\mathbf{S}^*$

while not converged **do**

1. Update $\mathbf{J}$, with $\mathbf{Z}$, $\mathbf{X}_0$, $\mathbf{S}$, $\mathbf{Y}$ fixed
$$\mathbf{J} \leftarrow \arg\min_{\mathbf{J}} \tfrac{1}{\mu}\|\mathbf{J}\|_* + \tfrac{1}{2}\|\mathbf{J} - (\mathbf{Z} + \mathbf{Y}/\mu)\|_{\mathrm{F}}^2$$
2. Update $\mathbf{Z}$, with $\mathbf{J}$, $\mathbf{X}_0$, $\mathbf{S}$, $\mathbf{Y}$ fixed
$$\mathbf{Z} \leftarrow \left(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \tfrac{\mu}{\gamma}\mathbf{I}\right)^{-1}\left(\tfrac{\mu\mathbf{J}-\mathbf{Y}}{\gamma} + \mathbf{X}_0^{\mathrm{T}}(\mathbf{X} - \mathbf{S})\right)$$
3. Update $\mathbf{X}_0$, with $\mathbf{J}$, $\mathbf{Z}$, $\mathbf{S}$, $\mathbf{Y}$ fixed
$$\mathbf{X}_0 \leftarrow \left(\mathbf{Z}^{\mathrm{T}} + \mathbf{I}\right)(\mathbf{X} - \mathbf{S})\left(\mathbf{Z}\mathbf{Z}^{\mathrm{T}} + \mathbf{I}\right)^{-1}$$
4. Update $\mathbf{S}$, with $\mathbf{J}$, $\mathbf{X}_0$, $\mathbf{Z}$, $\mathbf{Y}$ fixed
$$\mathbf{S} \leftarrow \arg\min_{\mathbf{S}} \tfrac{\lambda}{2\gamma}\|\mathbf{S}\|_1 + \tfrac{1}{2}\left\|\mathbf{S} - \left(\mathbf{X} - \tfrac{\mathbf{X}_0(\mathbf{Z}+\mathbf{I})}{2}\right)\right\|_{\mathrm{F}}^2$$
5. Update $\mathbf{Y}$ and μ
$$\mathbf{Y} \leftarrow \mathbf{Y} + \mu(\mathbf{Z} - \mathbf{J}), \quad \mu \leftarrow \min(\rho\mu, \mu_{\max})$$
6. check the convergence condition $\|\mathbf{Z} - \mathbf{J}\|_\infty < \varepsilon$

end while

3 Experimental Results and Discussion

In our experiments, simulated data and real data are both used. To evaluate the effectiveness of the proposed method, we compare our method with two other methods: the LR-based method proposed in [6] and the group LR-based method proposed in [7]. The qualitative and quantitative results of the three methods are reported and discussed.

3.1 Experiments on Simulated Data

In the simulated experiments, the HSI of Pavia University acquired by the reflective optics system imaging spectrometer (ROSIS) is used. There are 103 bands in it, with 610 × 340 pixels.

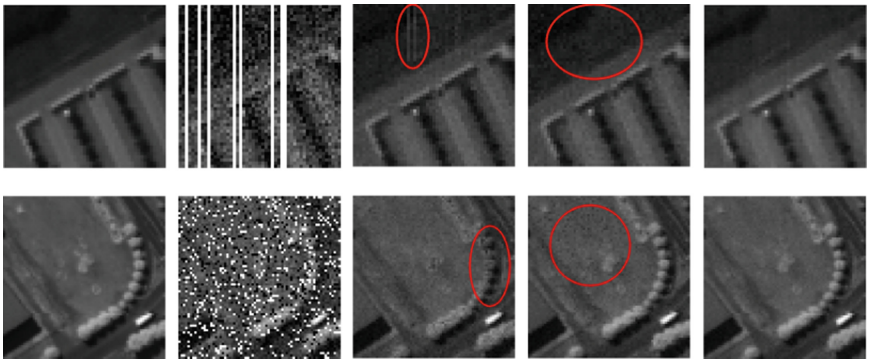
We add three typical noises into Pavia University in the experiments: the Gaussian noise, the salt & pepper noise, and the stripe noise. The stripe noise often appears in pushbroom systems. In the simulated data, the Gaussian noises with σ = 5% are added to all the bands, the salt & pepper noises with a percentage of 20% are added to ten randomly-selected bands, and the stripe noises are randomly added to ten bands with

(a) (b) (c) (d)

Fig. 1. Examples of noisy bands in the simulated Pavia University data, (a)–(b) are the original and noisy data of band 22, and (c)–(d) are those of band 84.

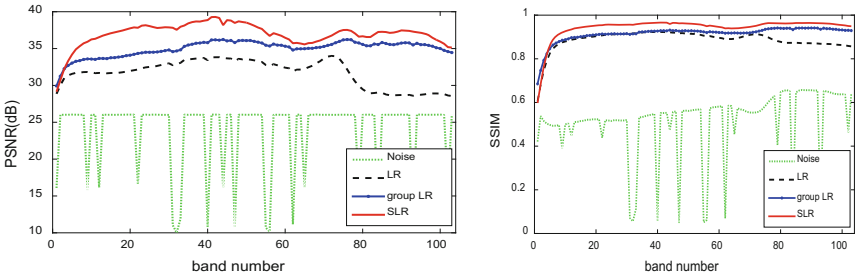
10-line stripes. Figure 1 shows two of the noisy bands: band 22 has the stripe noise and Gaussian noise, and band 84 has all the three kinds of noises.

The close-up denoising results of band 22 and 84 are shown in Fig. 2. It can be observed that the proposed method performs the best because it can eliminate all the noises thoroughly and reconstruct the structure well. LR fails to remove the stripe noise thoroughly in the enlarged area of band 22 and cannot reconstruct the spatial details in band 84, as shown in the red ellipse area in Fig. 2(c). Group LR performs much better than LR. However Fig. 2(d) shows that there are still some noises remaining in the red ellipse area both in band 22 and 84.



(a) original band (b) noisy band (c) LR (d) group LR (e) SLR

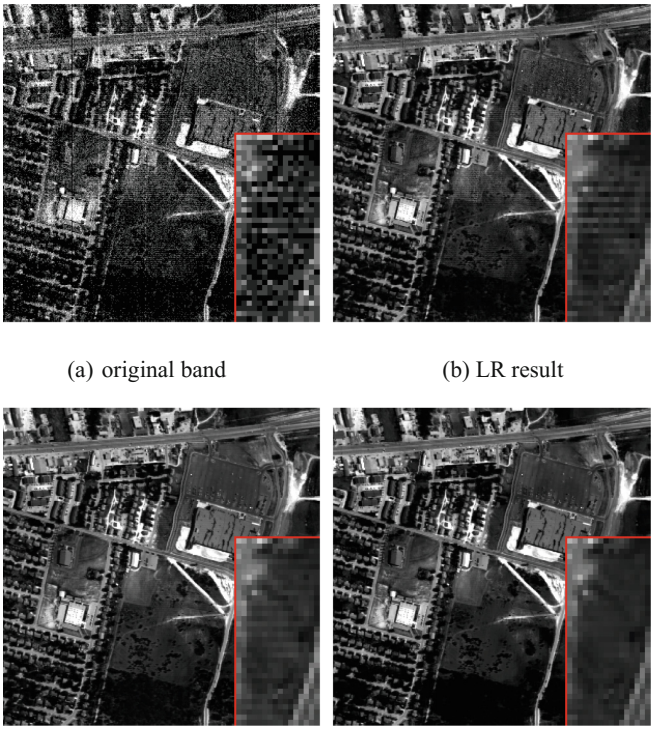
Fig. 2. Experimental results of band 22 and 84 in the simulated Pavia University data. Top: band 22; Down: band 84. (Color figure online)



(a) PSNR (b) SSIM

Fig. 3. Quantitative evaluation of the three methods

To better evaluate the performance of our proposed method, peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) indices are selected for the quantitative evaluation. They are both calculated band-by-band between the reconstructed data and the ground truth. The higher values of the two indices indicate better performance. Figure 3 shows the comparison results. It is obvious that our method achieves great



(a) original band (b) LR result

(c) group LR (d) SLR

Fig. 4. Experimental results of band 186 in urban (Color figure online)

improvement upon both LR and group LR. In average, our method gets an improvement of 5.28 dB and 1.98 dB in PSNR and that of 0.056 and 0.028 in SSIM upon LR and group LR.

3.2 Experiments on Real Data

A hyperspectral digital collection experiment (HYDICE) data, Urban of Copperas Cove, Texas (called as Urban for brief in the following) is used in our experiments as real noisy data. There are 307 $\times$ 307 pixels and 210 bands in the data. Bands 104–108, 139–151 and 207–210 are heavily polluted by the atmosphere and water absorption and thus are removed from the original data. In the Urban data, several bands are corrupted by heavy Gaussian noises and stripe noises, such as band 186 shown in Fig. 4(a).

Figures 4(b)–(d) show the denoising results by LR, group LR and our proposed method, with the sub-image in the downright corner showing the close-up of the area in the red rectangle. It is obvious that LR fails in eliminating the stripe noise. Group LR performs much better, but there is still some light stripe noise remaining. From Fig. 4(d) we can observe that, our proposed method can eliminate all the noises and reconstruct the spatial details simultaneously.

4 Conclusion

In this paper, we have proposed a novel HSI denoising method based on SLR representation. In this method, considering the intrinsic structure of the spectral space in HSI, SLR representation is introduced to reconstruct the corrupted HSI and the latent clean data is chosen as the dictionary instead of the heavily corrupted data itself to get a better representation. Experimental results both on simulated and real HIS data demonstrate that, our proposed method outperforms the state-of-art methods both in visual inspection and quality indices.

Acknowledgments. This work was supported in part by the National Nature Science Foundation (No. 61171117) and the Capital Health Research and Development of Special (No. 2014-2-4025) of China.

References

1. Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**(6), 417 (1933)
2. Othman, H., Qian, S.E.: Noise reduction of hyperspectral imagery using hybrid spatial-spectral derivative-domain wavelet shrinkage. IEEE Trans. Geosci. Remote Sens. **44**(2), 397–408 (2006)
3. Zhang, H.: Hyperspectral image denoising with cubic total variation Model. ISPRS Ann. photogramm. Remote Sens. Spat. Inf. Sci. **7**, 95–98 (2012)

4. Yuan, Q., Zhang, L., Shen, H.: Hyperspectral image denoising employing a spectral-spatial adaptive total variation model. IEEE Trans. Geosci. Remote Sens. **50**(10), 3660–3677 (2012)
5. Renard, N., Bourennane, S., Blanc-Talon, J.: Denoising and dimensionality reduction using multilinear tools for hyperspectral images. IEEE Geosci. Remote Sens. Lett. **5**(2), 138–142 (2008)
6. Zhang, H., He, W., Zhang, L., Shen, H., Yuan, Q.: Hyperspectral image restoration using low-rank matrix recovery. IEEE Trans. Geosci. Remote Sens. **52**(8), 4729–4743 (2014)
7. Wang, M., Yu, J., Sun, W.: Group-based hyperspectral image denoising using low rank representation. In: IEEE Processing of the ICIP, pp. 1623–1627 (2015)
8. Qian, Y., Ye, M.: Hyperspectral imagery restoration using nonlocal spectral-spatial structured sparse representation with noise estimation. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. **6**(2), 499–515 (2013)
9. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 171–184 (2013)
10. Favaro, P., Vidal, R., Ravichandran, A.: A closed form solution to robust subspace estimation and clustering. In: Processing of the IEEE Conference on Computer Vision Pattern Recognition, pp. 1801–1807, June 2011
11. Lang, C., Liu, G., Yu, J., Yan, S.: Saliency detection by multitask sparsity pursuit. IEEE Trans. Image Process. **21**(3), 1327–1338 (2012)
12. Cheng, B., Liu, G., Wang, J., Huang, Z., Yan, S.: Multi-task low-rank affinity pursuit for image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2439–2446, November 2011
13. Huang, H., Christodoulou, A.G., Sun, W.: Super-resolution hyperspectral imaging with unknown blurring by low-rank and group-sparse modeling. In: Proceedings of the ICIP IEEE, pp. 2155–2159 (2014)
14. Iordache, M.D., Bioucas-Dias, J.M., Plaza, A.: Sparse unmixing of hyperspectral data. IEEE Trans. Geosci. Remote Sens. **49**(6), 2014–2039 (2011)
15. Lin, Z., Chen, M., Ma, Y.: The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055
16. Cai, J.-F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**(4), 1956–1982 (2010)

A Least-Squares Ellipse Fitting Method Based on Boundary

Lei Liu(✉) and Xiangwei Meng

Department of Electronic and Information Engineering,
Naval Aeronautical and Astronautical University, Yantai 264001, China
ythyliulei@sina.com

Abstract. An ellipse fitting method based on boundary was proposed to estimate the relevant parameters of ship target in SAR image which can effectively solve the problem of the ship targets extraction and parameters estimation in SAR image. For the least-squares ellipse fitting method, all of the sample points on the boundary were involved in operation and causing deviation of the final results of ellipse fitting. For this kind of situation, a least-squares ellipse fitting method based on boundary was adopted. First extracting edge of the image and piecewise fitting ellipse. Then evaluating the ellipse and choosing suitable ellipse area for the target to be detected. The experiment shows that it is an effective method.

Keywords: Ellipse fitting · Least squares method · SAR image

1 Introduction

According to the need of modern high technology war and in order to fully grasp the details of the battlefield situation, Parameter information of the target has received more and more attention. The characteristics of ship wake are closely related to the movement characteristics of hull, so we can watch ships track or spectral characteristic to indirectly measure its characteristic parameters. The backscatter coefficient of ships on the radar image is a lot larger than ocean wave background and the pixels of the length and width of ship hull can be directly counted from SAR image. The size of the ship can be calculated based on the radar image resolution and the course can be directly determined from the track characteristics [1].

The ellipse is an important feature in real life. Elliptical extraction is a precondition for subsequent object recognition and measurement. It requests to achieve for ellipse extraction with robustness and accuracy. Basically there are three kinds of ellipse fitting method. They are ellipse fitting method based on HOUGH transform, invariant moment and the least squares method [2]. The least square method is an optimal estimation technology launched by the maximum likelihood when the random error obeys normal distribution [3–5]. It can make the sum of squares of the measurement error minimum, which is considered to be one of the most reliable methods to get a group of an unknown quantity from a set of measured value. Some literatures use the least square method to extract multiple elliptical targets in the images, but the effect of these methods is poorer in the actual image under complex background [6].

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 60–66, 2017.
DOI: 10.1007/978-981-10-3969-0_8

This article first introduced the basic way of the least squares ellipse fitting, and then introducing the boundary fitting method to propose a least square method based on boundary ellipse fitting method.

2 Ellipse Representation and Ellipse Fitting

2.1 Ellipse Representation

In the two-dimensional plane coordinate system, ellipse generally can be represented in two forms; one is a gen

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0 \tag{1}$$

The other is a more intuitive way which is expressed with five geometric parameters of the plane coordinate system. They are the center of the ellipse (x_c, y_c), the long half shaft and short half shaft (a, b), long axis angle $\theta(-\frac{\pi}{2}, \frac{\pi}{2})$, α and β. The arbitrary ellipse in the two-dimensional plane can be confirmed using the above parameters. The geometric meaning of the parameters is shown in Fig. 1. The two kinds of parameters representation can transform to each other through (2)–(10).

$$x_c = \frac{BE - 2CD}{4AC - B^2} \tag{2}$$

$$y_c = \frac{BD - 2AE}{4AC - B^2} \tag{3}$$

$$a = 2 \times \sqrt{\frac{-2F}{A + C - \sqrt{B^2 + \left(\frac{A-C}{F}\right)^2}}} \tag{4}$$

$$b = 2 \times \sqrt{\frac{-2F}{A + C + \sqrt{B^2 + \left(\frac{A-C}{F}\right)^2}}} \tag{5}$$

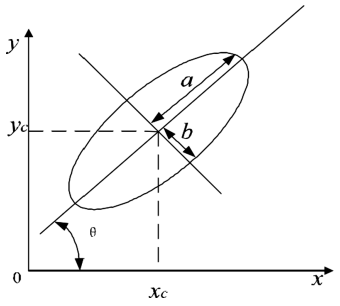


Fig. 1. Ellipse in the two-dimensional plane

$$\theta = \frac{1}{2}\arctan\frac{F}{A-C} \tag{6}$$

2.2 The Least Squares Fitting Ellipse

Assumes that the general form of the ellipse is as listed in (1), and the ellipse is expressed as two vector multiplication of implicit equation.

$$f(\alpha, X) = \alpha X = Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0 \tag{7}$$

In the type, α is (A, B, C, D, E, F) and the coefficient vector. $X_i = (x_i^2, x_i, y_i, y_i^2, x_i, y_i, 1)$. Because error $f(\alpha, X_i)$ in the point (x_i, y_i) is not zero, so $f(\alpha, X_i)$ is considered as algebraic distance from the point (x_i, y_i) to the implicit equation $f(\alpha, X)$. According to the principle of least squares, the curve fitting problem can minimize the sum of squares of algebraic distance to implement.

$$f(A, B, C, D, E, F) = \sum_{i=1}^{n} (Ax_i^2 + Bx_i y_i + Cy_i^2 + Dx_i + Ey_i + F)^2 \tag{8}$$

According to the extreme principle, to minimize the value of $f(A, B, C, D, E, F)$, there will be

$$\frac{\partial f}{\partial A} = \frac{\partial f}{\partial B} = \frac{\partial f}{\partial C} = \frac{\partial f}{\partial D} = \frac{\partial f}{\partial E} = \frac{\partial f}{\partial F} = 0 \tag{9}$$

This is a system of linear equations, then using algorithm for solving linear equations such as PCA gauss elimination and combining with the constraint conditions to obtain the value of the coefficient of equation A, B, C, D, E, F. The ellipse equation is obtained finally.

3 Ellipse Fitting Based on Boundary

Ellipse fitting by the least squares fit can only get a minimum error ellipse and it does not take into account the fitting degree of the ellipse and the original boundary. So it has a larger deviation between the ellipse and actual size and the algorithm results are often not satisfactory. In this case, it is necessary to evaluate and screen the fitting of the ellipse. In this paper, we proposed a least-squares ellipse fitting method based on boundary. First the SAR image was processed by regional elimination method based on level set segmentation and the boundary of the of ship targets in SAR image was extracted by the Canny operator. Then least squares method was used for the ellipse fitting of the target boundary [7].

Level set segmentation model (CV model) is a kind of flexible and effective image segmentation method which is suitable for regional uniform image segmentation [8].

The image slices in this paper can be divided into two parts; the target and the background. Energy equation is defined with the level set method and it can achieve segmentation by minimizing the energy functional.

The active contour (zero level set) is C while the image section I is divided into the target region $\Omega_{in} = \{x \in \Omega : \phi(x) > 0\}$ and background region $\Omega_{out} = \{x \in \Omega : \phi(x) < 0\}$, and it is homogeneous connected between areas. The domain is $\Omega \subset R^2$. The grayscale average is c_1 and c_2 respectively, and the energy function can be expressed as

$$E^{CV}(C, c_1, c_2) = \lambda_1 \int_{in} |I - c_1|^2 dx + \lambda_2 \int_{out} |I - c_2|^2 dx + v|C| \tag{10}$$

The first two are approaching, but $v|C|$ is constraints which represent the length of C and it is used to constraint the evolution of the curve. v, λ_1 and $\lambda_2 > 0$ is weight coefficient.

When $E^{CV}(C, c_1, c_2)$ in active contour C is in the boundary of the target, the type (10) can obtain the minimum value. The type uses the global information of the image, so it can get global optimal segmentation. Set ϕ is the level set function according to active contour C that is $\{C|\phi(x, y) = 0\}$. Further decomposing the type and the energy function can be expressed as

$$\begin{aligned} E^{CV}(\phi, c_1, c_2) = {} & \lambda_1 \int_{\Omega} |I - c_1|^2 H(\phi) dx + \lambda_2 \int_{\Omega} |I - c_2|^2 (1 - H(\phi)) dx \\ & + v \int_{\Omega} \delta(\phi) |\nabla \phi| dx \end{aligned} \tag{11}$$

$H(x)$ is 1d Heaviside function. $\delta(x) = H'(x)$ is Dirac function. Based on the euler - Lagrange equation, the partial differential equations of energy function can be obtained:

$$c_1 = \frac{\int_{\Omega} I(x, y) \cdot H(\phi(x, y)) dxdy}{\int_{\Omega} H(\phi(x, y)) dxdy} \tag{12}$$

$$c_2 = \frac{\int_{\Omega} I(x, y)[1 - H(\varphi(x, y))] dxdy}{\int_{\Omega} [1 - H(\varphi(x, y))] dxdy} \tag{13}$$

$$\frac{\partial \phi}{\partial t} = \delta(\phi)[v div(\frac{\nabla \phi}{|\nabla \phi|}) - \lambda_1 (I - c_1)^2 + \lambda_2 (I - c_2)^2] \tag{14}$$

The SAR image is binary processed after level set segmentation. Then the original image is transformed from gray image to binary image. In theory, every piece of binary area will correspond a goal. But because of various factors in the SAR image, there may be also a lot of noise, burr and small area. At the same time, the ship targets

are also existed as small white block areas in the sea. These small block areas need to be removed for subsequent process.

Region elimination can be realized by region merger. This method calculates the area of each region in segmentation image and the area which is less than a certain threshold will be contained to its surrounding regions. Through region merger, the number of the original binary image area decreased significantly and the noise of the small particles were eliminated; leaving only the target area. It brings great convenience for the target detection. Algorithm is described as follows:

(1) Target active contour for ships is obtained through the level set segmentation,.
(2) The grayscale image is binary processed.
(3) Label white connected area by scanning the whole image.
(4) Calculate each connected component of the white area and set the threshold value for S1 through the comparison of ship target area size.
(5) Remove the area which is less than S1 and get the binary image after region elimination.

The important parameter that can describe an ellipse is its semi-major axis a and short half axis b. The ellipse can be filtered by defining the ellipticity ρ and ellipse area difference.

In mathematics, ellipticity is defined as:

$$\rho = b/a \tag{15}$$

Ellipticity describes the extent of the ellipse tend to round. The size of the ellipse can be determined through the limitation on the length axis and the short axis. The area difference is:

$$\Delta Area = \left| Area_0 - Area_f \right| \tag{16}$$

$Area_0$ is the ideal ellipse area, $Area_f$ is the ellipse fitting area. Finally ellipse of minimum area difference will be chosen. The ellipse which has a larger area will be chosen as the optimal ellipse if the size of ellipse is not specified.

Algorithm principle is as follows:

(1) Carve up the target image by level set method; use the method of region elimination to deal with the image.
(2) Canny is utilized to extract edge image.
(3) If the number of boundary point is greater than the threshold, using the least squares fitting ellipse; otherwise choose the boundary again.
(4) Calculate the fitting rate and ellipticity.
(5) If the ellipticity is greater than the threshold and the fitting rate is greater than T_d, recording the oval ellipse E.
(6) Calculate the area difference between ellipse E and ideal ellipse. Calculate the area difference according to the formula (16).
(7) Execute steps (2)–(5) circularly and select the minimum area difference ellipse.

4 Experiment and the Results

Figure 2 is the original SAR images of ship target slices. Figure 3 is the images after level set segmentation and region elimination. Figure 4 is the ship target outline cut by Canny operator. Figure 5 is the minimum circumscribed ellipse by least squares.

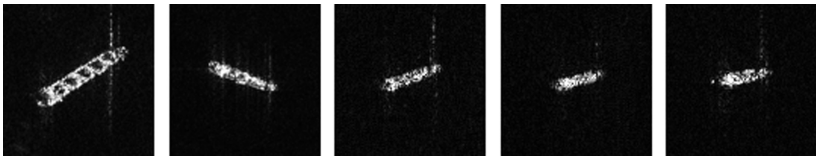


Fig. 2. Original SAR images of ship target slices

Fig. 3. Target slices based on level set segmentation and region elimination

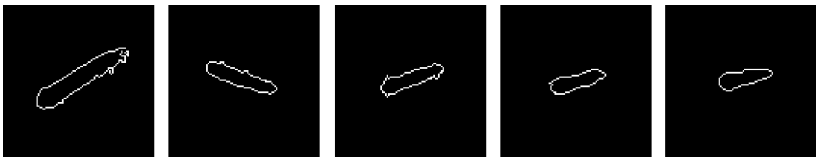


Fig. 4. Ship targets outline cut by Canny operator

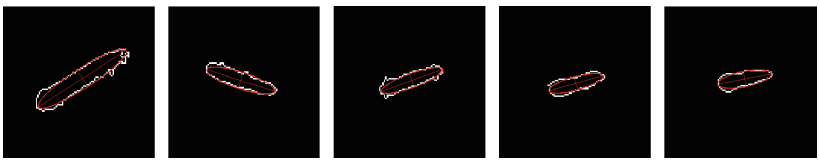


Fig. 5. The minimum circumscribed ellipse by least squares

The long axis of the ellipse corresponds to the length of the ship targets in SAR image and the short axis corresponding to the width. The area of the ellipse is the area of the target in the image. At the same time the ship course angle can be obtained by calculating the angle of the long axis and the horizontal axis. The experimental results are shown in Table 1:

Table 1. Ship target parameter estimation (resolution: 3*3 meters per pixel)

	Length	Width	Course angle(°)
1	67.3838	11.9309	32.1404
2	48.7097	10.0360	162.0453
3	42.8009	9.0298	20.0981
4	36.9080	8.3273	15.1051
5	35.2310	9.8931	11.3279

5 Conclusion

This paper studied the estimation problem of the parameters of ship targets in SAR image. Conventional least square ellipse fitting will use all sample points which will obtain an excessive fitting ellipse, and the error is bigger. It can't meet the needs of parameter estimation of ship target. The experiment has proved that the algorithm proposed in this paper eliminated the ellipse which does not conform to the requirement in the process of calculating the optimal ellipse. The practicability of the algorithm is enhanced and the accuracy is improved at the same time. The running time is shorter, so it can fully meet the requirements of parameter estimation. It has important significance to use the ellipse fitting method to estimate the parameters of the ship targets in SAR image.

References

1. Fitzgibbon, A., Pilu, M., Fisher, R.B.: Direct least square fitting of ellipses. IEEE Trans. Pattern Anal. Mach. Intell. **21**(5), 477–480 (1999)
2. Hough, P.V.C.: Method and Means for Recognizing Complex Patterns. USA, 3069654[P], 20 June 1962
3. Gander, W., Golub, G.H., Strebel, R.: Least-squares fitting of circles and ellipses. BIT Numer. Math. **34**(4), 558–578 (1994)
4. Ahn, S.J., Rauh, W., Cho, H.S.: Orthogonal distance fitting of implicit curves and surfaces. IEEE Trans. PAMI **24**(5), 620–638 (2002)
5. Halir, R., Flusser, J.: Numerically stable direct least squares fitting of ellipses. In: Proceedings of the Sixth International Conference in Central Europe on Computer Graphics and Visualization, pp. 125–132. Citesser, Czech Republic (2008)
6. AIzahrani, F.M., Chen, T.: A real-time edge detector: algorithm and VLSI architecture. Real-Time Imaging **3**(5), 363–378 (1997)
7. Awrangjeb, M., Lu, G., Fraser, C.S.: A fast corner detector based on the chord-to-point distance accumulation technique. In: Digital Image Computing: Techniques and Applications, pp. 519–525 (2009)
8. Touzi, R.: A review of speckle filtering in the context of estimation theory. IEEE Trans. Geosci. Remote Sens. **40**(11), 2392–2404 (2002)
9. Hu, H., Zhu, J.: Direct least squares ellipse fitting based on arc group. J. Hangzhou Normal Univ. **10**(6) (2011)
10. Wu, G., Hu, X., Zhang, L.: Application of weighted total least-square adjustment to the versine fitting of the railway curve. Comput. Eng. Appl. **50**(1), 232–234 (2014)

Training Convolutional Neural Networks Based on Ternary Optical Processor

Ruifen Zhang[1,2] and Shan Ouyang[1,2(✉)]

[1] School of Computer Engineering and Science,
Shanghai University, Shanghai, China
zhruifen@163.com, {zhangruifen, ouyangshan}@shu.edu.cn
[2] Computer Science Building, Shanghai University,
99 Shangda Road, Baoshan District, Shanghai City, China

Abstract. A novel platform and algorithms of Ternary Optical Computer (TOC) are proposed to training Convolutional Neural Network (CNN). It can significantly improve the concurrency and throughput of the training process of CNN. Reviewing the irrelevance data and the inherent parallelism module of the CNN, this paper discusses the preprocessing way of arbitrary number of two-dimensional data which include feature maps, convolutional kernels and mini-batches. Then strategies of parallel training of CNN based on the reconfigurable flexible arithmetic operator are proposed. All these arithmetic units are implemented by the optical Modified Signed Digit (MSD) adder and optical MSD multiplier, which are carry-free differing from the electronic computers. The massive data-bits of TOC are reconfigurable and redistributable, so fully parallel pipeline of the CNN can be sufficiently achieved. The computational complexity of the algorithms in time are analyzed. The result shows that TOC has great benefits comparing to the GPU and FPGA in concurrency, needed cycle and hardware resources resumed. This paper provides a new perspective to efficiently address computation-intensive and data-intensive issues.

Keywords: Massive data-bits · Reconfigurable and redistributable processor · Parallel processing · Convolutional Neural Network · Ternary Optical Computer

1 Introduction

The Convolutional Neural Network (CNN) introduced by Le Cun, is a multilayer neural network trained with the back-propagation algorithm to learn complex, high-dimensional, non-linear mappings from large collections of labeled samples [1]. In order to leverage the CNN ability to training data in the image and voice filed, greater complex models are required and rich training data are needed to avoid overfit. Therefore, parallel method to training CNN in a feasible amount of time and resources is a trend at present. A mixed high-end, analog-digital processor ANNA chip was first displayed, which could compute 64 8 * 8 sized convolutions by multiply-accumulate operators simultaneously [2]. Mapping the iterative convolution kernel to the restricted units of Field Programmable Gate Array (FPGA) can speed up the training process too. Its performance is decided by the execution efficiency of the convolution kernel units

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 67–75, 2017.
DOI: 10.1007/978-981-10-3969-0_9

and the design of the internal memory caches. Also a common alternative parallelism hardware solution to accelerate the application is Graphics Processing Units (GPU) [3], the model structure and the data structure of CNN contribute two aspects to analyze parallel processing. Distribute the training data like samples, weight matrixes and intermediate parameters to the several fixed-module GPU blocks, or asynchronous sharing database to accomplish parallel processing.

The Ternary Optical Processor (TOP) retains many crucial advantages benefited from it is a kind of structured computer: own ten millions extendable usable data-bits, reconfigurable and redistributable processor, and MSD adders with no carry and so on [4–7]. TOP is reconstituted as a structured processor in real time which contains many different isolated arithmetic units according to requirements of the users. Ten millions data-bits simultaneously manipulated within single executive instruction for different independent functions. Taking full advantage of CNN characteristics like multi-dimensional data, dense internal representations, two-dimensional convolution detectors, and iterative filter windows, combine the above unique advantages of the TOP can greatly improves parallelism performance of CNN. Optical algorithms for reconfigurable flexible convolution detector and sampling operator are discussed in detail. Fully pipeline optical training system for CNN is introduced briefly. The performance of optical method is presented in every perspective. This conceptual basis is only a first step towards a full parallelism implementing of Convolutional Network properties.

2 Related Work

2.1 Convolutional Neural Network

A typical application of CNN is handwritten recognition which is proposed by Le Cun. The vector sets of size-normalized images and corresponding ideal outputting class indices feed into the network. Data be trained through a cascade of linear operators and non-linearities. The hidden layer has two closely sub layers: convolution layer and sub-sampling layer, and it has two crucial properties: dense local receptive field and sharing weight. In the fed-forward propagating phase, output is obtained by calculating the every function between the input and the weight matrix in each layer. Sharing weights significantly reduce the number of training parameters and the complexity of the network. The latter tasks exactly give special focus on this phase.

2.2 High-Speed Hardware Used to Training CNN

Cude-convnet algorithm is a classical parallel solution for training CNN which utilizes the CUDA framework to program GPUs. It assigns the extract feature computations and the rectifiers in the same layer to the different GPU blocks to achieve the parallel execution of the data. Usual hardware designs based on FPGA needed to be changed cause any small change, that is absolutely non-flexibly. Efficient implementation of ConvNets on a low-end DSP oriented FPGA proposed by Farabet [8] which has been

applied in a face detection system fixed the problem of flexibility. It uses a single FPGA with an external memory module to train a 340 million connection network by a sequence of instructions for the ConvNet Processor (CNP) which consumes less than 15 W.

2.3 Ternary Optical Processor

TOC is an optical-electric hybrid computer which has the ability to tackle with structured data. It uses three optical states (horizontal polarization, vertical polarization and darkness) to present information. MSD adder uses a redundant value of the three ones to accomplish carry-free addition, by carrying out the three-valued logic transformation for each bit of the two operands [6]. It only costs three clock cycles for calculating the sum of regardless of the bits of two operands, and so achieves a fully parallel addition with on-carry. The basic implementation of MSD multiplier of TOC is as following: Firstly, the multiplier manipulates the M transformation to each bit of the operands to generate the partial product, then shift the partial product, finally uses MSD adder to conduct binary iterative addition to the partial product to get result. The n-bit MSD multiplier is implemented through n q-bit adders, meanwhile, q = n + s, the sample s is the additional bit ensuring the precision, and the multiplying process costs $3 \times \lceil \log_2 n \rceil + 2$ clock cycles.

3 Training Convolutional Neural Network Based on TOP

Parallel processing situation naturally exists during the fed-forward propagating phase of training CNN, which is pointed out by many researchers, and this is precisely the focus we will use optical method to accelerate. The parallel implementation involves two iterative sub-processes: convolutional operation and subsampling operation. They are aimed at obtaining the different representations of the samples, such as the presence or absence of edges at particular orientations and locations of image. The input data sets of convolution layer iterate on linear operators with a plurality of multi-scales weight matrixes. The sub-sampling operation is used to reduce the resolution of feature maps by calculating the partial average or maximum of mini-batches, which makes the CNN having certain tolerance for partial transformation. The kernel learning algorithm for convolution and sub-sampling will be introduced in first and second subsection. Fully pipeline parallel action of CNN is the dream of CNN training, explained in third subsection.

3.1 Implementing Convolution Based on TOP

First we fix some notations for convolution layer: Using a small Convolutional kernel W[m][m] to do feature map extraction on the $n \times n$ sized feature map S[n][n] with p sliding steps. Then get a result feature map C[g][g], and its dimension $g = \lceil (n - m)/p \rceil + 1$. We assume the images and kernels are in square for simplicity of notation, and later work trivially extended to non-square images and kernels. In the

forward phase the feature map extraction operators use the way of the 2-dimensional discrete convolution (*) and followed by non-linearities as in (1).

$$c_j = f_1\left(\sum s_i * w_{ij} + b_i\right). \tag{1}$$

f_1 denotes the non-linearity activation function. s_i is one of the input feature maps and means a 2-D vector of dimensions $n \times n$. w_{ij} is a trainable convolution weight kernel (filter bank). b_i is an offset coefficient, and c_j is the result feature map convoluted from s_i. The unique advancement of the TOP is that it can manipulate ten millions data bits once in two liquid crystal response time $2T_c$ [9]. By preprocessing the training data and convolution kernels to the correct structured data, we can act large amounts of multi-scale data through general data path of processor. As shown in Table 1, we transform the 2-D feature map S[n][n] into a linear discrete sequence ancillary data along stride p. This ancillary data contains g^2 blocks in $m \times m$ sized match to g^2 partial receptive fields. These values of each component are exactly mini-batch pixels which are uncorrelated with each other. The same goes for decompose the convolution kernel W[m][m] into a linear discrete sequence and each component is equal to the values of component 1, as shown in Table 1. That means g^2 components are assembled as a structured data, and components of ancillary data with respect to each other too in a linear space logically. We assume that optical q data-bits represent one pixel-information.

Table 1. Ancillary data of S[n][n] and W[m][m].

Ancillary data of S[n][n]						
Component 1	m^2	...	$m + 1$	m	...	1
	s[m − 1] [m − 1]		s[1][0]	s[0][m-1]	…	s[0][0]
Component 2	$2\ m^2$	...	$m^2 + m + 1$	$m^2 + m$	...	$m^2 + 1$
	s[m − 1] [p + m − 1]	…	s[1][p]	s[0] [p + m − 1]	…	s[0][p]
…						
Component g^2	g^2m^2	...	$(g^2 - 1)$ $m^2 + m + 1$	$(g^2 - 1)$ $m^2 + m$	...	$(g^2 - 1)$ $m^2 + 1$
	s[n − 1] [n − 1]	…	s[n − m + 1] [n + m-1]	s[n − m] [n − 1]	…	s[n − m] [n − m]
Ancillary data of W[m][m]						
	m^2		$m + 1$	m		1
Component 1	w[m − 1] [m − 1]	…	w[1][0]	w[0] [m − 1]	…	w[0][0]
Component 2	w[m − 1] [m − 1]	…	w[1][0]	w[0][m-1]	…	w[0][0]
…						
Component g^2	w[m − 1] [m − 1]	…	w[1][0]	w[0] [m − 1]	…	w[0][0]

There are four crucial serial steps to training convolution layer based on TOP:

Step 1: Compute the ancillary data S with the corresponding weight matrix W, outputting the partial product $S'[g^2 \times m^2]$.

TOP puts $g^2 \times m^2$ MSD multipliers together, and each has q data-bits. The $g^2 \times m^2$ 1-D vector of feature maps and kernels will respectively feed into the input of the main optical path and control optical path of TOP, meanwhile every pixel of the ancillary data corresponding to a q data-bits MSD multiplier. All components of structural data are parallelism processed in single instruction, and sharing one cycle of MSD multiplier as in (2).

$$t_1 = (3 \times \lceil \log_2 \mathrm{q} \rceil + 2) \times 2T_c. \tag{2}$$

An alternative strategy is that the TOP only reconstruct one m^2*q data-bits MSD multiplier. The g^2 components of feature map structured data sequentially as the input of the main optical path. Component 1 (convolutional kernel) is as the input of the controlling optical path and it will remain unchanged during the g^2 operations. Meanwhile every component of the ancillary data are matched to m^2 * q data-bits MSD multiplier. So that, each operation only increases the main optical path response time of liquid crystal. The cycle of these g^2 operations is as shown in (3).

$$t_1^{'} = (3 \times \lceil \log_2 m^2 \times \mathrm{q} \rceil + 2) \times T_c \times g^2 + T_c. \tag{3}$$

The first method sacrifices the bit space for shortening the consuming time and each operation cycle processes as much data bits as possible. The second method however sacrifices the time for reducing space usage and uses as few bits data module to processing the same functions of the large amounts data. For the two methods above, the users should balance the data bits and the operating cycle to getting the proper solution in practical situations.

Step 2: Iteratively process the intermediate result $s'[g^2 \times m^2]$ with binary tree method, and getting the sum $s''[g^2]$.

Treat each m^2 data of sequence $s'[g^2 \times m^2]$ as a component logically. Put components and its copies respectively into the main optical path and the control optical path. TOP will calculate the inner sums of each components for $\log_2 m^2$ times. At per iteration, TOP will reconfigure $g^2 \times \lfloor m^2/2^i \rfloor (i = 1, \ldots, \log_2 m^2)$ MSD adders with q-bits, make g^2 components manipulated at same time. The occupation of the data bits will decrease exponentially with the increasing of cycles. The total iterative cycle is t_2, as in (4).

$$t_2 = 2T_c \times 3 \times \lceil \log_2 \mathrm{m}^2 \rceil. \tag{4}$$

Step 3: Add offset coefficient b_i to $S''[g^2]$.

Copy b_i for g^2 times mapping to each value of $S''[g^2]$. TOP assemble g^2 q-bits MSD adder and complete the operations in $3 \times 2T_c$ cycles.

Step 4: Conduct the non-linear transformation to the last product and get the final result $C[g^2]$ which is the pixel values of the feature map C[g][g].

The key insight is that the data-bits of the convolution layer are verily flexible and scalable according to the real situation. Imagining that using 2304000 (24 * 24 * 25 * 16 * 10) data-bits partial of ten-millions data-bits of TOP can simultaneously produce 10 representations in 24 * 24 sized in one convolutional layer of MINIST which data set are 28 * 28 sized, if use 16 data-bits represent one pixel information. And no-carry-delay MSD adder and fast-calculation MSD multiplier are shining light on these procedures required cycles.

3.2 Implementing Sub-sampling Based on TOP

First we fix some notations for sampling layer: Sampling layer get the representation S[h][h] from the previous layer by conducting no overlapping sub-sampling operations. $k \times k$ sized mini-batches slide along the feature map C[g][g]. In the symbol S[h][h], $h = \lceil g/k \rceil$. The subsampling is as shown in (5).

$$s_j = f_2(\beta_i \times down(c_i) + b_i). \tag{5}$$

In (5) the f_2 is the rectified linear unit, such as ReLU, which is a simple half-wave rectifier $f(z) = \max(z, 0)$. c_i is the pixel of previous sub-layer. β_i is the weighting coefficient. b_i is a bias. s_j denotes the product. *down* () is the down-sampling function.

The way of preprocessing sampling layer data is similarity with the method of organizing convolutional parameters. C[g][g] is transformed into an auxiliary data sequence that is shown in Table 2. This discrete sequence composes h^2 blocks which are unrelated components. Each component covers one subsampling mini-batch as the no overlap style.

Table 2. Ancillary data of C[g][g].

Ancillary data of C[g][g]						
Component 1	k^2	...	$k + 1$	k	...	1
	c[k − 1] [k − 1]		c[1][0]	c[0][k − 1]	…	c[0][0]
Component 2	$2\ k^2$	...	$k^2 + k + 1$	$k^2 + k$	...	$k^2 + 1$
	c[k − 1] [2 k − 1]	…	c[1][k]	c[0][2 k-1]	…	c[0][k]
…						
Component h^2	h^2k^2	...	$(h^2 - 1)$ $k^2 + k + 1$	$(h^2 - 1)$ $k^2 + k$	...	$(h^2 - 1)$ $k^2 + 1$
	c[g − 1] [g − 1]	…	c[g − k][0]	c[g − k + 1] [k − 1]	…	c [g − k + 1] [0]

The following steps are computations of subsampling unit based on TOP.

Step 1: Calculate the *down*() function produce the intermediates $c'[h^2]$.

1-D auxiliary discrete data obtains h^2 components. Take it and its copy respectively into the main optical path and the controlling optical path. Calculate the internal sums of each component for $\lceil \log_2 k^2 \rceil$ cycles. In each cycle, the TOP need to put $h^2 \times \lfloor k^2/2^j \rfloor (j = 1, \ldots, \log_2 k^2)$ MSD adders with q-bits together to calculate the intermediate values using iteration binary tree.

These components share one operation cycle. The total cycle is (6).

$$t_3 = 3 \times 2T_c \times \lceil \log_2 k^2 \rceil. \tag{6}$$

Step 2: Adjust β_i to $c'[h^2]$ produce partial result $c''[h^2]$.

h^2 copies of β and $c'[h^2]$ are handed in data path of processor which contains h^2 MSD multiplier with q-bits at this time. This reform operation will be completed in one multiply instruction and its cycle is (7).

$$t_4 = 2T_c \times \lceil \log_2 h^2 \rceil. \tag{7}$$

Step 3: Plus the bias b_i to $C''[h^2]$ in one MSD adder cycle.

Step 4: Decoder conducts non-linear transformation function f_2, and produces the output feature map S[h][h].

In conclusion, this bit-flexible sampling basic unit can be easily used by bigger scale images.

3.3 Full Parallel Pipeline Strategy for CNN

Sequence-to-sequence learning with neural network gives the pipeline feature for CNN and all output information of the previous layer will be the input of the next. A simple CNN is built to show outperform in this section which implemented based on TOP. The principle of design the layer after layer pipeline will be shown.

The architecture has one input layer, two hidden layers, one full connection layer and output layer. Every hidden layer is C-S structured. Figure 1 shows a parallel pipeline design with three samples and each one is marked as A^i. The solid line mean the input of the main optical path and the dotted line denotes the controlling optical

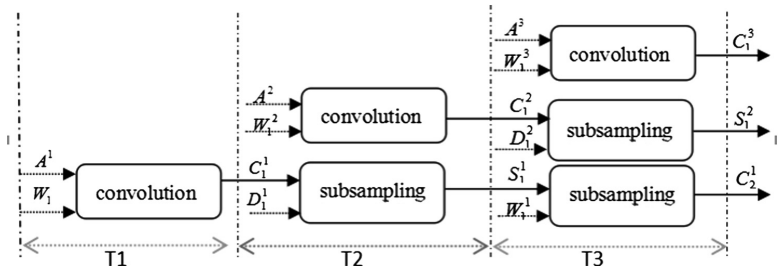


Fig. 1. Illustration of pipeline parallelism processing sample in TOP

path. C_j^k denotes the feature map of convolution layer. W_j^k is convolution kernel. S_j^k represents feature map of sampling layer. D_j^k means the neighbouring window. The index k and j represent the number j layer of training parameters in the number k sample set.

Every bit of processor is reconstructed according to the function during each pipeline cycle. Thus processor becomes a combination of many complex units. For example, the TOP is configured as convolution unit to meet the demand bits of (A^1, W_1^1) during the first cycle T1. Then convolution operator produces the representation C_1^1 of the first layer which is exactly the input of the cycle T2. TOP is reconfigured as a synthesis of convolution unit and sampling unit to simultaneously complete the convolution of (A^2, W_1^2) and sampling operations of C_1^1, D_1^1 within cycle T2. It outflows the results C_1^2 and S_1^1 to the next cycle. Calculations of cycle T3 involve different action in different layers for sample A^1, A^2, A^3. The rest can't be done in the same manner until complete the full-connection operations and obtain the final result.

The Sects. 3.1 and 3.2 clarify the preprocessing method which makes feature maps into a discrete structured data according to patch size. This approach ensures the ten millions-bits of the TOP having a unified data input channel. In this, structured data of feature maps compound to a larger but simple structured data. The larger structured data are passed to TOP through the general data channel. Meanwhile each component is mapped to corresponding data bit of the platform. Then all these structured data will be processed under a single TOP instruction.

4 Result

A general-purpose algorithm of accelerate CNN training with TOP is proposed, which yield over several orders of magnitude compared to existing state-of-the-art implementations, such as modern GPU, FPGA.

The data-bit of scalar processor is fixed and an operation instruction can only handle a single pixel. FPGA develops some arithmetic units whose data-bit is fixed. Designed multiply-accumulate computes one kernel data in each processing cycle, and to detect result score by reusing the units. A key advantage of CNN training on TOP is that massive pixel can be simultaneously processed. Considering the following scene: we use 16-bits MSD value to represent a pixel value of the MNIST handwritten digital image. In the task of performing 6 convolutions with 5 * 5 kernels on the input image size of 28 * 28, comparing with the FPGA only handling pixels in one window (5 * 5 * 16 bits) within a cycle, the TOP completes a bit calculation of 24 * 24 * 5 * 5 * 6 neurons (24 * 24 * 5 * 5 * 6 * 16 bits) only in a liquid crystal response period.

Considering the parallelism degree of processing feature map and pixel, TOP provides a strong force. GPU achieves a kernel by completing a convolution operation thread, and the modern GPUs check a full feature map or some kernel sized maps by using the way of cross accessing sharing memory. TOP however can flow out one or more complete feature map in one convolution operation cycle.

It is worth mentions that comparing the utilization rate of the hardware resources, the FPGA design computing unit with limited hardware gate array and a 5 * 5 sized convolution unit even consumes about 73% of the DSP resources of the Virtax6 (SX475T). TOP can be successfully reconstructed as a complex unit in one operation cycle in real-time and this unit include thousands 5 * 5 sized modules. Therefore, the data processing ability of the optical processor is several orders of magnitude to FPGA in theory.

5 Conclusion

Simple flexible and configurable convolution operator and sub-sampling unit were designed benefit from the TOP which can deal with rich structure data, has no-carry-delay MSD adder, and reconfigurable and redistributable processor. Technique of pre-process 2-D vector into 1-D discrete structure data is demonstrated for training convenience. Certainly, series of instructions are analyzed. These approaches do extendable to achieve concurrency such like CNN compute-intensive and data-intensive training models. Although conceptually straight forward, a number of challenges relating to TOP implementation needed to be addressed. In future work the new effective algorithm will be adopted to accelerate training and inference, such as Fourier domain.

References

1. Le Cun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: Advances in Neural Information Processing Systems (1990)
2. Sackinger, E., Boser, B., Bromley, J., LeCun, Y., Jackel, L.D.: Application of the ANNA neural network chip to high-speed character recognition. IEEE Trans. Neural Netw. **3**(2), 498–505 (1992)
3. Uetz, R., Behnke, S.: Large-scale object recognition with CUDA-accelerated hierarchical neural networks. In: IEEE International Conference on Intelligent Computing and Intelligent Systems, ICIS 2009, vol. 1. IEEE (2009)
4. Yi, J., Huacan, H., Lü, Y.: Ternary optical computer architecture. Phys. Scr. **T118**, 98 (2005)
5. Jin, Y., Ouyang, S., Song, K., Shen, Y.F., Peng, J.J., Liu, X.: Management of many data bits in ternary optical computers. Sci. Sin. Inf. **43**, 361–373 (2013). doi:10.1360/112012-260
6. Shen, Y.F., Pan, L.: Principle of a one-step MSD adder for a ternary optical computer. Sci. China Inf. Sci. **57**(1), 1–10 (2014)
7. Jin, Y., et al.: Principles structures and implementation of reconfigurable ternary optical processors. Sci. China Inf. Sci. **54**(11), 2236–2246 (2011)
8. Farabet, C., Martini, B., Akselrod, P., Talay, S., LeCun, Y., Culurciello, E.: Hardware accelerated convolutional neural networks for synthetic vision systems. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems, pp. 257–260. IEEE, May 2010
9. Shen, Y., Pengfei, H., Wang, H.: The computational complexity of arithmetic based on ternary optical computer. J. Inf. Comput. Sci. **8**(5), 850–857 (2011)

An Improved Algorithm for Video Abstract

Jianlei Zhang, Qin Li(✉), Wenfeng Shen, and Shengbo Chen

School of Computer Engineering and Science,
Shanghai University, Shanghai 200444, People's Republic of China
519832998@shu.edu.cn, 958583805@qq.com

Abstract. In this paper, we study the foreground object extraction, the trajectory extraction, the trajectory combination optimization and other key technologies of the surveillance video abstract generation technology. Put forward a kind of improved algorithm, through the pre-processing based on Focus Stacking, foreground object extraction of the foreground object shadow removal, the trajectory synthesis optimization. As a result, the foreground extraction accuracy rate was increased by **2.78%**; because of shadow removal of the foreground object, the collision rate of the foreground object in the synthesized video is reduced by **15.77%**; The use of Semi-Transparent Handling Collision (**STHC**) makes the trajectory of the foreground object is not interrupted, the video frame information is not lose and the compression rate is increased by about **10%**. The algorithm is applied in this paper, and the optimization effect is observed through the whole system test. As a result, the clarity of the synthesized video is increased, the integrity of the video's information is enhanced, and the compression rate of the video is improved.

Keywords: Video synopsis · Focus Stacking · Shadow remove · Trajectory synthesis

1 Introduction

In recent years, with the increase of people's demand of social public security, intelligent video surveillance technology [1] has been widely developed. Major public places are installed a large amount of surveillance cameras, forming a wide range of video surveillance network, to improve the security system of detection and alarm level, so that people can adjust and respond to emergencies according to their needs.

A large number of cameras record real-time video data. But the data utilization rate of these massive video is extremely low, because viewer mainly through the playback of the original video data to find the content of interest, which consumes a considerable time. Therefore, how to accurately find the interesting video clips in massive digital video has become the urgent needs of the industry and the great challenge.

Only within the scope of a camera monitor, if you want to know exactly what occurred within **24** h of the record, you need to watch the events section of the surveillance video from the beginning to the end. If only rely on manual to watch the video to find clues, it is bound to consume a lot of time. At the same time, some important clues and activities in videos may only appear in the monitor screen for a few seconds, so the use of artificial view is very easy to miss these important information

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 76–89, 2017.
DOI: 10.1007/978-981-10-3969-0_10

scattered in dozens of hours of video data. According to the IMF study, we can draw the conclusion that if the viewer watches the video continuous for **12** min, will miss the **45%** of the scene information; if the viewer watches the video continuous for **22** min, will miss the **95%** of the scene information [2]. So many fleeting important information will be neglected. Due to the traditional video browsing is time-consuming, high workload and low efficiency for browsing problems, so most of the monitoring video data almost never be browsed or reviewed. It is not active and efficient to respond to the state of public security in the monitoring area, only relying on Artificial back to browse after the occurrence of the incident.

Generally, people browse a large number of video with some purpose, and focus on a part of the key information in the video. When using surveillance video to find and analysis of crime, people tend to be more concerned about the picture of the active object in the surveillance video. But there are often a lot of "static" picture in the surveillance video. Although commonly used video player has quick forward function, but does not distinguish between static parts and activities in video, rather than frame skip constantly, leading to skip moving pictures that contain important clues.

So, how to condense a long video file into a short summary video, that viewers can analyze the abnormal events and potential hazards only through browsing the brief summary of the video quickly so as to shorten the response time, improve the efficiency from the accident to find tracking and assist emergency response personnel to deal with the incident in a timely manner.

With the increasing demand for video data processing and the increasing amount of video data, people want to be able to build a summary for a long period of video to browse it quickly and facilitate better use of it. The way to realize it is a hot and difficult point of the digital video technology: video abstract [3].

Video abstract is a summary through analyzing the structure and content of the video, extracting meaningful information from the original video, and recompose the meaningful information to condense to video semantic content which can be fully expressed. Video abstract is a summary of a static image or dynamic video sequences of long video content. Video abstract can be provided people for a summary of the main content of the original video, but its length is much shorter than the original video that it has more concise information. Video abstract technology can be condensed the dozens of hours of video for ten minutes even less according to user requirements to shorten the time of video browsing. Besides, with the technology of video object feature retrieval, video abstract can be used to help people to analyze the semantic characteristics of the underlying video and quickly locate the target of interest. Video abstract and retrieval technology can help users to fully excavate the meaningful information in the massive video surveillance video, and to improve the analysis and response efficiency.

Video abstract can analyze and deal with the video through intelligent video technology (including motion detection, the human body detection, vehicle inspection, the sensitive area, the sensitive area, wandering detection, cross the line detection, etc.). Removing still frame, invalid and redundant information in the video, extracting the video segments contain valid information and saving video information effectively to the database can let In-depth analysis, query and management; And then, re-planning and organization of these effective information in time, space, density to constitute a shorten time of the video file. Video files can be used for regular broadcast control, and

people can click on the moving objects in the video. Smooth switch playback controls between synopsis video files and original video can greatly reduce the time of watching video, and allow user to focus on key information, improving the efficiency of video browsing.

In this paper, the technology of video abstract that consist of foreground extraction module based on **ViBe** [4], trajectory extraction module, trajectory optimization module was analyzed and studied. Three optimization methods are proposed: first, pre-process for the original video based on **Focus Stacking**; second, shadow removal of foreground object; third, proposed STHC algorithm. Results of this paper can be directly used in video abstract browsing under video surveillance system.

2 Related Work

The goal of video abstract system is to help users to browse video surveillance more effectively and view any sports events in just a few minutes. It can take all events that occur within **24** h in the form of condensed video clips and fully displayed in just a few minutes. Video synopsis can present multiple objects moving and events occurring at the same time, however they appeared at different times in original video that can be displayed by clicking on any object or event of the fragment.

According to the different forms of video summary results, video synopsis can be divided into two categories: static video abstract and dynamic video abstract.

2.1 Static Video Abstract

Static video abstract [5], is a series of static semantic units extracted from the original video stream to represent the video content. The static semantic unit can be a key frame, title, slide, etc., which can be summarized to express the static characteristic information of the surveillance video.

At present, static video abstract research is mainly based on key frame selection method. Key frame also known as storyboards, is extracted from the original video to represent main information. This method allows users to quickly browse the contents of the original video through a small number of key frames and provide rapid retrieval [6].

Classic key frame selection algorithm [7–10] mainly uses the color, motion vector and other visual features to distinguish the difference between the frames. However, the difference of the frames depends on the choice of threshold.

The video abstract based on the key frame has the advantages of simple structure and easy browsing. But video content is easy to lose and the compression rate of video abstract is very low.

2.2 Dynamic Video Abstract

Dynamic video abstract [11], is a much shorter video than the original video. It is by extracting the object from the video frame and then re combining these objects into a new video, so it is also called video abstract based object.

Video abstract object based is put forward in recent years and mainly used in surveillance video field [12]. The object extraction include background modeling, moving target detection and tracking technology and visual analysis technology. The recombination of the object is determined by user attention degree, compression ratio, multi video fusion and other factors [13–17]. This method can effectively preserve the characteristics of video content with time, reduce the redundancy of time and space. But object extraction is difficult and it is difficult to solve the problem of video abstract generation in complex scenes.

At present, video abstract technology in the field of video surveillance mainly focuses on these two kinds of forms: static video abstract based on key frames and dynamic video abstract based on object. Two can greatly shorten the length of video, to facilitate the video viewing, analysis and retrieval. The minimum unit of video summary based on the key frame is "frame", which is smaller and easier to transmit, but it can't represent the motion track of each target, which is not conducive to video object retrieval. The video skimming based on object of smallest unit is "object", can maximum reduce the temporal redundancy and spatial redundancy information, and video retrieval upper development provide object structure, can quickly respond to emergencies in the security monitoring, locate the events related to the "object", but there is a complex processing, generation problem of video skimming.

In this paper, the algorithm is improved based on the existing object based video abstract, which makes it possible to generate a high quality, not lost object information, high compression rate of synopsis video. Main research contents: first, research video pre-processing algorithm and use Image Fusion algorithm to make the video more clear; second, proposed a shadow removal algorithm for optimize foreground extract to make the foreground extraction more accurate; third, proposed **STHC** algorithm to solve trajectory collision problem and improve the compression ratio of video abstract.

3 Improved Video Abstract Algorithm

In order to achieve the high quality of picture, little lose of the object information and high compression rate of video abstract, we will start with the following aspects, such as Fig. 1.

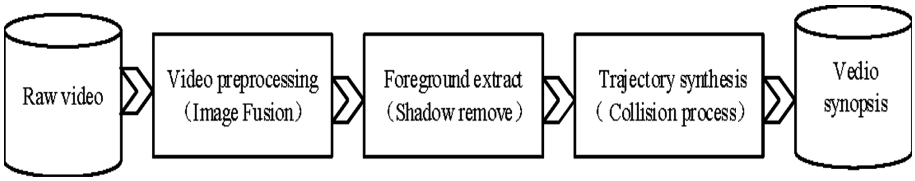


Fig. 1. Video synopsis processing

A. The Method for Improving Foreground Extraction Accuracy of Video Abstract

When using surveillance camera, the video will inevitably have jitter and illumination change. Result in the video footage is not clear, fuzzy and exposure, etc. In order to solve this problem, we proposed image fusion algorithm based on Focus Stacking [18–23].

We use the following algorithm to achieve this goal.

Definition 1. *Contrast: Using Laplacian filter to filter the gray image, will produce a contrast index* ***C****. It is assigned a weight that represents the edge or texture of an image.*

Definition 2. *Saturation: In order to make the image look clear and define a saturation of* ***S****, calculated for each pixel in the standard deviation on the* ***R***, ***G***, ***B*** *channel.*

Definition 3. *Well-exposedness: The first observation of each* ***RGB*** *channel of the original light intensity, we use Gaussian curve* $e^{(-\frac{(i-0.5)^2}{2\sigma^2})}$ *to separate each channel and produce metric variables* ***E****, which* $\sigma = 0.2$, ***i*** *is to measure light intensity variable.*

In this way, we can use the weighted linear combination, using the Formula 1 to calculate the pixel point weight $W_{i,j,k}$.

$$W_{i,j,k} = (C_{ij,k})^{\omega_C} * (S_{ij,k})^{\omega_S} * (E_{ij,k})^{\omega_E} \tag{1}$$

In Formula 1, **(i, j)** for the location of the pixel point, **k** for the **k-th** image, ω_C, ω_S, and ω_E for the weight index. When it is **0**, the corresponding metric is not included in the calculation.

We will calculate the weighted average value of each pixel to integrate two consecutive video frames. In order to get a stable result, we standardized the **N** weight maps, using the Formula 2.

$$\hat{W}_{ij,k} = [\sum_{k'}^{N} W_{ij,k'}]^{-1} W_{ij,k} \tag{2}$$

Output image **R** can be calculated using the Formula 3:

$$R_{ij} = \sum_{k=1}^{N} \hat{W}_{ij,k} I_{ij,k} \tag{3}$$

This will get a picture of the quality of the source video data, such as Fig. 2. This provide a reliable guarantee for the accuracy of the foreground object extraction from the video file.

Fig. 2. Focus stacking

B. The Method for Improving Compression Ratio of Video Synopsis

Optimization of Foreground Object Extraction. After decades of development, digital image processing technology has become more and more mature. But in the application of video abstract, for foreground object extraction, shadow makes the interesting region extraction of target is too large so that a video frame can not contain too many objects. In other words, The larger the foreground object, the greater the trajectory collision probability. In order to solve the collision problem proposed new challenges. Therefore, to remove the shadow from the foreground, has great academic value and practical significance for video abstract algorithm.

In this paper, the foreground extraction of video abstract add the shadow removal function [24] to make the extraction of moving foreground target more accurately and reduce the interest region of fore ground object. The ultimate goal is retaining the complete information of the foreground objects in video frames reducing the trajectory collision probability of foreground object. As the result, it makes video compression rate higher.

In Fig. 3, from the raw video data to the shadow removal of foreground object, which makes the foreground position more accurately. So that the trajectory collision problem can be reduced a lot, and more convenient to deal with.

The basic idea is to use a texture-based method to detect large area shadows:

- *First, using color feature selection of a plurality of possible regional, in order to select the shadow of each pixel, and then use the texture gradient to calculation each region to judge whether it is a shadow.*
- *Second, the association between the detection area and the background, because the shadow tends to keep the bottom of the texture, so there is a high correlation between shadow region and the background texture.*

In the second step, for each selected area each point $\mathrm{p} = (x, y)$ to calculate the light gradient $|\nabla_p|$, (Formula 4) and direction gradient θ_p (Formula 5).

$$\nabla_\mathrm{p} = \sqrt{\nabla_x^2 + \nabla_y^2} \tag{4}$$

$$\theta_\mathrm{p} = \tan^{-1} 2(\nabla_y / \nabla_x) \tag{5}$$

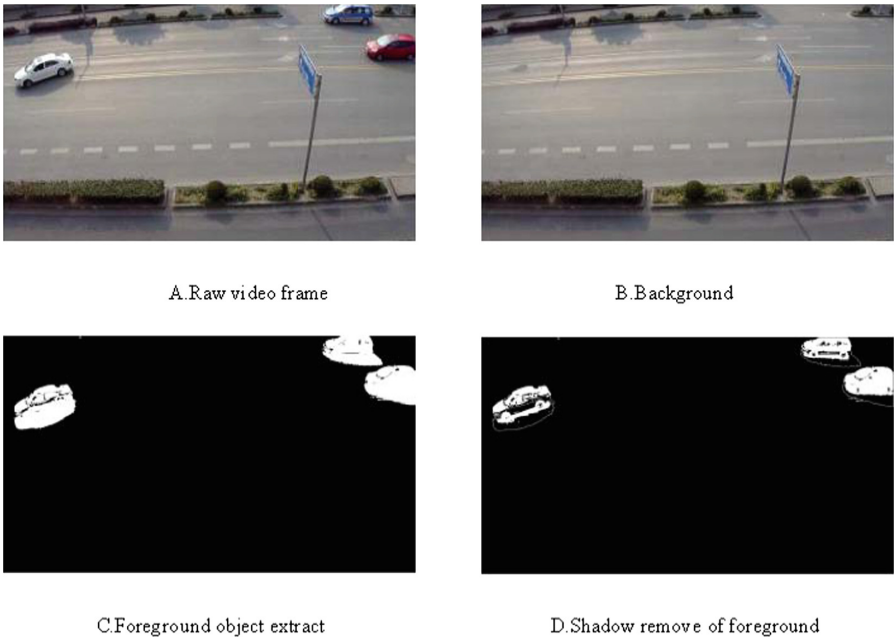


Fig. 3. Shadow remove

In Formula 4, ∇_x is the horizontal gradients of the point **p**, ∇_y is the vertical gradient of the point **p**, $\tan^{-1}2$ is a variant of the $\tan^{-1}$, return a angle value of a $[-\pi, \pi]$. When ∇_p less than the threshold τ_m, then considered here is the shadow, the first such a rough judgment.

Then in each point $p = (x, y)$ for the following calculations, where the **F** for the video frame tag, **B** as the background image tag, continue to calculate correlation direction gradients $\Delta\theta_p$ (Formula 6) between the original video frame and the background image.

$$\Delta\theta_p = \cos^{-1} \frac{\nabla_x^F \nabla_x^B + \nabla_y^F \nabla_y^B}{\sqrt{(\nabla_x^{F^2} + \nabla_y^{F^2})(\nabla_x^{B^2} + \nabla_y^{B^2})}} \tag{6}$$

Correlation direction gradients between the frame and background $c = \frac{\sum_{p=1}^{n} H(\tau_a - \Delta\theta_p)}{n}$ to evaluate, τ_a is gradient threshold, **n** for the selected shadow pixel number, **H(.)** as a unit step function, if $\tau_a - \Delta\theta_p$ is greater than or equal to **0**, just $H(.) = 1$; if less than **0**, $H(.) = 0$. According to the definition τ_c of a shadow threshold, compare **C** and τ_c $, if $C > \tau_c$, so the candidate area is shadow and should be removed from the foreground mask.

This method can reduce the area of the foreground object and increase the spatial redundancy, which can increase the foreground object trajectory in synopsis video. So that we can raise the compression rate in video abstract.

Trajectory Optimization for Video Skimming Synthesis. In the synthesis of foreground and background, in order to improve the compression rate of the video, you have to put multiple objects in the same video frame, but this will inevitably produce track collision of object. Usually, there is a track combination optimization algorithm based on clustering. The energy function is used to define a value of the target conflict and lost in the time and space. According to the start and end of the target space position, using K-means to cluster; then planning these targets in space-time location in the video abstract based on the segment tree [25].

However we use Semi-Transparent Handling Collision (**STHC**) algorithm to solve this problem. After the foreground object extract from surveillance video and re-combination in synopsis video, there is a trajectory collision, as shown c image in Fig. 4. When using STHC algorithm, two objects can be seen at the same time in the track intersection as shown d image in Fig. 4. This module is mainly used to solve the problem of trajectory collision during the synthesis of foreground object and background, which avoids the trajectory classification, and any track can be placed in the same video frame, so that the compression rate of the video abstract is increased.

$$\mathrm{Out} = in_A * 0.5 + in_\mathrm{B} * 0.5 \tag{7}$$

In Fig. 4, **a** graph is the original video object trajectory of **A**; **b** graph is the trajectory of the original video object **B**; **c** is the trajectories synthesis of **A** and **B** in video abstract; **d** graph can also see two objects (semi-translucent) when the object trajectory collision happened, using the Formula 7. In dealing with the trajectory, the use of **STHC** algorithm make any trajectory can be compressed in the same video frame, so as to improve the video compression ratio, thus forming a good video abstract.

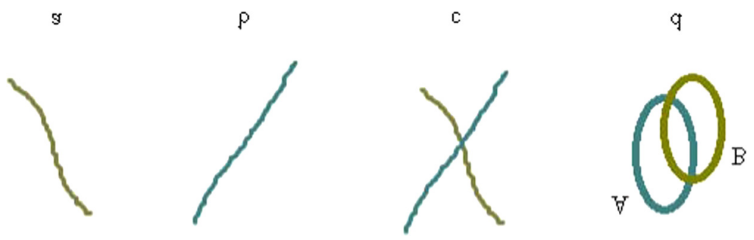


Fig. 4. Trajectory collision optimization

4 Application Effect

The video abstract algorithm proposed in this paper, that through optimizing each module of video summarization, result in the whole effect of video abstract algorithm is improved.

4.1 The Effect of Improved Foreground Extraction Algorithm

In the video pre-processing, the Fusion Image algorithm that based on Focus Stacking is used to pre-process each frame of the video. So that the video is clearer and the accuracy rate of foreground extraction is improved. Fusion image algorithm is designed in this paper based on **Enblend** [23], and the algorithm is integrated into the pre-processing module of video abstract algorithm. The experimental results are shown in Fig. 5.

In Fig. 5, **O** image represents the original video frame, $\mathbf{O_1}$ image and $\mathbf{O_2}$ image respectively expressed the adjacent two frames local magnification of original video, **D** image is the result of after the Image Fusion. It is clearly that the head of the white car become more clearer from the blur than in the original video frame. This is very beneficial for the foreground object extraction. Thus, the final effect of the synthesis video is better.

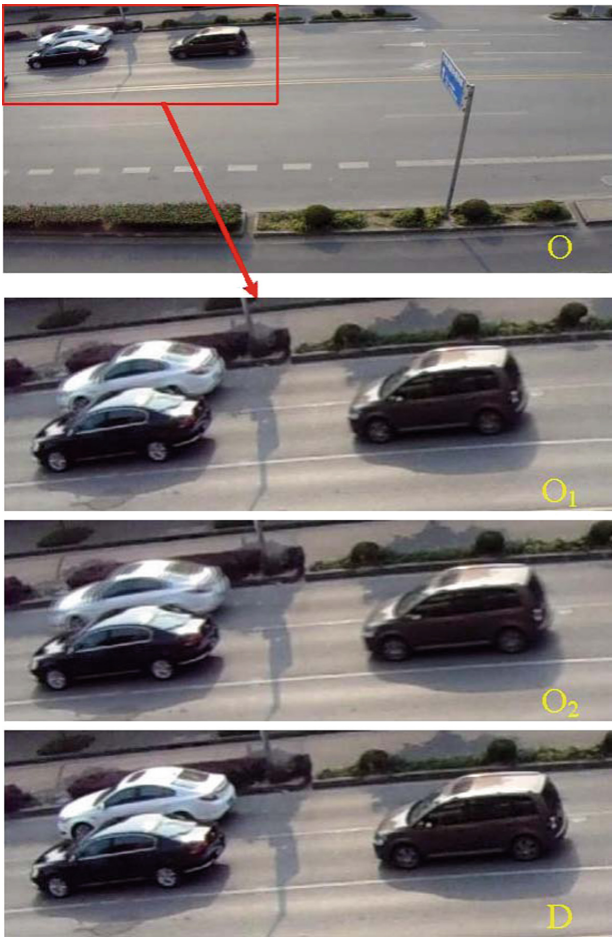


Fig. 5. Preprocessing effect of fusion image algorithm

In the experiment using **2** h of **720P** video that the frame rate is **25** frames per second, the total frames is **180000** frames. The foreground extraction using the **ViBe** algorithm for the integrated frame is **173430** frames, with the combination of Fusion Image and **ViBe** algorithm bring about the integrated frame achieved **178434** frames. Such as Table 1, the accuracy rate of foreground object extract is improved **2.78%**, so this method is effective for the pre-processing of video abstract.

Table 1. Accuracy rate

Algorithm	Accuracy rate
ViBe	96.35%
Improved ViBe	96.35%

4.2 The Improvement Effect of Compression Rate for Video Abstract

(1) The Effect of Shadow Remove Algorithm. Because the size of a picture is limited, it can contain the foreground object is also limited. So in order to make the image can contain more foreground objects, improve the compression rate, it is necessary to remove the shadow of the foreground object. We design a shadow removal algorithm suitable for video abstract based on Large region texture-based method [24], and integrate it into the foreground object processing module of video abstract. As a result, the accuracy rate of the foreground objects extract is promoted and for the trajectory synthesis module provides reliable guarantee.

As shown in Fig. 6, the **O** image is the original video frame, and the O_1 image is the foreground object extracted from the video frame, and the **D** image is the foreground object that the shadow had removed. Shadow removal can reduce the area of the foreground object, and reduce the collision probability for the trajectory synthesis of the foreground objects, thus making the synopsis video more beautiful and clear.

Also using the road monitoring video analysis, the original algorithm is used to synthesize the trajectory leading to 42048 collision frames, however the improved algorithm reduces the collision frame to 13662 frames. Such as Table 2, the rate of object collision is reduce by 15.77%. By reducing the collision rate, the video frame can contain more foreground object on the same frame so that the compression rate of video abstract increased by about 50%.

(2) The Effect of STHC Algorithm. For trajectory synthesis, the use of STHC algorithm, can further improve the compression rate of about 10%. In the previous algorithm of collision avoidance and the traffic lights (an object move first, another object wait for the former past then to move) algorithm, will bring track interruptions problem [26, 27]. In order to make the trajectory synthesis of foreground object continuous and uninterrupted, the paper proposed the STHC method, in same time, two objects can be seen in the same position, so that the trajectory and video frame information is not lost.

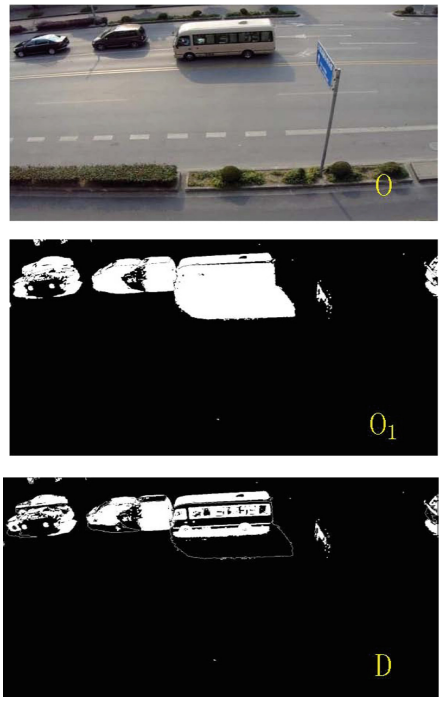


Fig. 6. The effect of shadow remove algorithm

Table 2. Collision rate

Algorithm	Accuracy rate
Old algorithm	23.36%
Shadow remove algorithm	7.59%

In Fig. 7, 20 images respectively show the whole two people cross walk and happening collision, among them, 6–18 images presented the effect of STHC algorithm. The two people independent cross walk in the synopsis video, but actually they are from different video frames at different time, the purpose to do so is to improve the compression ratio of the video.

Previously, researchers took the collision detection and prediction algorithm to reduce the collision, so that obtain a better video synopsis. But the STHC algorithm in this paper can omit this process, through STHC allows collision, and the video information is not lost, result in the trajectory without disruption in the synopsis video. As for a further comment, the STHC algorithm, compared with the traffic light algorithm, result in the synopsis video compression rate increased by about 10% in dealing with the track collision.

An optimized video abstract algorithm is proposed in this paper, there are many advantages: first, increase the clarity of the synopsis video; second, enhanced the integrity of the video information; third, improved the compression rate of the video.

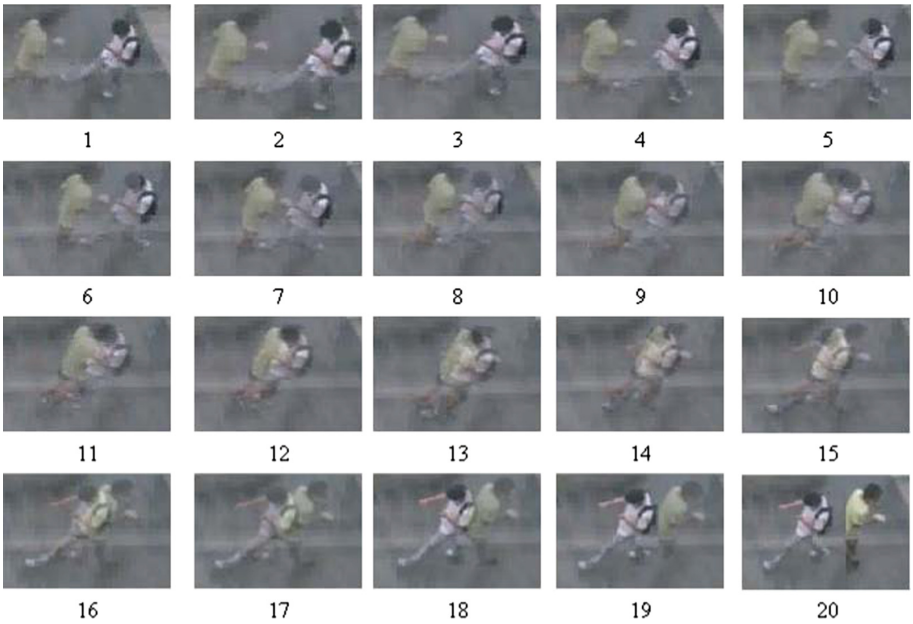


Fig. 7. Trajectory collision processing of video

5 Conclusion and Future Work

The proposed algorithm effectively improves the video abstract extraction. Mainly in three aspects: first, the accuracy rate of foreground extraction was increased by 2.78%; second, because of the shadow removal of the foreground object, so that the rate of track collision in the synthesis of video was reduced 15.77% and improves the compression ratio; third, using the Semi-Translucent Handling Collision (STHC) algorithm result in that the trajectory of foreground object continuous and uninterrupted, the video frame information is not lost and the compression was effectively improve by about 10%.

In this paper, because of the optimization of each process module, leads to increasing the amount of computation task of video summary. In order to get a good synopsis video in a reasonable time, it also need parallel computing and big data technology to speed up the video abstract algorithm. In future work, we will start from big data algorithm of video summarization, use big data related technologies to solve the synthesis technology of video abstract, so that achieve reducing the time required to generate the synopsis video and reduce storage space for video storage.

Acknowledgments. This work is funded by "Peak disciplines achievements in 2015 of the School of Film and Television Art Technology of Shanghai University" and "Shanghai University Material Genetic Engineering Institute" (No. 14DZ2261200). Thanks for the support of the high performance computing center.

References

1. Huang, K., Chen, X., Kang, Y., Tan, T.: A survey of intelligent video surveillance technology. J. Comput. Sci. **6**, 1093–1118 (2015)
2. Kang, M.: Research of video abstract algorithm based object. Ph.D. dissertation, Xi'an Electronic and Science University (2014)
3. Wang, J., Jiang, X., Sun, T.: Summary of video synopsis technology. Chin. J. Image Graph. **19**(12), 1940–1943 (2014)
4. Olivier, B., Marc, V.D.: ViBe: a universal background subtraction algorithm for video sequences. IEEE Trans. Image Process. **20**(6), 1709–1724 (2011). A Publication of the IEEE Signal Processing Society
5. Wu, Q., Shi, P.: Analysis of video abstract technology. J. Commun. Univ. China Nat. Sci. Edn. **15**(2), 54–58 (2008)
6. Bhaumik, H., Bhattacharyya, S., Dutta, S., Chakraborty, S.: Towards redundancy reduction in storyboard representation for static video summarization. In: International Conference on Advances in Computing, Communications and Informatics, pp. S56–S57 (2014)
7. Zhu, X., Wu, X., Fan, J., Elmagarmid, A.K., Aref, W.G.: Exploring video content structure for hierarchical summarization. Multimedia Syst. **10**(2), 98–115 (2004)
8. Yeh, C.H., Kuo, C.H., Liou, R.W.: Movie story intensity representation through audiovisual tempo analysis. Multimedia Tools Appl. **44**(2), 205–228 (2009)
9. Zhang, S.H., Li, X.Y., Hu, S.M., Martin, R.R.: Online video stream abstraction and stylization. IEEE Trans. Multimedia **13**(6), 1286–1294 (2010)
10. Winnemöller, H., Olsen, S.C., Gooch, B.: Real-time video abstraction. ACM Trans. Graph. **25**(3), 1221–1226 (2006)
11. Zhang, L., Cao, Y., Ding, G., Yong, W.: A computable visual attention model for video skimming. In: IEEE International Symposium on Multimedia, pp. 667–672 (2008)
12. Rav-Acha, A., Pritch, Y., Peleg, S.: Making a long video short: dynamic video synopsis. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 435–441 (2006)
13. Pritch, Y., Ratovitch, S., Hendel, A., Peleg, S.: Clustered synopsis of surveillance video. In: 6th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 195–200 (2009)
14. Pritch, Y., Rav-Acha, A., Gutman, A., Peleg, S.: Webcam synopsis: peeking around the world. In: ICCV 2007, Riode Janiero, pp. 1–8 (2007)
15. Vural, U., Akgul, Y.S.: Eye-gaze based real-time surveillance video synopsis. Pattern Recogn. Lett. **30**(12), 1151–1159 (2009)
16. Li, T., Mei, T., Kweon, I.S., Hua, X.S.: Video^M: multi-video synopsis. In: IEEE International Conference on Data Mining Workshops, pp. 854–861 (2008)
17. Li, T., Mei, T., Kweon, I.S., et al.: Multi-video synopsis for video representation. Sig. Process. **89**(12), 2354–2366 (2009)
18. Qian, Q., Gunturk, B.K.: Extending depth of field and dynamic range from differently focused and exposed images. Multidimens. Syst. Sig. Process. **27**, 1–17 (2015)
19. Qian, Q., Gunturk, B.K., Batur, A.U.: Joint focus stacking and high dynamic range imaging. In: Proceedings of SPIE, pp. 866 004–866 004–7 (2013)
20. Dezeeuw, P., Gledhill, L., Cardwell, M.W.: Motor controlled macro rail for close-up focus-stacking photography (2012)

21. Brecko, J., Mathys, A., Dekoninck, W., Leponce, M., Vandenspiegel, D., Semal, P.: Focus stacking: comparing commercial top-end set-ups with a semi-automatic low budget approach. A possible solution for mass digitization of type specimens. Zookeys **464**, 1–23 (2014)
22. Zhang, C., Bastian, J., Shen, C., Van den Hengel, A., Shen, T.: Extended depth-of-field via focus stacking and graph cuts. In: 2013 20th IEEE International Conference on Image Processing (ICIP), pp. 1272–1276 (2013)
23. Demandolx, D., Ricard, D.A., Dideriksen, T.L., Chiu, K.G.: Combining multiple images in bracketed photography (2013)
24. Sanin, A., Sanderson, C., Lovell, B.C.: Shadow detection: a survey and comparative evaluation of recent methods. Pattern Recogn. **45**(4), 1684–1695 (2012)
25. Jiang, D.: Research on video abstract generation based on clustering mining. Ph.D. dissertation, Zhejiang University (2010)
26. Sun, L., Xing, J., Ai, H., Lao, S.: A tracking based fast online complete video synopsis approach. In: International Conference on Pattern Recognition, pp. 1956–1959 (2012)
27. Xu, L., Liu, H., Yan, X., Liao, S., Zhang, X.: Optimization method for trajectory combination in surveillance video synopsis based on genetic algorithm. J. Ambient Intell. Hum. Comput. **6**(5), 1–11 (2015)

The Prediction of CTR Based on Model Fusion Theory

Jiehao Chen(✉), Shuliang Wang, Ziqian Zhao, and Jiyun Shi

School of Software, Beijing Institute of Technology,
South Zhongguancun Street. 5, Beijing 100081, China
{cjh, slwang2011}@bit.edu.cn, 1264733941@qq.com,
340604636@qq.com

Abstract. Online advertising makes it possible to show different ads to different customer groups according to their own characteristics, which will definitely prove the efficiency of ads, and we manage to accurate advertising by predicting the CTR of ads based on varieties of algorithm and models. This essay presented a kind of merged model of GBDT and LR, whose accuracy doesn't heavily depend on the effect of building features artificially. In the GBDT part of the new model, the ways to build the decision trees made it possible to recognize the effective combination of features, on the other hand, the LR part of model makes it possible to deal with large amount of data. At the same test condition, the new model performed better than LR at the range of 1.41% to 1.75% with the standard of MSE, AUC and Log Loss. The results of the experiment show that GBDT model did a great job on building features for LR model without much help from human, which provides a new thought to improve the current CTR prediction models.

Keywords: CTR prediction · Gradient Boosting Decision Trees · Logistic Regression · Model fusion

1 Introduction

With the widespread of the Internet, the online advertising has come into being. Compared with the traditional advertising, the online advertising has the ability to record, track and research into the customers' consumption habits and preferences, which makes it able to target ads at specific users.

No matter what type of advertisement, the result that whether the customer clicked the ads or not is an important evaluating indicator. The prediction of CTR plays a key part in the procedure of accurate advertising.

Through years of practices and experiments, there are many models and algorithms that can achieve satisfactory results with a large amount of training data as well as appropriate manually-built features. However, building an effective feature requires huge time and efforts on selecting, processing and constructing features from the raw data, which also depends heavily on the work of data analysts and scientists. The perform of these prediction models have obviously positive correlation with the validity of the artificial features.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 90–100, 2017.
DOI: 10.1007/978-981-10-3969-0_11

The problem is, we usually don't have enough resources to figure out the best way to build features by trial and error in many CTR prediction scenarios. In order to solve this problem, we did a lot of research on the existing models and finally find a hybrid model to reduce the dependence on human experience.

We merged a kind of decision tree model, Gradient Boosting Decision Tree (GBDT) model, with the traditional Linear Regression (LR) model. In the GBDT part of the new model, the ways to build the decision trees made it possible to recognize the effective combination of features, on the other hand, the LR part of model makes it possible to deal with large amount of data.

In this essay, we chose three types of assessment indicators to evaluate the fusion model, Mean Square Error, Area Under Curve and Log Loss. Under the same experiment condition, the merged model perform better than the LR model in the scale of 1.41–1.75%.

This article first described the theories and methods to combine these two models in Sect. 2. Then Sect. 3 of this essay did an overview of the experiment. And we showed the results of this experiments and had some discussions in Sect. 4. At last, the article presented a brief conclusion.

2 Proposed Model and Approach

2.1 Logistic Regression

Logistic Regression [1] is a classic linear classifier designed to calculate the probability of samples for the positive and negative class. It's simple and linear structure gives it the ability to deal with large amount of training data.

$$P(Y = 1|x, w) = \frac{e^{w^T x+b}}{1 + e^{w^T x+b}} = \frac{1}{1 + e^{-(w^T x+b)}}. \tag{1}$$

$$P(Y = 0|x, w) = \frac{1}{1 + e^{w^T x+b}}. \tag{2}$$

In the formulas above, P stands for probability, x stands for the input vector which is insist of the features of the predicted sample while w is the parameters of the LR model which describes the weight of each feature. b is also the parameter of the model which is used as an offset.

However, a linear classifier can't perform well while facing the nonlinear data which is very common in many application scenarios. In other words, the simple Logistic Regression model is able to handle classification problems rather than the regression problems that predict the CTR. We are supposed to take several pretreatments to achieve a more satisfying result.

2.2 Two Major Ways to Improve the Linear Classifier

One major way to improve the accuracy of a linear classifier is transforming the input features into a vector. For categorical features, we can use a technique called One-Hot Encoding [2] to transfer it into a vector which only contains 0 s and 1 s.

For instance, assuming a categorical feature that describes the type of the content of the ads [sports, health, entertainment, food]. Instead of simply replace it by the vector of [1, 2, 3, 4] which add some redundant sequence information to the model, we can get a vector [0001, 0010, 0100, 1000] through One-Hot Encoding where '0001' stands for 'sports' and so on. That is to say, we have encoded a feature that contains four states to four binary features which makes it much easier to input to the linear classifier.

As for continuous features, such as the time when the ads are shown and how long they have been shown, we can first split them into categorical ones according to some logical relation, then deal them with One-Hot Encoding.

The other way is to merge a nonlinear model with the linear one to deal with the nonlinear data [3]. In this essay, we managed to achieve this goal by introducing a kind of Decision Tree model, the Gradient Boosting Decision Tree [4] model.

2.3 Gradient Boosting Decision Trees

The trees in Decision Tree models can be divided into three main types, classification trees, regression trees and CART, which is the short of Classification and Regression Trees. Classification trees are usually used to predict the category of samples while the regression trees are used to predict the specific numbers. CART is a combination of these two kinds of decision tree, which is also the type of tree that is used in the Gradient Boosting Decision Tree model.

The trees in CART are binary trees. When they reach a sufficiently depth, every samples that belongs to one leaf node can be seen as a class that shares a common value. The key to the build this binary tree is that during each split, we choose the way which achieves the smallest coefficient of heterogeneity, as well as the way that makes the left and right sons become more distinguish classes. Below is a brief procedure of building a CART [5] (Fig. 1).

CART procedure

```
For n = 1 to N do:
//N is the number of features that are not used yet
  For each Split Way do:
    Calculate Coefficient of Heterogeneity g
        If g < gmin
            gmin = g
            Best Split Way = Current Split Way
  EndFor
EndFor
```

Fig. 1. A brief procedure of CART.

There are different split ways for continuous or categorical features. In order to ensure both the accuracy and efficiency of the model, we adopted the following split ways in this experiment.

For continuous features, we first ranked all the different values in the order from height to low, then went through every mid-value of two neighboring values as the split point. If the feature value set contains N different values, there will be N − 1 distinct split ways.

On the other hand, for each categorical features, we first build a set consist of distinct feature values, then split it into one of its non-void proper sub-set and the complementary set of the sub-set. If the feature value set contains N different values, the number of its non-void proper sub-sets will be $2^N - 2$, and half of the sub-sets complement with the other half, which provides us $2^{N-1} - 1$ different split ways.

There are three common ways to calculate the coefficient of heterogeneity, Gini (3), Twoing (4) and LSD (5).

$$Gini(D) = 1 - \sum_{i=1}^{n} p_i^2 \tag{3}$$

$$\Delta i(t) = \frac{p_l p_r}{4}\left[\sum_{i=1}^{n} |p(i|t_l) - p(i|t_r)|\right]^2 \tag{4}$$

$$f(x) = \sum_{i=1}^{n} (a_i^T x - b_i)^2 \tag{5}$$

The first two are often used to solve the classification problems, on the other hand, the LSD is used on regression problems. In this experiment, we choose the LSD to be our way to calculate the coefficient of heterogeneity.

In the usual usage of CART, we are supposed to build the binary tree as deep as we can until it reaches a certain depth or gets enough leaf nodes. But in GBDT, due to the idea of boosting, we need a series of weak classifiers rather than a strong one. The way we build the binary trees adopts the method based on gradient descent, the brief algorithm is as follows (Fig. 2).

GBDT procedure

```
For m = 1 to M do:
  //M is the number of trees that we need
  For n= 1 to N do:
    //N is the number of features that are not used
yet
    Calculate the prediction value of current trees p
    Calculate residual
    Real value = residual
  EndFor
  Build next CART
EndFor
```

Fig. 2. A brief procedure of GBDT.

We use the least square method to calculate the residual. Below is the loss function.

$$L(y, F) = \frac{(y - f)^2}{2}. \tag{6}$$

Its first order derivation is as follows.

$$\tilde{y}_i = \left[\frac{\partial L(y, F(x_i))}{\partial F(x_i)}\right] F(x). \tag{7}$$

What we can find after some formalization is that the residual can be describes as shown below, which means the residual equals to the target value minus the predict value in the current iteration, which is also the target value in the next iteration.

$$\tilde{y}_i = y_i - F_{m-1}(x_i). \tag{8}$$

When it comes to the prediction of CTR after finishing training the GBDT, the procedure can be briefly described as follows. We can build a new feature vector for the LR model using the intermediate result of the GBDT model.

As shown in the Fig. 3, this trained GBDT model contains n SubTrees and m leaf nodes in total, where SubTree 1 has three leaf nodes, L_1, L_2 and L_3, and its depth is three, while SubTree n has two leaf node L_{m-2}, L_{m-1}, and its depth is two. Assuming Sample X's feature vector is [13:00, Food, …, App], it will fall in the L_2 leaf node in SubTree 1 and the L_{m-2} leaf node in SubTree n. And the final prediction value equals to the sum of these leaf nodes' target values.

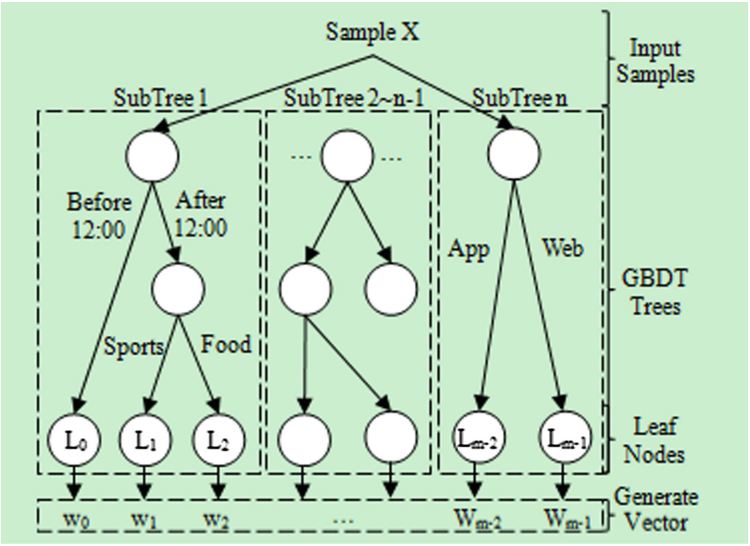


Fig. 3. The structure of GBDT and how to generate vector.

While generating the new feature vector, each leaf node decides the state in the corresponding dimension. If the sample falls in the leaf node, the state is 1, otherwise is 0. Use Sample X as an example, the vector will be [0, 0, 1, …, 1, 0] which has m dimensional.

In this experiments, we first transformed the original features into vectors through One-Hot Encoding, then combined it with the vector that is generated by the GBDT model, and finally input the joint vector to LR model to increase its accuracy.

3 Experiments

3.1 Data Descriptions

The data we used in this experiment is comes from the Click-Through Rate Prediction match on the data mining match platform, Kaggle. The original dataset is provided by Avazu, which contains 10 days detailed ads display record. We chose several key fields to build our own experiment dataset (Table 1).

Table 1. Dataset description.

Field name	Field description	Field type
id	The id of the ads	Categorical
click	Clicked-1, otherwise-0	Categorical
hour	When the display happened	Continuous
C1	Anonymous field	Categorical
banner_pos	On which part of the webpage the ad was shown	Categorical
site_category	The category of site	Categorical
app_category	The category of app	Categorical
device_type	The type of device	Categorical
device_conn_type	The type of device conn	Categorical

Due to the limitation of memory, our own experiment dataset contains 3999999 samples from one day. Then we did a random split at the ratio of 9:1 to build our train and test dataset. The CTR in both the train and test dataset is similar, which can ensure the training effect and meet testing standard.

3.2 Evaluation Indicator

Mean Square Error: The MSE [6] of an estimator measures the average of the squares of the errors. It is always non-negative, and values closer to zero are better. The computational formula is as follows.

$$MSE = \frac{\sum_{i=1}^{N} (Y_i - P_i)^2}{N}. \tag{9}$$

N stands for the total number of the samples. Y_i is the real value of the i_{th} ample (values 1 when it is clicked, otherwise 0), P_i is the prediction value that the model gives to the i_{th} sample.

Area Under Curve: The curve in AUC [7] usually means the Receiver Operating Characteristic, whose horizontal axis stands for false negative rate and vertical axis stands for true positive rate (Table 2).

$$TruePositiveRate = \frac{N_{tp}}{N_{tp} + N_{fn}}. \tag{10}$$

$$FalsePositveRate = \frac{N_{fp}}{N_{fp} + N_{tn}}. \tag{11}$$

N_{tp}, N_{fn}, N_{fp} and N_{tn} are the amount of true positive, false negative, false positive and true negative results. The value of AUC ranges from 0.5 to 1, the bigger the value is, the better the prediction is.

Table 2. Receiver operating characteristic.

True value	Predict	
	1	0
1	True positive	False positive
0	True negative	True negative

Log Loss: Also called cross-entropy loss or logistic loss. It is the logarithm of the likelihood function for a Bernoulli random distribution. This error metric is used where contestants have to predict that something is true or false with a probability (likelihood) ranging from definitely true to equally true to definitely false.

$$LogLoss = \frac{\sum_{i=1}^{N} (Y_i * \log(P_i) + (1 - Y_i) * \log(1 - P_i))}{-N}. \tag{12}$$

N stands for the total number of the samples. Y_i is the real value of the i_{th} sample (values 1 when it is clicked, otherwise 0), P_i is the prediction value that the model gives to the i_{th} sample.

3.3 Experiment Procedure

Here is the main procedure of the experiment:

(1) Build the test and train dataset according to the way described in Sect. 3.1.
(2) Train the GBDT model to find the best parameters to fit the scenario. Record the best performance that the GBDT model achieved alone. Generate the feature vector with the best parameters.

(3) Train the LR model to find the best parameters to fit the scenario. Record the best performance that the LR model achieved alone.
(4) Train the LR model with the feature vector and best parameters. Record the best performance of the fusion model.
(5) Analyze the results and come to a conclusion.

Supplementary Explanation: Due to the effect of sampling rate, the training samples of each experiment are not the exactly same ones. So the final result is an average of three times of experiments.

4 Results and Discussions

4.1 Results

After several experiments, we found the best parameters for LR and GBDT model in our experiment environment (Table 3).

Table 3. Parameter settings.

GBDT parameters		LR parameters	
Parameter name	Value	Parameter name	Value
Sampling rate	0.8	Learning rate	0.05
Iteration	10	Sampling rate	0.8
Max depth	4	Sample name	360000

Here are the detailed highest result of comparison experiments (Table 4).

Table 4. Experiments results.

Indicators	Models		
	LR	*GBDT*	*LR + GBDT*
MSE	0.14117947	0.13960281	0.13961235
AUC	0.56146128	0.60591896	0.56155468
Log Loss	0.45544369	0.44955803	0.44932354

As it shows in Fig. 4. Using feature vectors generated by GBDT model decreased the MSE, which means the improvement of the prediction accuracy.

Figure 5 shows the results measured by AUC, however, it is not able to prove that the accuracy has been increased by merge GBDT and LR models. We will have a brief discussion about this in the next chapter.

Figure 6 shows that the Log Loss of the combined model is lower than both the LR and the GBDT model, which surely proved the accuracy is improved.

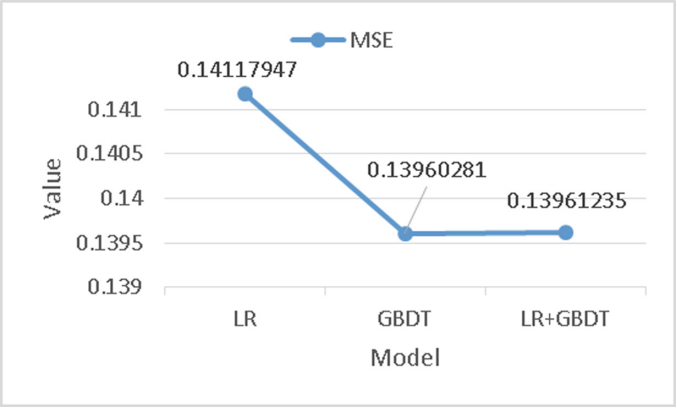


Fig. 4. The structure of GBDT and how to generate vector.

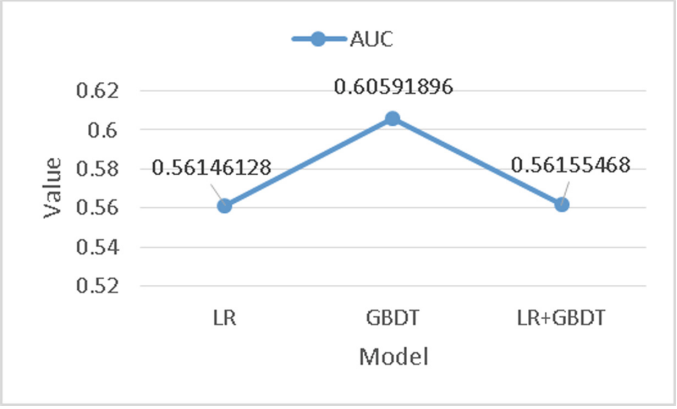


Fig. 5. The results measured by AUC.

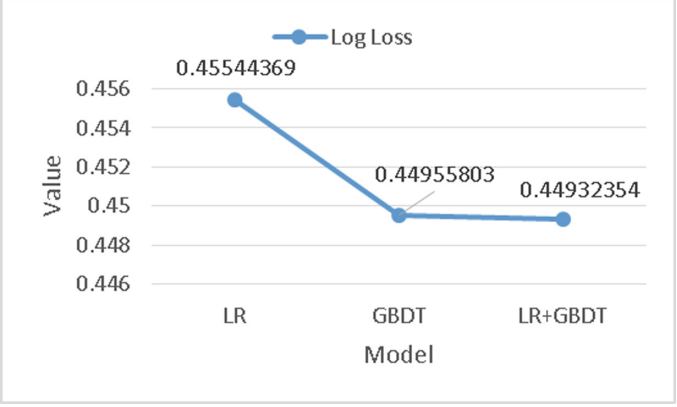


Fig. 6. The results measured by Log Loss.

4.2 Discussions

As mentioned above, Fig. 5 shows the results measured by AUC, however, it is not able to prove that the accuracy has been increased by merge GBDT and LR models. We did a further research about the theory of AUC, in order to find out why it doesn't prove our thoughts.

What we can learn about AUC in Sect. 3.2 is that, unlike the MSE and Log Loss, this indicator doesn't concern about the exact prediction value. The only variable which matters to AUC is the result of the prediction, that is to say, the ads will be clicked (1) or not (0). For example, the LR model predicts that the probability ads X will be clicked is 0.6, and the probability provided by merged model is 0.7. Assuming that ads X is truly clicked, it is obviously that the result of the merged model is better, and the MSE and Log Loss of the combined model can reflect the fact clearly, while the AUC of these two model will be the same because both of them have made a correct prediction.

So it comes to a conclusion that AUC cares more about the correctness of the prediction results while MSE and Log Loss cares more about the values that the models output, which is absolutely essential when we are dealing with the problems in the field of Computational Advertising.

5 Conclusions

This essay presented a kind of merged model of GBDT and LR. In the GBDT part of the new model, the ways to build the decision trees made it possible to recognize the effective combination of features, on the other hand, the LR part of model makes it possible to deal with large amount of data. At the same test condition, the new model performed better than LR at the range of 1.41% to 1.75% with the standard of MSE, AUC and Log Loss.

With the increasing diversity of advertising scenarios, using only one model to predict the CTR of ads becomes less and less effective. The results of the experiment in this essay show that GBDT model did a great job on building features for LR model without much help from human, which provides a new thought to improve the current CTR prediction models. In the future, the tendency of Computational Advertising will be using multiple basic models to further prove the efficiency and effectiveness.

References

1. Daniel, T.L.: Data Mining Methods and Models. Wiley-IEEE Press, New York (2006)
2. Harris, D., Harris, S.: Digital Design and Computer Architecture, 2nd edn. Morgan Kaufmann, San Francisco (2012)
3. He, X., Pan, J., Jin, O., et al.: Practical lessons from predicting clicks on ads at Facebook. In: 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1–9. ACM Press (2014)

4. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. J. Ann. Stat. **29** (5), 1189–1232 (2000)
5. Breiman, L.I., Friedman, J.H., Olshen, R.A., et al.: Classification and regression trees (CART). J. Biom. **40**(3), 17–23 (1984)
6. Harvey, D., Leybourne, S., Newbold, P.: Testing the equality of prediction mean squared errors. Int. J. Forecast. **13**(2), 281–291 (1997)
7. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. J. IEEE Trans. Knowl. Data Eng. **17**(3), 299–310 (2005)

An Improved Algorithm of LEACH Protocol Based on Node's Trust Value and Residual Energy

Miaoyuan Huang(✉), Enjian Bai, Xueqin Jiang, and Yun Wu

2999 North Renmin Road, Shanghai, China
qiaokeliyal88@163.com

Abstract. In previous clustering algorithms for wireless sensor networks such as LEACH, the residual energy of the node is not considered into the election of the cluster head, and there is no security mechanism for identifying malicious nodes. In response to this phenomenon, we propose a WSN clustering routing algorithm based on node's trust value and residual energy to eliminate the negative influence of malicious nodes as well as to balance energy consumption of the network. The simulations show that the proposed algorithm prolongs the network lifetime and has a better performance in the communication quality.

Keywords: LEACH protocol · Cluster head · Trust value · Residual energy

1 Introduction

Nowadays, with the development of sensor technology and radio frequency technique, the low energy consumption and low-cost wireless sensors have been applied on a large scale. So, the wireless sensor network (WSN) came into being accordingly. The correct routing and data transfer are the necessary conditions to guarantee the normal work of the WSN, while the routing protocol on network layer is responsible for providing key routing services. Therefore, the security of routing protocol directly influences the safety and availability of wireless sensor network (WSN), so it is a crucial issue in WSN security research.

So far, different protocols have been designed for WSNs. Among these routing protocols, the hierarchical-based routing protocols have more advantages than the flat-based routing protocols [1]. The hierarchical-based routing is more flexible and it has higher scalability, above all, it is an efficient way to lower energy consumption within a cluster, so it can improve the network performance greatly. Hierarchical-based routing is mainly two-layer routing where one layer is used to select cluster heads and the other for routing [2]. The network is divided into several clusters, each cluster contains one cluster head and many normal nodes, the normal nodes are used for data collecting, while the cluster head is responsible for collecting data from normal nodes and sending information to the base station after aggregating the data. Due to the characteristic of self-organizing, hierarchical-based routing has a better network performance with lifetime extension.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 101–109, 2017.
DOI: 10.1007/978-981-10-3969-0_12

LEACH [3] is the earliest low-energy adaptive protocol based on clustering. It saves power because the normal nodes are not involved in data transfer, furthermore, the problem faced in cluster heads of COUGAR [4] is solved here as cluster head is selected randomly over time in-order to balance the node energy dissipation rate [5]. LEACH extends the lifetime of the network greatly; Then, LEACH-C [6] protocol was put forward to use a centralized clustering algorithm which has the same steady-state as LEACH. It avoid the random distribution of cluster heads; Younis and Fahmy. proposed the HEED protocol [7] to generate distributed evenly clusters after several iterations of clustering. In MCHCA algorithm [8], the mobile cluster heads move to the best location for each round, and the best location depends on the residual energy of nodes in each cluster to achieve the minimum differences in residual energy between nodes.

In general, the current hierarchical-based routings haven't taken full consideration of energy of cluster head and the security mechanism to eliminate the negative influence of malicious nodes. The algorithm presented in this paper has a malicious node identifying mechanism and an improvement in clustering.

2 System Model

2.1 Network Model

In this paper, we consider a sensor network consisting of one hundred sensor nodes randomly deployed over a field of 100 m × 100 m, assuming that the nodes and the base station are all stationary after deployment.

2.2 Energy Consumption Model

The energy consumption model used in this paper mainly considers the energy consumption of transmitters and receivers. There are two channel models, free space (d^2 power loss) and multi-path fading (d^4 power loss), depending on the distance between the transmitter and receiver.

The energy spent for transmission of an l-bit packet over distance d is:

$$E_{TX}(l,d) = \begin{cases} lE_{elec} + lE_{fs}d^2, d < d_0 \\ lE_{elec} + lE_{mp}d^4, d \geq d_0 \end{cases} \quad (1)$$

where l represents the packet size, E_{elec} represents the consumption of transceiver for each bit. If the distance between two nodes exceeds the threshold $d_0 = \sqrt{E_{fs}/E_{mp}}$, the energy consumption of transmission will increase sharply, so it turns to multi-path fading model.

The energy spent for reception of an l-bit packet is:

$$E_{RX} = lE_{elec} \quad (2)$$

Assuming that there are n nodes existing in the net currently, the average energy of the network is:

$$E_{avg} = \frac{\sum_{i=0}^{n} E_i}{n} \tag{3}$$

3 The Proposed Algorithm

3.1 Calculation of Trust Value

Trust value is the evaluation of one node's reputation. It is valued by the node's neighbor nodes and the information exchange of the other nodes. The higher trust value means the better reputation and that the node is more reliable. So the algorithm in this paper let the nodes with high trust values have more competitive advantages to become the cluster heads. In the other hand, the lower trust value of the node is, the more suspicious of the node is. And if the trust value is below a pre-set threshold, the node should be considered as a malicious node.

The trust value is calculated periodically. For example T_i^k represents the trust value of node i in the k-th time period:

$$T_i^k = \frac{R_t}{R_l \times L} \tag{4}$$

where R_t represents throughput rate of node i, R_l represents the packet loss rate, while L represents the latency of the transmission.

The trust value in each time period has different weight in the calculation of trust value, which depends on the distance of time. The more recent of time, the greater the weight of T_i^k is. In order to achieve it, an attenuation function is set as follows:

$$f_k = \rho^{m-k}, \quad 0 < \rho < 1, \ 1 \le k \le m \tag{5}$$

then, this function is used to set the weight of T_i^k:

$$T_i = \frac{\sum_{k=1}^{m} f_k T_i^k}{\sum_{k=1}^{m} f_k} \tag{6}$$

Assuming that there are n nodes existing in the net currently, the average trust value of the network is:

$$T_{avg} = \frac{\sum_{i=1}^{n} T_i}{n} \tag{7}$$

3.2 Clustering Algorithm

Election of Cluster Head. In the traditional algorithm LEACH, cluster heads are selected randomly without taking the residual energy and the behavior of the node into consideration. So, the new algorithm improve it by adding a precondition of being a cluster head, and modify the election threshold $Th_i(t)$. Firstly, a trust value threshold is set as *TH*, and the precondition of being a cluster head is that the node's trust value exceeds threshold *TH*. Node satisfying the precondition will get a random number in the range of *0–1*. If the random number is less than the election threshold $Th_i(t)$, then the node will be select as a cluster head. The formula for calculating the $Th_i(t)$ of node i is as follows:

$$Th_i(t) = \begin{cases} \frac{p}{1-p(r \bmod \frac{1}{p})} \frac{E_i}{E_{avg}} \frac{T_i}{T_{avg}}, & node_i \in G \\ 0, & else \end{cases} \tag{8}$$

among which p represents the percentage of cluster heads in the network, E_i represents the residual energy of node i, T_i represents the trust value of node i, while E_{avg} represents the average energy of the network and T_{avg} represents the average trust value of the network. In this formula, $node_i \in G$ means that node i haven't been selected as a cluster head in the epoch (every $1/p$ rounds it begins a new epoch).

The energy consumption is quite high for being a cluster head, so this algorithm adds energy parameter to the election threshold calculating to let the node who has more energy get higher possibility for election. This will avoid the untimely death of the node who has low power, so that it can extend the lifetime of the network greatly. But usually, a malicious node will claim that it has high residual energy to defraud the chance of being a cluster head, so that it can attack the network. So the trust value threshold *TH* is set in the new algorithm to avoid this, because the malicious nodes usually have low trust values due to the bad behavior, then, if their trust values are below *TH*, they will be denied the chance to be selected as a cluster head. Meanwhile, the threshold *TH* can also eliminate the normal nodes who are not suitable to become the cluster head because of the low trust values. In Eq. (8), the factor T_i/T_{avg} is added into it. Because trust value is the evaluation of one node's reputation, it is calculated from past and it can be used to predict the behavior of the node in the future. A node with high trust value is supposed to behave well, so it will have higher possibility for election due to the factor T_i/T_{avg}.

Through introducing the factors of trust value and residual energy, the good nodes get higher possibility of being selected as a cluster head, which will improve the communication quality and extend the lifetime of the network.

However, the two factors T_i/T_{avg} and E_i/E_{avg} in Eq. (8) will make the number of cluster heads in every round can't meet the expectation: $n \times p$. So, in this algorithm, after election of cluster head in each round, it will check if the number of cluster heads reaches $n \times p$. If not, sort the other nodes by the value ($E_i \times T_i$) in descending order, and select the top ($n \times p$-*cluster*) nodes as cluster heads (*cluster* refers to the number of cluster heads in present). It can ensure the communication quality by controlling the number of cluster heads.

Control of Cluster-Heads Distance. In order to avoid the random distribution of cluster heads, the cluster-heads distance should be limited to a value. The distance of any two cluster heads should be above this value to achieve a uniform distribution of cluster heads. The method of setting the limit value is as below:

Assuming that a field of M × M is supposed to be divided into k clusters, the radius of each cluster (r) should be in the interval $[\frac{M}{2\sqrt{k}}, \frac{M}{\sqrt{k\pi}}]$. So, we determine the lower limit value for cluster-heads distance as d_{min}:

$$d_{\min} = 2r_{\min} = \frac{M}{\sqrt{k}} \tag{9}$$

During the election of cluster heads, if there is a cluster-head distance below d_{min}, one of the two nodes who has lower value of ($E_i \times T_i$) will lose the chance of being cluster head in this round.

Mechanism for Identifying Malicious Nodes. The worse one node behaves, the lower trust value it has, then it is more suspicious. If one node's trust value is too low, we should suspect it as a malicious node. So another trust value threshold *TL* is set for identifying malicious nodes. Given that normal nodes may behave badly occasionally, in order to avoid misidentifying the normal node as malicious node because its trust value fall below *TL* occasionally, the algorithm in the paper adopt the twice identifying trust model, specific as follows: check every nodes per round. If a node is found to have a trust value below the threshold *TL* for the first time, it will be added to a suspect list. If a node is found to have a trust value below the threshold *TL* and it is already in the suspect list, which means it is suspected for the second time, then this mechanism identify it as a malicious node and keep it out from the communication of the network.

4 Simulation and Analysis

4.1 Simulation Settings

Some experiments have been done with *MATLAB* to compare our algorithm with LEACH and LEACH-ED in the performances such as the lifetime and communication quality of the network. One hundred sensor nodes are randomly deployed over a field of 100 m * 100 m. In the simulation, 6 malicious nodes with higher initial energy and bad behavior are deployed randomly in the field. We assume that the malicious nodes attack the network by dropping packets.

The main parameter settings are shown in the following Table 1:

4.2 Results and Analysis

Figure 1 shows the states of nodes while the network is running. The node which is plot as '+', represents the normal node; and the 'o' represents the advanced node which has more energy than the normal node; the '×' which is blue and in the middle of the field represents the base station and the black '×' represents that the node is selected as

Table 1. The main parameter settings

Parameter	Value	Parameter	Value
Amount of nodes (N)	100	Packet length	4000 bit
Distribution area	100 m × 100 m	Control packet length	100 bit
Location of BS	(50 m, 50 m)	E_{elec}	5×10^{-10} J/bit
Initial energy of node	3 J	E_{fs}	1×10^{-13} J/(bit•m^2)
Percentage of cluster heads (p)	0.1	E_{mp}	1.3×10^{-15} J/(bit•m^4)
Maximum round	5000		

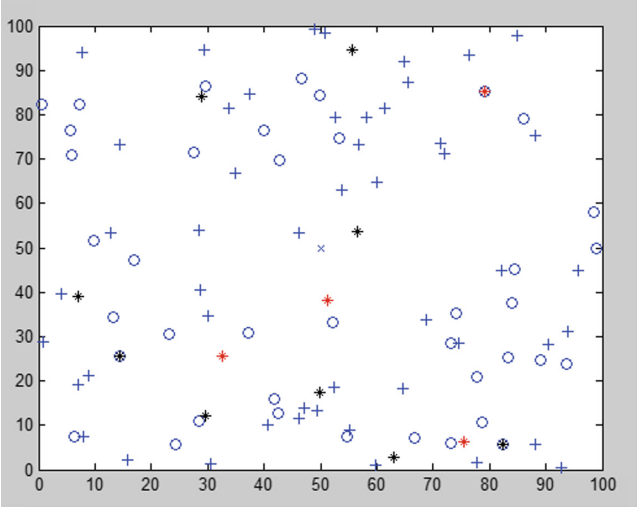


Fig. 1. The states of nodes during communication (Color figure online)

a cluster head; when the algorithm identifying a node as a malicious node, it will be plot as a red '*'.

Simulation Results in Lifetime of Network. The number of survived nodes in each round is counted during the simulation, the result is shown in Fig. 2:

It is clearly observed that the numbers of survived nodes decrease in different rate in different algorithm. Among them, in LEACH, the death of nodes begins first, while in LEACH-ED and algorithm in this paper the first deaths begin relatively late. And apparently, in the three algorithm, the numbers of survived nodes decrease sharply after the first deaths, which indicate that the residual energy do not vary much between nodes of the network, so the time of nodes' deaths are close. It is remarkable that from about the thousandth round, the number of survived nodes in the improved algorithm remains flat for a period of time, and it's survived nodes are more than that in the other two algorithms. The lifetime in the improved algorithm is extended.

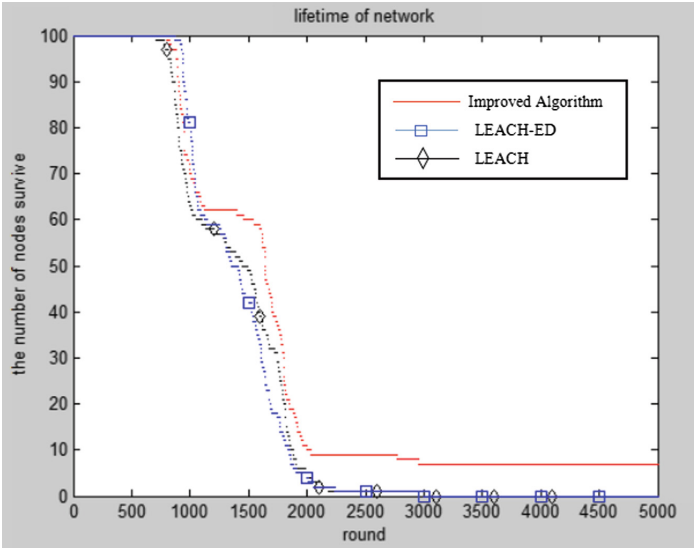


Fig. 2. The numbers of survived nodes change with time in different algorithm

Simulation Results in Quality of Communication. Assuming that the malicious nodes only transmit invalid packets before they are identified, so the malicious nodes seriously affect the communication quality of the network. Besides, the normal nodes lost packets sometime, which also affect the communication quality. Assuming that the node's packet loss rate is related to its trust value, the amount of data transmitted to the base station can be increased by selecting the nodes who have high trust values as cluster heads. In this experiment, the ratio of valid packets delivered successfully over the expected number of packets delivered to base station (PDR) is used to measure the communication quality of the network. Figure 3 shows the change of PDR with time in each algorithm:

It is clearly observed that the communication quality is significantly higher in the improved algorithm than in the other two algorithms. Because the new algorithm has improved in three aspects: first, the twice identifying trust model is used to find the malicious nodes and keep them out from the network, so that they can not affect the communication quality; second, through setting a trust value threshold *TH*, the nodes have the chance to compete for cluster heads only when their trust values exceed *TH*, which makes the cluster heads have high quality on average; finally, factor of trust value is added into the calculation of election threshold $Th_i(t)$, which further improve the quality of cluster heads on average. Our algorithm gives more chances to the nodes who behave better in several ways, so the amount of valid packets delivered to the BS can be increased a lot which means a good communication quality. In the LEACH-ED, the PDR is always lower than that in LEACH, because the malicious node usually has a high level of energy in order to compete for the cluster head, however, the LEACH-ED gives more chances to the nodes with high energy, which makes the malicious nodes

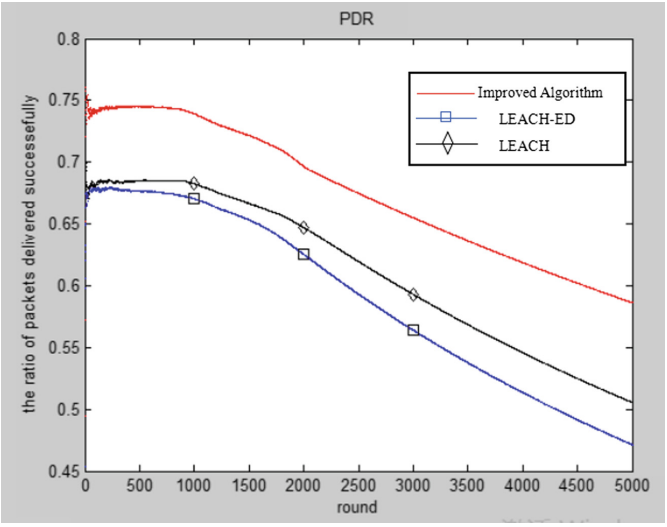


Fig. 3. The change of PDR with time in each algorithm

become cluster heads more easily. When the malicious nodes become cluster heads, they will attack the network, let alone the influence on communication quality.

5 Conclusion

With the rapid development of WSN, the security and energy efficiency become two important parts. The paper proposes a new algorithm which combines the energy model with the trust model and creates mechanism for identifying malicious nodes, so the security and energy efficiency are both improved. The simulation and analysis have proved that the new algorithm has a better performance than the traditional algorithms.

References

1. Khan, W.Z., Saad, N.M, Aalsalem, M.Y.: An overview of evaluation metrics for routing protocols in wireless sensor networks. In: 2012 4th International Conference on Intelligent and Advanced Systems, pp. 588–593 (2012)
2. Li, X.H, Hong, S.H., Fang, K.L.: A heuristic routing protocol for wireless sensor networks in home automation. In: 2009 5th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1–8 (2009)
3. Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless sensor networks. In: Proceedings of Hawaii International Conference on System Sciences. Hawaii (2000)
4. Yao, Y., Gehrke, J.: The cougar approach to in-network query processing in sensor networks. In: SIGMOD Record (2002)

5. Chakraborty, S., Khan, A.K.: Evaluation of wireless sensor network routing protocols with respect to power efficiency. In: 2013 5th International Conference on Computational Intelligence and Communication Networks, pp. 123–128 (2013)
6. Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. IEEE Trans. Wirel. Commun. **1**(4), 660–670 (2002)
7. Younis, O., Fahmy, S.: HEED: a hybrid energy-efficient distributed clustering approach for ad hoc sensor networks. IEEE Trans. Mob. Comput. **3**(4), 660–669 (2004)
8. Tao, Z., Jiang, S.F.: Clustering algorithm for wireless sensor networks with mobile cluster heads. Comput. Eng. Appl. **52**(5), 75–78 (2016)

Red Preserving Algorithm for Underwater Imaging

Chunbo Ma(✉) and Jun Ao(✉)

Guangxi Key Laboratory of Precision Navigation Technology and Application, Guilin University of Electronic Technology, Guilin, People's Republic of China
machunbo@guet.edu.cn

Abstract. The Gray World algorithm can remove the green or blue cast in underwater images. However, when one component of RGB is very little, it would lead to supersaturate and color distortion. In this paper, an improved Gray World algorithm, called Red Preserving is proposed. The minimum color component has been least changed, and the green or blue cast in the underwater images is suppressed. Experiments show that the proposed algorithm is simple and efficient. Compared to the classic Gray World algorithm, it can better suppress the cast and restore the images.

Keywords: Gray World · Green-blue cast · Red Preserving · Underwater · Image processing

1 Introduction

When light travels through water, the absorption varies greatly depending on the wavelength and water quality [1]. Figure 1 illustrates the absorption of sea water to the light with different wavelength. We can see that the green-blue light has the strongest penetrate capability, and the red light's penetrate capability is the weakest. That is why the green or blue cast is existed in underwater images. Contrary to this fact, traditional image enhancement tools, e.g., high pass filtering and histogram equalization are typically spatially invariant. Since they don't model the spatially varying distance dependencies, traditional methods are of limited utility in countering visibility problems. As the development of imaging equipment underwater, how to calibrate color in the underwater images by image processing technology is one of the research hotspots.

Currently, the major issues in underwater image processing include enhancing contrast, image denoising, color correction, and so on. In this paper, we will emphasize on how to efficiently and automatically remove the green or blue cast, and correct color distortion. At present, the classic algorithms, such as gray world, perfect reflection and single or multiscale retinex, are widely used in underwater imaging. According to the shortcomings of different algorithms, researchers are making their effort to improve them. For example, the underwater images processed by Gray World algorithm will tend to be faint red. It means that the processed images contain too much red component. To solve this problem, this paper proposed a new algorithm, called Red preserving. In this algorithm, the color components are inversed to its proportion. Since the Red component is relatively very little in underwater images, it basically keeps unchanged.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 110–116, 2017.
DOI: 10.1007/978-981-10-3969-0_13

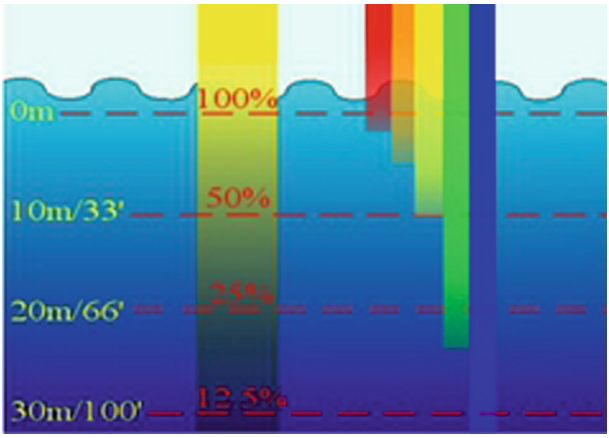


Fig. 1. Absorption characteristics of water

2 Related Works

As mentioned in [2], the single-scale retinex [3] can't simultaneously provide dynamic range compression and tonal rendition. Its subsequent version, the multiscale retinex with color restoration can solve this problem. This algorithm is quite automatic and simple, and is well approximate the performance of human vision. The properties make the multiscale retinex with color restoration used for smart camera and other wide dynamic range color imaging systems.

The gray world algorithm is suitable for the image that has sufficiently varied colors and the average of each color component tend to the same gray. Then the color of the illuminant is determined by the shift from gray of the measured average on the three channels.

The key issue of the white balance algorithm is looking for a white patch in the image. And the chromaticity of white patch is been considered as that of the illuminate. The white patch is evaluated as the maximum found in each of the three image bands separately [4]. The scaling coefficients are now obtained comparing these maxima with the values of the three channels of the chosen reference white.

Edwin Land proposed an image enhancement method, called Retinex method based on color theory. This method [5] tries to simulate the adaptation mechanisms of the human vision, performing both color constancy and dynamic range enhancement. It has some relationship with the white patch algorithm, and the different of these two methods are that the Retinex take into account the spatial relationships in the scene.

In 2013 Hitam et al. [6] proposed a new technique called hybrid Contrast Limited Adaptive Histogram Equalization (CLAHE) on the RGB and HSV color spaces. Experiment results show that the method considerably improves the visual quality of underwater images by enhancing contrast, as well as dropping noise and artifacts.

In 2012, Shamsuddin et al. [7] presented an enhancement technique for underwater images. Their work concentrated on color diminished. In the same year, Chiang and Chen [8] researched on the underwater image enhancement by wavelength

compensation and dehazing. Their new algorithm gives back the attenuation difference along the broadcast path, and takes the pressure of the possible presence of a false light source into consideration.

Compared the light attenuation in atmosphere to that in water, Wen et al. [9] presented a new underwater optical model, and then proposed an effective enhancement algorithm with the derived model to improve the perception of underwater images.

In 2010, Carlvaris-Bianco et al. [10] estimate the depth of the scene based on the difference in attenuation among the different color channels, and proposed a new algorithm for underwater imaging. Although the algorithm can reduce the effect of haze, but it is still not perfect. The absorption in different environment will influences the dehaze results.

Until now, how to remove the cast in underwater imaging is still an open problem. Although some algorithms has better effects, but it needs more processing time. Designing a simple and efficient algorithm to dehaze underwater images still has lots of work to do.

3 Algorithm

3.1 The Gray World Algorithm

The Gray World algorithm assumes that the image contains sufficiently varied colors and the average surface color in a scene is gray. From the perspective of physics, the Gray World algorithm assumes that the average of the reflective mean value of the objects in image is constancy. This idea is applied to reduce the influence of the ambient light.

Let R, B and G denote three independent channels of an image. $M * N$ denote the number of the pixels. This algorithm can be expressed as follows.

(1) Compute the average of the total image pixels, and denote the result as $C_{average}$.

$$C_{average} = \frac{\sum R + \sum G + \sum B}{3 * M * N} \tag{1}$$

(2) The gain of each channel can be express as

$$\begin{cases} k_R = \frac{M*N*C_{average}}{\sum R} \\ k_G = \frac{M*N*C_{average}}{\sum G} \\ k_B = \frac{M*N*C_{average}}{\sum B} \end{cases} \tag{2}$$

(3) The R, G and B of the output image are as follows.

$$\begin{aligned} outGW_R &= R * k_R \\ outGW_G &= G * k_G \\ outGW_B &= B * k_B \end{aligned} \tag{3}$$

The Gray World is a basic, simple and efficient algorithm which has been widely used in digital camera. However, it has some shortcomings since it is not designed for processing underwater image. From the analysis of the following Sect. 4, we can see that by using this algorithm, the processing results of the underwater image show faint red. It means that the Red component is over computed and the color is distorted. This shortcoming is what we want to overcome. Then the Red Preserving algorithm is proposed.

3.2 The Red Preserving Algorithm

As we have mentioned above, our goal is to attenuate the green or blue cast in underwater images. The algorithm called Red Preserving is designed for the underwater environment in this section. The detailed steps are as follows.

(1) Compute the Total, which is the sum of the three channels R, G and B.

$$Total = \sum R + \sum G + \sum B \tag{4}$$

(2) Calculate the ratio of R, G and B, respectively.

$$\begin{aligned} Ratio_R &= \frac{\sum R}{Total} \\ Ratio_G &= \frac{\sum G}{Total} \\ Ratio_B &= \frac{\sum B}{Total} \end{aligned} \tag{5}$$

Then we have following formulation.

$$\begin{aligned} 1 - Ratio_R &= \frac{\sum G + \sum B}{Total} \\ 1 - Ratio_G &= \frac{\sum R + \sum B}{Total} \\ 1 - Ratio_B &= \frac{\sum R + \sum G}{Total} \end{aligned} \tag{6}$$

(3) The R, G and B in corrected image are as follows.

$$\begin{aligned} out_R &= R * (1 - Ratio_R) = R * \left(\frac{\sum G + \sum B}{Total} \right) \\ out_G &= G * (1 - Ratio_G) = G * \left(\frac{\sum R + \sum B}{Total} \right) \\ out_B &= B * (1 - Ratio_B) = B * \left(\frac{\sum R + \sum G}{Total} \right) \end{aligned} \tag{7}$$

The following verification experiments are composed of two steps. The first step is to process the underwater images by using Gray World and Red Preserving, respectively. The second step is to process separately the outputs of the last step by using

CLAHE with same parameters. The experiment results show in Fig. 2. Let take the Red component as example. The different between formulation (7) and (3) is how to calculate the corrected Red component by using the proportion of the original Red component in image. Compared to Red Preserving, the Gray World algorithm usually tends to obtain larger Red value, and makes the hue of the image faint red. On this point we can see from the marked area in the middle column in Fig. 2.

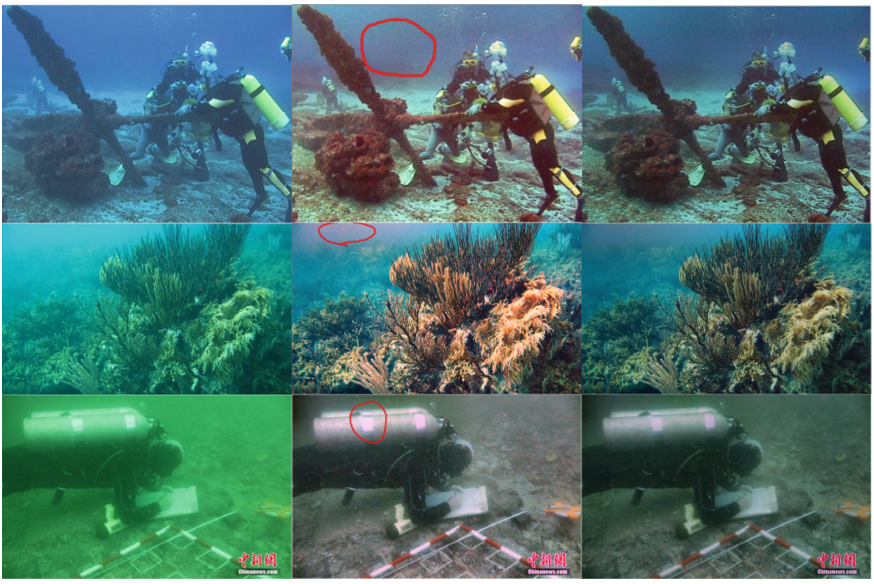


Fig. 2. From left to right, original images, images processed using Gray World algorithm, images processed using Red Preserving Algorithm

4 Analysis

(1) Protect low brightness color.
Assume that R is the low brightness color, that means the distribution of the Red is relative sparse, and the sum of value is low. Then we have following expression.

$$\lim_{R\to 0} out_R = \lim_{R\to 0} R * \left(\frac{\sum G + \sum B}{\sum R + \sum G + \sum B}\right) = R \tag{8}$$

In other words, if R is very little, then the Red Preserving algorithm will keep its value basically unchanged. However, in classic Gray World algorithm, it is very different. Let $outGW_R$ denote the R channel output in Gray World algorithm, we have

$$\lim_{R\to 0} outGW_R = \lim_{R\to 0} R * \left(\frac{\sum R + \sum G + \sum B}{3 * \sum R}\right) \tag{9}$$

When the R is very little, the out_R maybe too large to cause color distortion. This will lead to change the hue of the image.

(2) The relation between the Red Preserving and the Gray World algorithm. As described in formulation (9), the $outGW_R$ is

$$outGW_R = R * \left(\frac{\sum R + \sum G + \sum B}{3 * \sum R}\right) \tag{10}$$

Then with formulation (10), the out_R in Red Preserving can be re-presented as

$$out_R = R * \left(1 - \frac{R}{3 * outGW_R}\right) \tag{11}$$

With formulation (11), we come to the conclusion that out_R and $outGW_R$ are nonlinear relationship.

(3) Efficiently remove color cast.

The Red Preserving algorithm can efficiently remove the color cast. As we know, the green is the major color that superimpose on the image in the environment of underwater. From the formulation (7), we can see that the output brightness of a channel is decided by the other two channels. Since the brightness of Red and Blue are relatively low compared to Green in underwater environment, then the Green will multiply a relatively small coefficient, and the green cast is suppressed.

5 Conclusions

Gray World algorithm is widely used in digital camera and some portable imaging devices, since it is simple and efficient in color calibration. However, considering the deficiency of Gray World algorithm, we propose an improved Gray World algorithm, named Red Preserving. Experiments show that this algorithm overcomes the deficiency of Gray World, at the same time keeps the properties of simple and efficiency.

References

1. Torres-Méndez, L.A., Dudek, G.: Color correction of underwater images for aquatic robot inspection. In: Rangarajan, A., Vemuri, B., Yuille, A.L. (eds.) EMMCVPR 2005. LNCS, vol. 3757, pp. 60–73. Springer, Heidelberg (2005). doi:10.1007/11585978_5
2. Jobson, D.J., Rahman, Z., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE Trans. Image Process. **6**(7), 965–976 (1997)

3. Jobson, D.J., Rahman, Z., Woodell, G.A.: Properties and performance of a certer/surround retinex. IEEE Trans. Image Process. **6**(3), 451–462 (1997)
4. Funt, B., Barnard, K., Martin, L.: Is machine colour constancy good enough? In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1406, pp. 445–459. Springer, Heidelberg (1998). doi:10.1007/BFb0055683
5. Land, E.: The Retinex theory of color vision. Sci. Am. **237**(6), 108–128 (1978)
6. Hitam, M.S., Yussof, W.N.J.H.W., Awalludin, E.A., Bachok, Z.: Mixture contrast limited adaptive histogram equalization for underwater image enhancement. In: Proceedings of Computer Applications Technology (ICCA), pp. 1, 5, 20–22 (2013)
7. Shamsuddin, N.B., Wan, F.B.W.A., Baharudin, B.B., Kushairi, M.: Significance level of image enhancement techniques for underwater images. IVIC **1**, 490–494 (2012)
8. Chiang, J.Y., Chen, Y.C.: Underwater image enhancement by wavelength compensation and dehazing. IEEE Trans. Image Process. **21**(4), 1756–1769 (2011). A Publication of the IEEE Signal Processing Society
9. Wen, H.C., Tian, Y.H., Huang, T.J., Gao, W.: Single underwater image enhancement with a new optical model. In: IEEE International Symposium on Circuits and Systems (ISCAS 2013), pp. 753–756 (2013)
10. Carlevaris-Bianco, N., Mohan, A., Eustice, R.M.: Initial results in underwater single image dehazing. Oceans **27**, 1–8 (2010)

Estimating Gas Source Location Based on Distributed Adaptive Deflection Projected Subgradient Method

Zhemin Zhuang(✉), Fenlan Li, and Ye Yuan

Department of Electronics Engineering, Shantou University, Guangdong, China
{zmzhuang, lifenlan, yuanye}@stu.edu.cn

Abstract. A novel method based on distributed adaptive deflection projected subgradient is proposed in this paper. By used of attenuation model of a gas source in wind field, the method is able to process the distributed information from sensors by using the developed algorithm, replace the original gradient direction with deflection subgradient direction, and utilize the deflected subgradient projection hyperplanes as the searching areas in the process of relaxed projection, so as to obtain the gas source position. A related simulation provided in the paper illustrates that this method can not only provide good convergence property and accurate localization results, but also save large amount of energy.

Keywords: Wireless Sensor Network · Gas source localization · Distributed · Deflection · Projected subgradient

1 Introduction

Wireless Sensor Network (WSN), is composed by a large number of randomly distributed sensor nodes, and it has the merits of low-cost and low-power through self-organizing means. These tiny sensors nodes which consist of sensing, data processing and communicating components are collaborated to perform different tasks.

Gas source localization technology in WSN has two types at present: centralized localization schemes and distributed localization schemes. The centralized localization schemes require each sensor node to send measured information to the center, and consequently estimates the source position. In these schemes, not only the localization results become unsatisfactory, but also they consume too much energy and easily make the local networks paralyze. The distributed schemes can solve these problems. In the schemes, the data of each sensor node is processed, the center is essentially liberated and the amount of communication is reduced, thus the distributed scheme can enhance the efficiency and prolong the lives of the networks. The algorithms of distributed scheme include the Incremental Gradient (IG) algorithm [1], the distributed subgradient projection algorithm [2], the Projection of onto Convex Sets (POCS) algorithm [3], and so on. However, depending on the choice of the unknown initial values, these distributed algorithms may fail due to local minima and poor convergence. Especially, it's difficult to realize POCS algorithm because the exact expression of the orthogonal projection is hard to obtain. Consequently, a Distributed Adaptive Deflection Projected

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 117–126, 2017.
DOI: 10.1007/978-981-10-3969-0_14

Subgradient Method (DADPSM) for gas source localization is proposed. The method can process the distributed information from the sensors, and utilize the deflection projected subgradient hyperplanes as the searching areas in the process of relaxed projection to achieve the gas source position. Computer simulation results show that this proposed approach has better gas source localization performance than the results of other distributed methods.

2 Gas Concentration Attenuation Model

With the presence of wind, the dispersal behavior of gas is, as well as diffusion, also characterized by wind [4]. To obtain the gas concentration attenuation model for source localization, we make the following assumptions:

The gas propagates outwardly at a constant rate, and there is no environmental change throughout the propagation. We don't consider the effects of temperature and obstacles, meanwhile, the velocity and direction of wind are both fixed.

A single gas source is located at a stationary position $\vec{g} = (x_0, y_0)$, which is also randomly placed in WSN field, and the measured concentration at the source is at a constant of $Q(mg/s)$.

A set of N sensor nodes are stationary, randomly placed in WSN field at the positions of $\vec{r}_i = (x_i, y_i)$, $i = 1, 2, \ldots, N$. Supposing that the sensors detect the presence of the gas source if their concentration measurements are above T, which is named as the detection threshold.

Based upon the above conditions, we assume that there are M sensor nodes which are employed to detect the presence of the source. By Fick's law of diffusion [5], the gas concentration received by the i th sensor node is expressed as

$$c_i = \frac{Q}{4\pi k|\vec{r}_i - \vec{g}|}\exp\left\{\frac{(\vec{r}_i - \vec{g})\cdot\vec{v} - |\vec{r}_i - \vec{g}||\vec{v}|}{2k}\right\} + \varepsilon_i \quad i = 1, 2, \ldots, M, M \leq N \tag{1}$$

where $k(m^2/s)$ is the diffusivity, which is influenced by environment temperature, $\vec{v} = (v_x, v_y)$ is the velocity of wind, $|\vec{r}_i - \vec{g}|$ is the distance between the ith sensor node and the source, ε_i is the additive Gaussian noise [6, 7].

In terms of the measured concentration of sensor node $\vec{r}_i$, we obtain the distance function between the ith sensor node and the source position $\vec{g}$.

$$d_i(x_0, y_0) = \frac{2k}{|\vec{v}|}Lambertw\left(\frac{|\vec{v}|Q}{8\pi k^2 c_i} \times \exp\left(\frac{v_x(x_i - x_0) + v_y(y_i - y_0)}{2k}\right)\right) \tag{2}$$

where *Lambertw* ( ) is the Lambert W function. If the measurements of concentration C_i are disturbed by noise, the corresponding ovals don't intersect at the source position. To achieve its position, we formulate the objective function (3) for localizing the source:

$$J(\vec{g}) = \sum_{i=1}^{M} [|\vec{r}_i - \vec{g}| - d_i(x_0, y_0)]^2 \tag{3}$$

as a matter of fact, the localization of the gas source $\vec{g}$ is also the problem to estimate.

$$\arg\min_{\vec{g}} J(\vec{g}) = \arg\min_{\vec{g}} \sum_{i=1}^{M} [|\vec{r}_i - \vec{g}| - d_i(x_0, y_0)]^2 \tag{4}$$

In the disturbed case $\boldsymbol{J}(\vec{g})$ it doesn't become 0 for $\vec{g}$ being the intersection of the ovals. We firstly apply adaptive projection sub-gradient method to solve this optimization problem.

3 Adaptive Projected Subgradient Method (APSM) for Gas Source Localization

We define $d(\vec{a}, \vec{b}) = \left\|\vec{a} - \vec{b}\right\|$ in the real Hilbert space $(\boldsymbol{\gamma}, <\bullet, \bullet>)$, and $\vec{a}, \vec{b} \in \boldsymbol{\gamma}$ is the distance between two points $\vec{a}$ and $\vec{b}$ in γ. $(\vec{g}_k)_{k\in K}$ is assumed to be the estimation of gas source position at the kth iteration by the ith sensor node, and K denotes the maximum iteration. Let's define $d(\vec{g}_k, \boldsymbol{C}_k) = \|\vec{g}_k - \boldsymbol{P}_{\mathbf{C}_k}(\vec{g}_k)\|$, $\forall \vec{g}_k \in \boldsymbol{\gamma}$ indicates the distance between $\vec{g}_k$ and the closed convex set C_k. If the true position of gas source $(\vec{g}_k)_{k\in K} \subset \boldsymbol{\gamma}$ is satisfied, a sequence of the source position estimation $\vec{g} \in \boldsymbol{C}_k$ can be obtained iteratively as Eq. (5) [8–10],

$$\vec{g}_{k+1} = \vec{g}_k + \lambda_k [P_{C_k}(\vec{g}_k) - \vec{g}_k], \quad \forall k \in K \tag{5}$$

where $\lambda_k \in [0, 2]$. Since the projection onto the convex sets C_k is difficult to be obtained in most cases, it is necessary to substitute a new approximate method for the orthogonal projection $\boldsymbol{P}_{C_k}(\vec{g}_k)$ in Eq. (5). Thus the projected subgradient technology is introduced to simplify the projection.

The closed convex set $\boldsymbol{C}_k = \{\vec{x} \in \boldsymbol{\gamma} : \varphi_k(\vec{x}) \le 0\}$ is defined, here $\varphi_k(\vec{x}), \vec{x} \in \boldsymbol{\gamma}$ is a convex function. If (6) is satisfied,

$$<(\vec{x} - \vec{g}_k), \vec{t}> + \varphi_k(\vec{g}_k) \le \varphi_k(\vec{x}), \quad \forall \vec{x} \in \gamma \tag{6}$$

$\vec{t}$ is called a sub-gradient of $\varphi_k(\vec{x})$ at point $\vec{r}_k$. Taking the affine half space as

$$H^-(\vec{x}) = \{\vec{x} \in \gamma : f(\vec{x}) \le 0\} = \{\vec{x} \in \gamma : <(\vec{x} - \vec{g}_k), \vec{t}> + \varphi_k(\vec{g}_k) \le 0\} \tag{7}$$

If $\vec{g}_k \notin C_k$, then $\vec{g}_k \notin H^-(\vec{x}) \subset C_k$, and the hyperplane $H^-(\vec{x})$ always separates $\vec{g}_k$ and C_k. Thus the projection onto the convex set C_k can be expanded to the projection onto the half space $H^-(\vec{x})$, which has a simple closed-form expression. If $\varphi_k(\vec{x})$ is continuous at point $\vec{t}$, $\vec{g}_k$ has a unique value, i.e. the selected subgradient $\nabla\varphi_k(\vec{g}_k)$, and

then the projection from $\vec{g}_k$ to $H^-(\vec{x})$ has the form of Eq. (8) is the projection equation of APSM.

$$P_{H^-(\vec{g}_k)}(\vec{g}_k) = \begin{cases} \vec{g}_k & ,\vec{g}_k \in H^-(\vec{g}_k) \\ \vec{g}_k + \frac{-\varphi_k(\vec{g}_k)}{\|\nabla\varphi_k(\vec{g}_k)\|^2}\nabla\varphi_k(\vec{g}_k) & ,\vec{g}_k \notin H^-(\vec{g}_k) \end{cases} \tag{8}$$

According to Eq. (4), the set $D_i = \left\{\vec{g} \in R^n : \|\vec{g} - \vec{r}_r\|^2 \le d_i^2(x)\right\}$ is defined, thus, we can solve the gas source localization problem by letting the estimator be the intersection of the sets D_i. We define the following convex function

$$\varphi(\vec{g}_k) = \|\vec{g} - \vec{r}_r\|^2 - d_i^2(x_0, y_0), \vec{g}_k \in R^n \tag{9}$$

and the convex set is as below

$$C_k(d_i(x_0, y_0)) = \left\{\vec{g}_k \in R^n : \|\vec{g}_k - \vec{r}_r\|^2 \le d_i^2(x_0, y_0)\right\} = \{\vec{g}_k \in R^n : \varphi(\vec{g}_k) \le 0\} \tag{10}$$

For the reason that the closed convex set C_k contains the true position of gas source with high probability, the issue of estimation of $\vec{g}$ is changed into the issue of projection to the convex set C_k. Here the gradient operator is $\vec{t} = \nabla\varphi(\vec{g})$, then the half space based on this convex function can be defined as

$$H^-(\vec{g}) = \left\{\vec{g}_k \in R^n : (\vec{r}_i - \vec{g}_k)^T\vec{t} + \varphi(\vec{g}_k) \le 0\right\} \tag{11}$$

which satisfies $\vec{g}_k \notin H^-(\vec{g})$ and $C_k \subset H^-(\vec{g}_k)$ if $\vec{g}_k \notin C_k$. Hence the projection onto $C_k(d_i(x_0, y_0))$ can be expanded to the projection onto the half space $H^-(\vec{g})$, which has the same expression as Eq. (9). The iterative update of gas source position can be obtained by

$$\vec{g}_{k+1} = \vec{g}_k + \lambda_k[P_{H^-(\vec{g}_k)}(\vec{g}_k) - \vec{g}_k] \tag{12}$$

4 Distributed Adaptive Defection Projected Subgradient Method (DADPSM) for Localization

At some stage in the actual computation, the whole WSN may carry a large amount of communication as shown in Fig. 1, all the measured information should be sent to the center **S**, and the network congestion can increase the response time of WSN, however reduce efficiency and easily cause the local network to paralyze. As a result, a distributed localization algorithm to deal with these problems is employed. Different from the centralized localization method, the distributed localization method doesn't require all sensor nodes to participate in calculation, and the computation is performed at each of sensor nodes. First of all, we assume that the network is in a cyclic fashion as shown in Fig. 2. Messages are passed between nodes in the order $1, 2, \ldots, M-1, 1, 2, \ldots, M-1, M$.

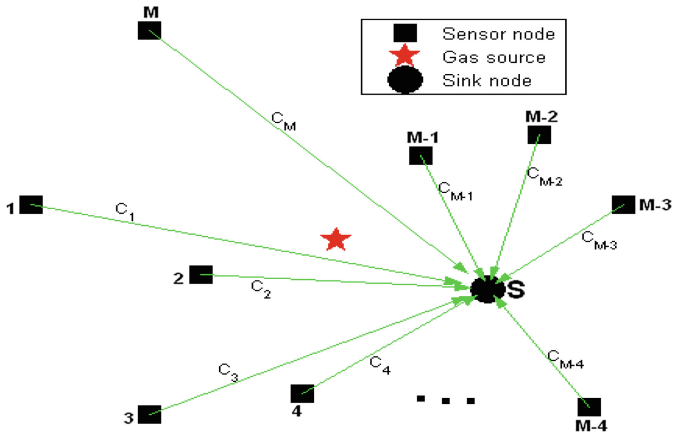


Fig. 1. Centralized network structure

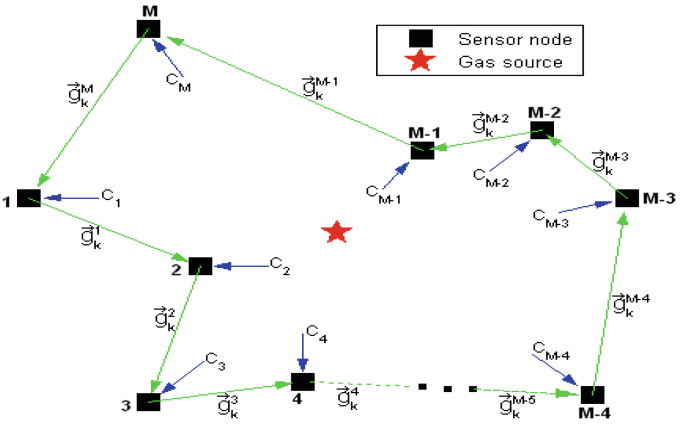


Fig. 2. Distributed network structure

In Fig. 2, c_i is the measured concentration by the ith sensor node, and $\vec{g}_k^i$ denotes the estimation of gas source position at the kth cycle by the ith sensor node. According to Eq. (4), the objective function for localizing the source contains M component functions as below:

$$\arg\min_{\vec{g}} J(\vec{g}) = \arg\min_{\vec{g}} \sum_{i=1}^{M} \left[|\vec{r}_i - \vec{g}| - d_i(x_0, y_0)\right]^2 = \arg\min_{\vec{g}} \sum_{i=1}^{M} J_i(\vec{g}) \tag{13}$$

Hence in the algorithm proposed in each cycle, each sensor node estimates the gas source location by the Eq. (14),

$$\vec{g}_k^i = \arg\min J_i(\vec{g}_k^{i-1}, c_i), 1 = 1, 2, \ldots, M, \text{ where } M \leq N \tag{14}$$

Namely, the $(i-1)$th sensor node sends its estimated location $\vec{g}_k^{i-1}$ to the ith sensor node, and then the ith sensor node updates the estimation of gas source position using its own gas concentration data and the received $\vec{g}_k^{i-1}$. After obtaining the new estimated location $\vec{g}_k^i$, the ith sensor node sends it to the $(i+1)$th sensor node, and this procedure is continued until it reaches convergence. However, to the APSM for localization, the main point of this algorithm is to gain the subgradient $\nabla\varphi(\vec{g})$, if c_i is disturbed by noise, the objective function (10) becoming non-convex has multiple local optima and saddle points, and the APSM may stagnate at one of these suboptimal solutions instead of converging to the optimal one (i.e., poor convergence). Meanwhile, from Eq. (5), when the direction of subgradient $\nabla\varphi\left(\vec{g}_k^i\right)$ at the point $\vec{g}_k^i$ forms an obtuse angle with the previous subgradient $\nabla\varphi\left(\vec{g}_k^{i-1}\right)$ at the point $\vec{g}_k^{i-1}$, the next estimated position will be very close to the point $\vec{g}_k^{i+1}$. Obviously, it's not helpful for the estimated result to be improved by this iteration $\vec{g}_k^{i-1} \to \vec{g}_k^i \to \vec{g}_k^{i+1}$, so the convergence rate of it becomes slow. Here, the subgradient $\nabla\varphi\left(\vec{g}_k^{i-1}\right)$ is obtained by the ith sensor node using $\vec{g}_k^{i-1}$ at the kth cycle. In order to form an acute angle between the next iterative direction and the current direction of subgradient, we propose a scheme which is to apply the deflection subgradient ϕ_k^i to take place of $\nabla\varphi\left(\vec{g}_k^{i-1}\right)$ from the current sensor node i at the kth cycle.

$$\phi_k^i = \nabla\varphi\left(\vec{g}_k^{i-1}\right) + w_k^i \times \phi_k^{i-1}, \quad w_k^i > 0 \tag{15}$$

Where $\boldsymbol{w}_k^i$ is a deflection factor, and if $i = 0$, $\boldsymbol{\Phi}^i = 0$.

The nearer the distance between the gas source position and the sensor nodes, the less the influence is caused by the noise, so the weight affecting the estimation of gas source position should be larger. Therefore, we introduce a concept of weighting ratio to reflect the distributed character of sensor nodes and gas source in the network field. Simultaneously, with regard to the estimated source location $\vec{g}_k^i = \left(x_{0,k}^i, y_{0,k}^i\right)$, the subgradient of each dimension possesses of different weights during the iteration, and the large one can speed up the convergence. So we modify the deflection factor $\boldsymbol{w}_k^i$:

$$w_k^i = \left(\frac{\left\|\nabla\varphi(x_{0,k}^{i-1})\right\|}{\left\|\nabla\varphi(x_{0,k}^{i-1})\right\| + \left\|\nabla\varphi(y_{0,k}^{i-1})\right\|}, \frac{\left\|\nabla\varphi(y_{0,k}^{i-1})\right\|}{\left\|\nabla\varphi(x_{0,k}^{i-1})\right\| + \left\|\nabla\varphi(y_{0,k}^{i-1})\right\|} \right) \tag{16}$$

Meanwhile, the Eq. (10) will be replaced by

$$P'_{H^-(\vec{g}_k^{i-1})}(\vec{g}_k^{i-1}) = \begin{cases} \vec{g}_k^{i-1} & , \vec{g}_k^{i-1} \in H^-(\vec{g}_k^{i-1}) \\ \vec{g}_k^{i-1} + \frac{-\varphi(\vec{g}_k^{i-1})}{\left\|\phi_k^i\right\|^2}\phi_k^i & , \vec{g}_k^{i-1} \notin H^-(\vec{g}_k^{i-1}) \end{cases} \tag{17}$$

To avoid the algorithm exhibits an oscillation phenomenon, and to consider the modulus of each dimension parameter's projection, we obtain Eq. (18) by modifying Eq. (15),

$$\vec{g}_k^i = \vec{g}_k^{i-1} + \lambda_k \left[\frac{P'_{H^-(\vec{g}_k^{i-1})}(x_{0,k}^{i-1}) - x_{0,k}^{i-1}}{\left\| P'_{H^-(\vec{g}_k^{i-1})}(x_{0,k}^{i-1}) \right\|}, \frac{P'_{H^-(\vec{g}_k^{i-1})}(y_{0,k}^{i-1}) - y_{0,k}^{i-1}}{\left\| P'_{H^-(\vec{g}_k^{i-1})}(y_{0,k}^{i-1}) \right\|} \right], \quad i = 1, 2, \ldots, M \quad (18)$$

Where M is the number of active sensor nodes, $\vec{g}_1^0 = \left(x_{0,1}^0, y_{0,1}^0\right)$ can be arbitrarily initialized, and $\vec{g}_{k+1}$ denotes the estimation of gas source position at the kth cycle.

5 Computer Simulation Results

The proposed DADPSM is employ to solve the gas source localization problem. In the simulations, 40 sensor nodes are deployed uniformly at random in a 50 m * 50 m field, and each sensor node is used to measure the gas concentration. The detection threshold T is adjusted against the current noise evaluation μ to guarantee constant false ratio, meanwhile, at least three sensor nodes is needed to uniquely find $\vec{g} = (x_0, y_0)$. Thus T is needed to be set to a suitable value. The gas source is located at $\vec{g} = (17\,\text{m},\ 21\,\text{m})$, and Q = 24 mg/s. The gas diffusivity k = 0.08 m^2/s, the wind velocity $|\vec{v}| = 0.03\,\text{m/s}$ and wind direction = 30° are known by anemoscope.

Figures 3 and 4, have performed simulations to compare the performance of our method (DADPSM) with POCS, APSM and IG, where T = 0.8 mg/m^3 and the background noise has $N(0, 0.02)$ distributed, the initial points of these algorithms are all set to the origin(0, 0).

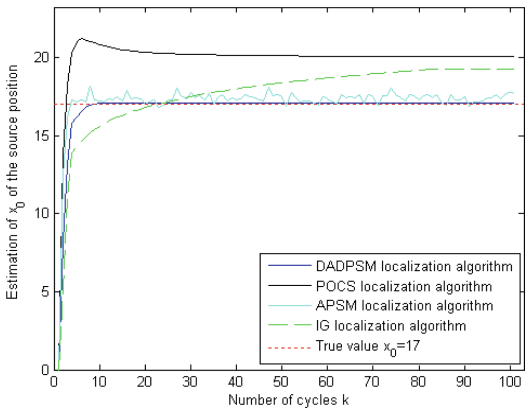


Fig. 3. X-coordinate's estimation of source's position by different distributed localization algorithms

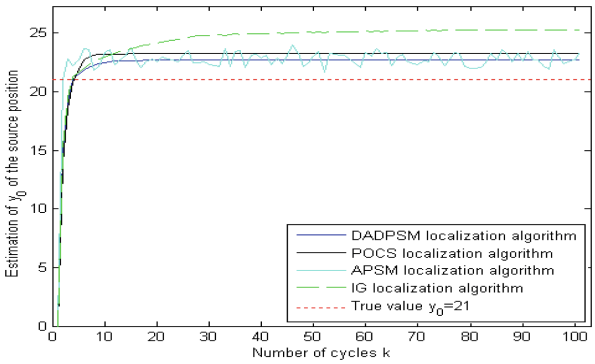


Fig. 4. Y-coordinate's estimation of source's position by different distributed localization algorithms

Figures 5 and 6 show the localization error comparison of different algorithms with different numbers of active sensor nodes, and different background noise, respectively.

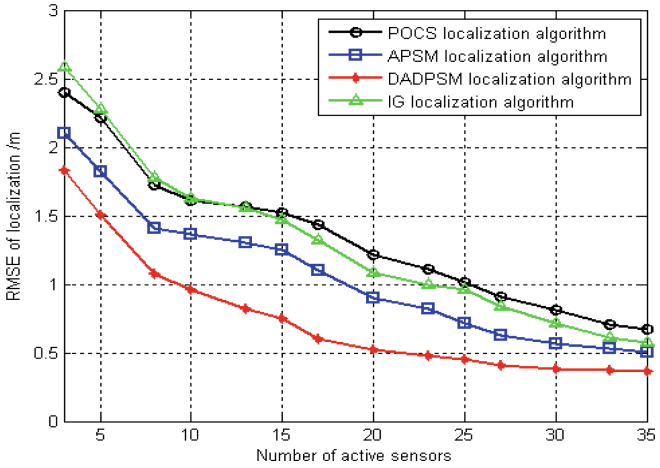


Fig. 5. RMSE vs. active sensor nodes

Figure 5 illustrates that the RMSE of the four distributed algorithms decreases with the increasing of active sensor nodes, however when the number of the nodes increases to a certain extern, the RMSE descending becomes unobvious. The cause for this is to suppose that the participation of many sensor nodes in the computation reduces the RMSE, however much more background noise is applied to localize the gas source. Nevertheless, in case of same number of sensor nodes, the DADPSM localization method performs better results than other three distributed localization methods. Figure 6 shows that the accuracy of localization is affected by background noise, the RMSE of estimated source position increases with the enhancement of the background noise variance.

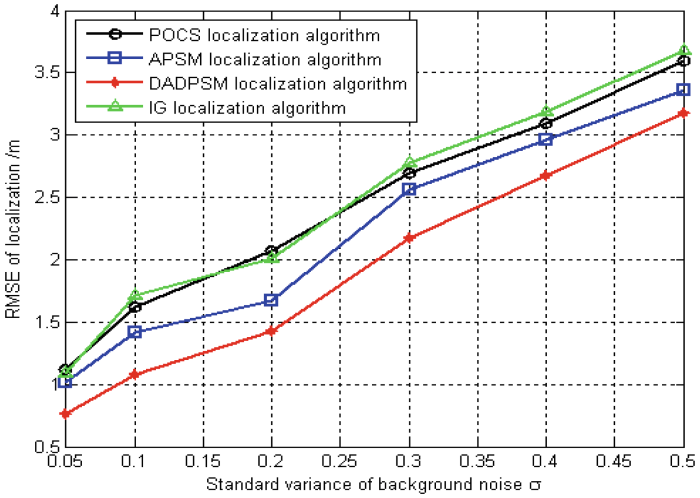


Fig. 6. RMSE vs. the variance of background noise

The localization results and time-consuming of these four distributed localization algorithms are shown in Table 1.

Table 1. Comparison of localization results and time-consuming of different algorithms under different noise and source's position

	Background noise standard variance $\sigma = 0.1$		Background noise standard variance $\sigma = 0.2$	
	Gas source position (8 m, 13 m)	Time consuming	Gas source position (15 m, 20 m)	Time consuming
POCS algorithm	(7.1502 m, 14.5622 m)	3.7871 (s)	(17.3287 m, 21.5218 m)	4.3071 (s)
APSM algorithm	(9.3055 m, 13.6393 m)	3.8555 (s)	(16.1222 m, 22.0931 m)	4.3294 (s)
DADPSM algorithm	(9.1568 m, 13.2784 m)	3.5268 (s)	(16.6746 m, 21.1654 m)	4.2835 (s)
IG algorithm	(9.6280 m, 14.0381 m)	3.9162 (s)	(17.0231 m, 21.8142 m)	4.5162 (s)

Table 1 shows that the proposed DADPSM algorithm is the most robust and can implement rapidly. The larger the background noise standard variance, the more the time-consuming of each algorithm. The reason of this is that the objective function for localization has more local optimal saddle points, and then each distributed localization algorithm is more difficult to reach the gas source location.

6 Conclusion

This paper constructs an objective function for localization based on the attenuation model of a gas source in the wind field, since the objective function is quite complex, it's easy of the network to be paralysis by some centralized localization method.

However, this paper proposed a distributed localization approach called DADPSM, which can converge to the vicinity of the true gas source with less time, yields much better results for the source position and performs very well in energy-saving. For the communication, the active sensor nodes only communicate with neighboring sensor nodes in the distributed localization algorithm, but in centralized localization algorithm, all the information must be sent to the center, so the former consumes less energy. Future research is directed to the multiple gas source localization with the proposed distributed localization approach in WSN.

Acknowledgements. This work was financially supported by the Foundation of China (No. 61471228) and the Key Project of Guangdong Province Science & Technology Plan (No. 2015B020233018).

References

1. Wang, C.L., Wu, D.S.: Decentralized positioning and tracking based on a weighted incremental subgradient algorithm for wireless sensor networks. In: Proceedings of IEEE Vehicular Technology Conference, Canada, pp. 1–5 (2008)
2. Sundhar, R.S., Nedic, A., Veeravalli, V.: Distributed subgradient projection algorithm for convex optimization. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Taiwan, pp. 3653–3656 (2009)
3. Blatt, D., Hero, A.O.: Energy-based sensor network source localization via projection onto convex sets. IEEE Trans. Sig. Process. **54**, 3614–3619 (2006)
4. Fukazawa, Y., Ishida, H.: Estimating gas-source location in outdoor environment using mobile robot equipped with gas sensors and anemometer. In: Proceedings of IEEE Sensors Conference, New Zealand, pp. 1721–1724 (2009)
5. Crank, J.: The Mathematics of Diffusion. Oxford Uni. Press, Oxford (1956)
6. Zhang, Y., Wang, L.: A particle filtering method for odor-source localization in wireless sensor network with mobile robot. In: Proceedings of 8th World Congress on International Control and Automation, China, pp. 7032–7036 (2010)
7. Ampeliotis, D., Berberidis, K.: Low complexity multiple acoustic source localization in sensor networks based on energy measurements. Sig. Process. **90**, 1300–1312 (2010)
8. Yukawa, M., Slavakis, K.: Signal processing dual domain by adaptive projection subgradient method. In: Proceedings of International Conference Digital Signal Processing, Greece (2009)
9. Nedic, A., Ozdaqlar, A., Parrilo, P.: Constrained consensus and optimization in multi-agent networks. IEEE Trans. Autom. Control **55**, 922–938 (2010)
10. Yamadaa, I., Ogura, N.: Adaptive projected subgradient method for asymptotic minimization of sequence of nonnegative convex functions. Numer. Funct. Anal. Optim. **25**, 593–617 (2004)
11. Stark, H., Yang, Y.: Vector Space Projections—A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics. Wiley, New York (1998)

System Locating License Plates with Shadow Based on Self-adaptive Window Size Technique

Jingyu Dun(✉) and Sanyuan Zhang

College of Computer Science and Technology,
Zhejiang University, Hangzhou, China
{dunjingyu, syzhang}@zju.edu.cn

Abstract. Most of the existing license plate localization algorithms have a parameter that is related to the size of the license plate. There is no parameter that is suitable for all the cases. In this paper, an algorithm is proposed to automatically compute the size-related parameter. Then a hierarchical system based on the self-adaptive parameter is proposed to locate license plates. Both connected component based methods and vertical edge based methods are used. The parameter is first used as the local window size to suppress the shadow. Then it is used to connect the discrete vertical edges to form a license plate region. The proposed system is used to locate license plates with shadow, and experiments are taken on images with different resolutions. The total localization accuracy achieves 94.40%. It can compete with the state-of-the-art methods and need not determine the optimal parameter by trial and error.

Keywords: License plate localization · Self-adaptive window size · Shadow suppression

1 Introduction

License plate localization is an important part of an intelligent transportation system, because the license plate is a vital component of a vehicle. It can be used to track vehicles, supervise vehicle behavior, and so on. There are three significant properties of a license plate: (1) it is composed of several characters with similar size; (2) it is rich in textures; (3) it has a rectangular shape and a fixed aspect ratio for a specific country. In fact, most of the existing license plate localization methods are based on the three properties. Connected component (CC) based methods [1, 9, 16] use the first and third properties. Vertical edge (VE) based methods [5, 8, 14] and the machine learning (ML) based methods [3, 4, 13, 17] typically rely on the second and third properties. However, no matter what properties they use, most of them have a parameter that is related to the size of the license plate. That is, by selecting a parameter value, the algorithm is adapted to detecting license plates whose size is in a specific range.

Image binarization is the first step of CC-based methods. Generally speaking, the thresholding algorithms are divided into two categories: global thresholding methods [7, 12] and local thresholding methods [2, 6, 10, 15]. Given the complex illumination

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 127–137, 2017.
DOI: 10.1007/978-981-10-3969-0_15

condition of the image, global thresholding methods are inferior to local thresholding methods. However, when using the local thresholding methods, a block size should be manually set. This parameter determines how much pixels around the central pixel are used to compute the local threshold value. Its value should be carefully selected depending on the size of the characters.

VE-based methods usually connect the discrete vertical edges by mathematical morphological operations. The mathematical morphological operations have a parameter that defines the neighbor pixels used by the operator. This parameter also depends on the size of the character and the distance between the characters.

The sliding window technique is typically used by the VE-based methods and the ML-based methods. The size of the window and the search step are related to the size and position of the license plate. If multi-scale search is used, the algorithms will be more robust. However, the scaling ratio still influences the size of the detected license plates.

By the analyses above, a technique is proposed to compute the size-related parameter at each pixel automatically in order to avoid the tedious manual work to select the optimal parameter. A hierarchical localization system is then proposed based on the technique. First, a CC-based method is used. If there is no result, a VE-based method is applied. Finally a window based method without multi-scale search is used. The order of the three process is important since the restrictions change from strict to loose. In this way, the system can achieve high accuracy with low false positive rate. The system is tested on images with different resolutions, and the experimental results show that the proposed system can compete with the state-of-the-art method.

The paper is organized as follows: the proposed system is presented in Sect. 2; the experiment and comparison are shown in Sect. 3; Sect. 4 summarizes the paper and makes a conclusion.

2 The Proposed System

A hierarchical localization system is proposed in this paper. Figure 1 shows the flow chart of the system. A technique is proposed to automatically compute the window size at each pixel. The window size is used in the following detection process to make the system flexible to locate license plates which vary a lot in size. The system is elaborated in detail in the following paragraphs.

2.1 Window Size Surface Generation

To find a suitable window size automatically, a simple technique is used in this paper. For pixel (x, y), firstly, scan horizontally from (x, y) to its left. Let

$$f(l) = \sum_{i=0}^{l} p(-i+x, y)/(l+1) - p(x, y), \tag{1}$$

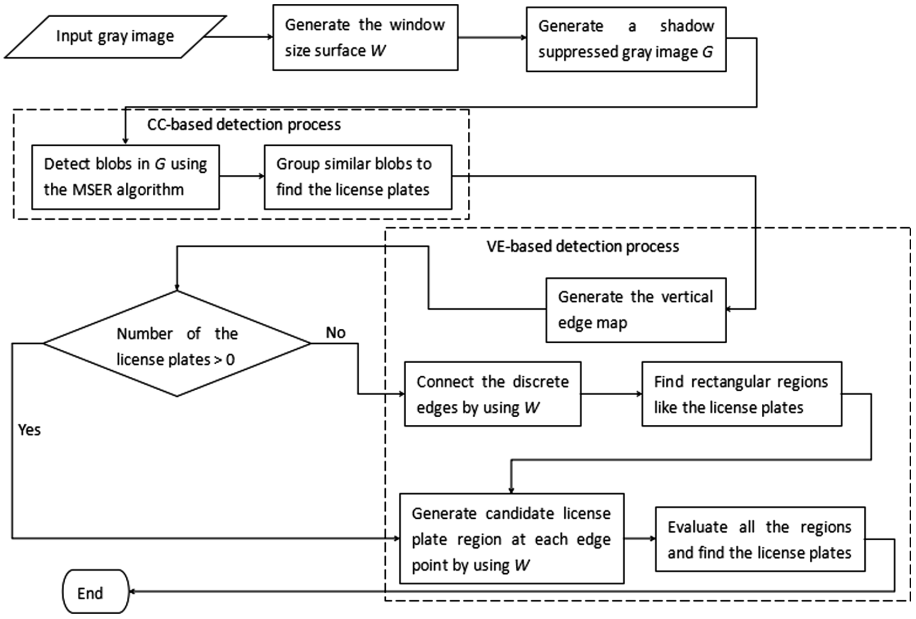


Fig. 1. The flow chart of the localization system.

where $p(x, y)$ denotes the gray scale value at point (x, y), and l stands for the half window size. If $f(l) > 0$ and $f(l-1) > f(l)$, then the window size at pixel (x, y) is set as $2l + 1$. $f(l) > 0$ ensures that the lighter background is included. $f(l-1) > f(l)$ means that the scanning reaches the edge of the next dark stroke.

Similarly, scan horizontally from (x, y) to its right. Let

$$f(l) = \sum_{i=0}^{l} p(i+x, y)/(l+1) - p(x, y). \tag{2}$$

If $f(l) > 0$ and $f(l-1) > f(l)$, then a new window size is computed. Choose the smaller one as the final window size at pixel (x, y). Figure 2 shows the window computed by the technique for the pixel marked in red.

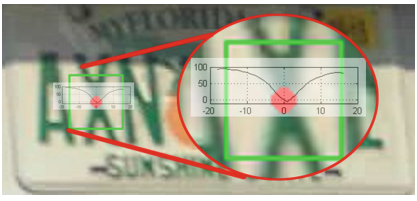


Fig. 2. An example of the automatic window size selection technique. (Color figure online)

Generally, the range of l is not limited. The process ends when the conditions are satisfied or the scanning reaches the bound of the image. However, if the stroke width of the detected characters has an upper limit, the process can be accelerated. $l \leq 50$ is used in this study.

The process mentioned above computes the window size for all the darker pixels. Inverse the image and repeat the process to compute the window size for the lighter pixels if lighter objects are to be detected.

2.2 Shadow Suppression

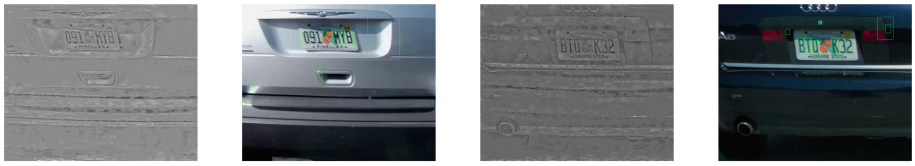
After obtaining the window size surface W, the mean matrix M and standard deviation matrix S are computed by using Eqs. (3) and (4).

$$M(x,y) = \frac{1}{W(x,y) \times W(x,y)} \times \sum_{i=-W(x,y)/2}^{W(x,y)/2} \sum_{j=-W(x,y)/2}^{W(x,y)/2} I(x+i, y+j) \tag{3}$$

$$S(x,y) = \frac{1}{W(x,y) \times W(x,y)} \times \sqrt{\sum_{i=-W(x,y)/2}^{W(x,y)/2} \sum_{j=-W(x,y)/2}^{W(x,y)/2} (I(x+i, y+j) - M(x,y))^2} \tag{4}$$

$W(x, y)$ and $I(x, y)$ denote the window size and the gray scale value at pixel (x, y), respectively. The integral image is used to improve the time efficiency. Subtract M from the original image I, and then divide the result by S to obtain the differential image G as Eq. (5) shows. In G, the shadow is suppressed. Figure 3(a) and (c) show the results of Eq. (5).

$$G = (I - M)/S \tag{5}$$



(a) (b) (c) (d)

Fig. 3. (a) and (c) are the differential images G; (b) and (d) are the blobs detected by using MSER on G.

2.3 CC-Based Detection Process

CC-based detection methods usually remain two kinds of regions: regions whose aspect ratio is similar with the characters and regions whose aspect ratio is similar with a license plate. To make the restriction tight, only the first kind of regions are detected. Since the shadow is suppressed in the last step, a global thresholding method is

sufficient. The maximally stable extremal regions (MSER) algorithm [11] uses multi-threshold values to find stable regions. By using this algorithm on G, candidate character regions will be generated.

The blobs detected by MSER algorithm are always overlapped with each other. A blob is a candidate character if there are blobs with a similar height around it. Algorithm 1 shows the process of removing the overlapped regions.

The candidate blobs detected after removing the overlapped regions are shown in Fig. 3(b) and (d). Finally, blobs with a similar height are grouped to locate the license plate. Because the regions have been sorted by their x coordinate values, two regions can be grouped if they have similar height and similar y coordinate values and are close to each other in the horizontal direction.

Algorithm 1 Remove overlapped regions

```
1: Sort all the regions by their x coordinate values;
2: for i := 1; i ≤ regions.size; i++ do
3:   for j := i + 1; j ≤ regions.size; j++ do
4:     if r_i ⊂ r_j ∧ r_i.area > 0.9*r_j.area ∨ r_j ⊂ r_i ∧ r_j.area > 0.9*r_i.area then
5:       Remove the smaller region;
6:     end if
7:   end for
8: end for
9: for i := 1; i ≤ regions.size; i++ do
10:   for j := i + 1; j ≤ regions.size; j++ do
11:     if r_i ⊂ r_j ∨r_j ⊂ r_i then
12:       Extend r_i horizontally, count the number of blobs which have similar height with r_i in
           the enlarged region, and record it as C1;
13:       Extend r_j horizontally, count the number of blobs which have similar height with r_j in
           the enlarged region, and record it as C2;
14:       if C1 > C2 then
15:         Remove r_j ;
16:       end if
17:       if C1 < C2 then
18:         Remove r_i ;
19:       end if
20:       if C1 = C2 then
21:         Remove the region whose neighbor regions have a larger variance in height;
22:       end if
23:     end if
24:   end for
25: end for
```

2.4 VE-Based Detection Process

The CC-based method is used first since it generates fewer false license plates than the VE-based detection process. If there is no detection result, the VE-based detection process is used. The simplest way to connect the discrete edge points is using the mathematical morphological operations. However, the operator size has to be carefully selected. A large size may connect the license plate with the edges around it. A small size will make the license plate broken. And once a value is selected, the algorithm can

only detect license plates whose size is in a specific range. In our work, since the window size is computed at first, it can be used to guide the connection of the edges. The connection process is shown in Algorithm 2. The parameter λ is set as 1.5 in our work.

The detection process shown in Algorithm 2 and the CC-based method use strict restrictions to detect the license plate. If characters are hard to detect separately and the license plate touches some textures around it because they have similar window size, then it will be left out. Therefore, after the two detection processes, another method is used for supplementary. The method is based on a dynamic window technique. For a point on the character, its window size is related to the size of the character. So the window size at each edge point is used to estimate the size of the license plate. Then the license plate region computed at each pixel is evaluated by its edge density and the average window size. The process is shown in Algorithm 3. α is set as the ratio of the standard license plate width and the character width, and it is 10 in this paper. β is the standard aspect ratio of the license plate. *thr* is set as the 60% of the maximal edge density.

Algorithm 2 Detection process by connecting discrete edges

1: Generate the gradient image by using the Sobel operator on the shadow suppressed image G;
2: Perform Otsu's method on the gradient image to get the vertical edge map E;
3: **for** $i := 0$; $i < E.rows$; i++ **do**
4: **for** $j := 1$; $j < E(i).size$; j++ **do**
5: $w_1 := W(E(i, j).x, E(i, j).y)$, $w_2 := W(E(i, j\text{-}1).x, E(i, j\text{-}1).y)$;
6: **if** $|w_1 - w_2| < \lambda \min(w_1, w_2) \wedge w_1/2 + w_2/2 > E(i, j).x - E(i, j\text{-}1).x$ **then**
7: Connect $E(i, j)$ and $E(i, j\text{-}1)$;
8: **else**
9: Disconnect $E(i, j)$ and $E(i, j\text{-}1)$;
10: **end if**
11: **end for**
12: **end for**
13: Find all the regions which have a similar aspect ratio with the license plate in E;
14: Select the region with the highest edge density as the license plate;

Algorithm 3 Detection process by dynamic windows

1: Get the initial vertical edge map E as it is stated in Algorithm 2;
2: **for** each edge point in E **do**
3: $w := W(E(i, j).x, E(i, j).y)$;
4: $x := E(i, j).x$;
5: $y := E(i, j).y$;
6: $width := \alpha \times w$;
7: $height := width/\beta$;
8: Compute the average window size avg_w of the edge points in the extended window (x, y, $width$, $height$)
9: **if** $edge\_density > thr \wedge |avg\_w - w| < \min(avg\_w, w)/2$ **then**
10: Reserve the extended window (x, y, $width$, $height$);
11: **end if**
12: **end for**
13: Merge the overlapped windows;
14: Select the regions whose aspect ratio is similar with the standard aspect ratio as the license plates;

3 Experiment and Comparison

3.1 Experiment

Image database: The proposed system is tested on four image sets. Each of these images contains one or more American license plates. Most of the license plates are partly covered by shadow. More information is listed in Table 1.

Table 1. The information of the image database.

Data set	HD1	HD2	SD	LD
Size of images	1920*1080	1280*720	720*480–800*600	352*240
Size of license plates	48*14–461*127	101*27–372*110	62*18–367*99	29*11–109*56
Number of images	269	170	329	191
Number of license plates	328	170	329	191

The proposed system is implemented in C++ and run on a computer with a 4.0 GB memory and a 2.8 GHz processor. In addition, to accelerate the computation of the window size, GPU is used for its powerful parallel computing ability. Table 2 shows the experimental results of the proposed method. Note that the images in set "LD" are re-sized to 704*480, in order to use the same detection process as the other sets.

Table 2. Experimental results.

Data set	HD1	HD2	SD	LD
Location rate	91.46%	97.65%	94.22%	96.86%
Time (ms)	685.554	271.712	104.595	105.660
Total location rate	94.40%			
Total false positive rate	33.31%			

3.2 Comparison

Niblack's method [10], Sauvola's method [15], the NICK method [6] and MSER are used to find CCs on the original gray image in order to make a comparison with the proposed CC-based algorithm. To get the best performance, the parameters of the three local thresholding methods are searched in an enumerated way. The value of w increases from 3 to 51 with a step of 2, and k increases from 0.01 to 3 with a step of 0.02. Then select the values when the total location rate is the highest.

The comparison is shown in Table 3. LR denotes the location rate, FPR denotes the false positive rate, and F-measure = 2*(1-FPR)*LR/(1-FPR + LR). The highest location rate and the largest f-measure value of each data set are in bold. It shows that Niblack's method obtains the highest location rate among the three local thresholding

Table 3. Compare other CC-based methods with the proposed CC-based method.

Method		The proposed	[10] $w = 11$, $k = 0.55$	[15] $w = 7$, $k = 0.11$	[6] $w = 9$, $k = 0.13$	MSER
HD1	LR	82.62%	**83.23%**	62.80%	72.26%	68.90%
	FPR	28.87%	72.48%	60.15%	51.83%	40.21%
	F-measure	**0.76**	0.41	0.49	0.58	0.64
HD2	LR	**92.94%**	86.47%	78.82%	79.41%	65.88%
	FPR	23.67%	58.36%	17.28%	15.09%	20.00%
	F-measure	**0.84**	0.56	0.81	0.82	0.72
SD	LR	87.54%	**89.97%**	75.38%	82.37%	57.14%
	FPR	17.48%	36.75%	19.74%	12.86%	21.99%
	F-measure	**0.85**	0.74	0.78	0.85	0.66
LD	LR	**91.62%**	87.96%	88.48%	81.15%	64.40%
	FPR	22.91%	38.01%	14.21%	18.42%	17.45%
	F-measure	0.84	0.73	**0.87**	0.81	0.72
Total LR		**87.62%**	86.84%	74.36%	78.39%	63.75%
Total FPR		23.37%	57.58%	36.12%	30.73%	28.52%
Total F-measure		**0.82**	0.57	0.74	0.74	0.67

methods. However, its f-measure value is less than the other two methods because it locates more false plates. MSER is a special global thresholding method, and as we have stated, the global thresholding methods are inferior to the local thresholding methods when considering uneven illumination conditions. Overall, the performance of the proposed method is better than the other methods since it locates more license plates and fewer false plates. At the same time, the proposed method does not need a manual process to select the optimal parameters. Figure 4 shows a comparison of the binary images created by Niblack's method, MSER and the proposed method. The binary images confirm the results listed in Table 3.

To confirm the performance of the proposed system, a ML-based method is used to make a comparison. It uses a cascade classifier with the Haar-like features to detect the license plate. The cascade classifier is used to detect all kinds of objects and obtains good results. And the Haar-like features are the most common features that are used to represent the objects. The ML-based method has two important parameters: the scale factor *sf* which specifies how much the image size is reduced at each image scale and a parameter *kN* specifies how many neighbors each candidate rectangle should have to retain it. The first parameter is related to the size and position of the license plate. The second parameter controls the false positive rate. To make the comparison fair, we range *sf* from 1.01 to 1.5 and *kN* from 1 to 8, and record the results on all the data sets. Then the results are sorted by the total localization accuracy in a descending order. Figure 5 shows the top 34 results of the ML-based method. Since the accuracy increases accompanying with the raise of the false positive rate, it can be seen from Fig. 5 that when the accuracy of the ML-based method is higher than that of the proposed system, almost all the f-measure values are lower than that of the proposed

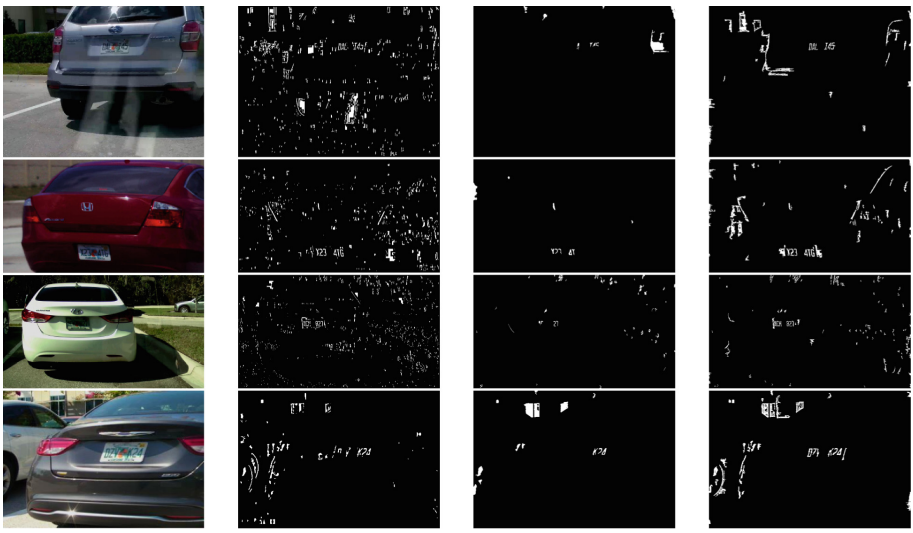


Fig. 4. Binary images created by three different algorithms. Images in (a) are the original images; Images in (b) are the binary images generated by Niblack's method; Images in (c) are the binary images generated by MSER; Images in (d) are the binary image generated by the proposed algorithm.

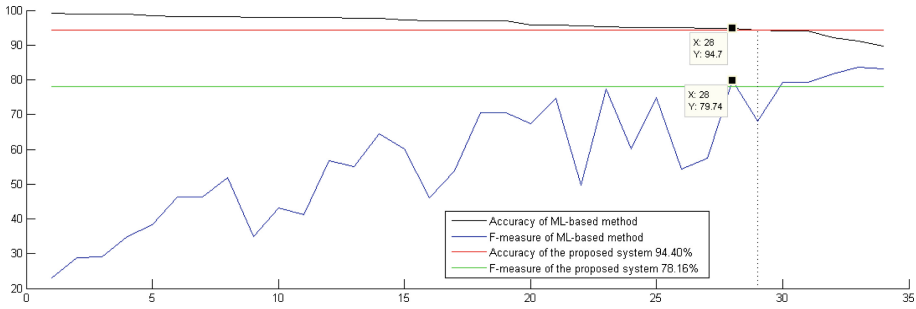


Fig. 5. Localization accuracy and f-measure of the top 34 results obtained by the ML-based method, and the accuracy and f-measure of the proposed system.

system and vice versa. The 28th result (sf = 1.02, kN = 8) is the only one that has both higher location rate and higher f-measure value than the proposed system. It means that the proposed system finds a balance between the location rate and the false positive rate. At the same time, it should be noted that the proposed system does not need to select optimal parameter manually. It is more stable than the ML-based method.

4 Conclusions

Shadow generated by uneven illumination can destroy the integrity of objects, thus making the localization of such objects a thorny issue. At the same time, most of the existing localization algorithms have a size-related parameter. Therefore, when the license plates are partly covered by shadow and their size varies a lot, selecting an optimal parameter becomes a tedious work. A new technique is proposed in this paper to compute the window size surface automatically. The proposed localization system building on this technique then avoids setting the size-related parameter by trial and error and becomes more flexible. The experiment shows that the proposed system can compete with the state-of-the-art method.

The main disadvantage of the proposed system is that the time efficiency decreases a lot when the size of the image increases. This problem will be considered in our future research.

Acknowledgments. This research work was supported by China Natural Science Foundation (No: 61272304) and Zhejiang Provincial Natural Science Foundation of China (No. LY15F020024).

References

1. Anagnostopoulos, C.N.E., Anagnostopoulos, I.E., Loumos, V., Kayafas, E.: A license plate-recognition algorithm for intelligent transportation system applications. IEEE Trans. Intell. Transp. Syst. **7**(3), 377–392 (2006)
2. Bernsen, J.: Dynamic thresholding of grey-level images. In: Proceedings of Eighth International Conference on Pattern Recognition, Paris, pp. 1251–1255 (1986)
3. Chen, Z., Chang, F., Liu, C.: Chinese license plate recognition based on human vision attention mechanism. Int. J. Pattern Recogn. Artif. Intell. **27**(08), 1350024 (2013)
4. Dehshibi, M.M., Allahverdi, R.: Persian vehicle license plate recognition using multiclass adaboost. Int. J. Comput. Electr. Eng. **4**(3), 355–358 (2012)
5. Jiao, J., Ye, Q., Huang, Q.: A configurable method for multi-style license plate recognition. Pattern Recogn. **42**(3), 358–369 (2009)
6. Khurshid, K., Faure, C.: Comparison of Niblack inspired binarization methods for ancient documents. In: Document Recognition and Retrieval XVI, DRR, Document Recognition and Retrieval Conference, pp. 1–10 (2009)
7. Kittler, J., Illingworth, J.: Minimum error thresholding. Pattern Recogn. **19**(1), 41–47 (1986)
8. Lalimi, M.A., Ghofrani, S., Mclernon, D.: A vehicle license plate detection method using region and edge based methods. Comput. Electr. Eng. **39**(3), 834–845 (2013)
9. Li, B., Tian, B., Li, Y., Wen, D.: Component-based license plate detection using conditional random field model. IEEE Trans. Intell. Transp. Syst. **14**(4), 1690–1699 (2013)
10. Niblack, W.: An introduction to digital image processing. Master's thesis, Strandberg Publishing Company (1985)
11. Nistér, D., Stewénius, H.: Linear time maximally stable extremal regions. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 183–196. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88688-4_14

12. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
13. Peng, Y., Xu, M., Jin, J.S., Luo, S., Zhao, G.: Cascade-based license plate localization with line segment features and haar-like features. In: 2011 Sixth International Conference on Image and Graphics (ICIG), pp. 1023–1028 (2011)
14. Rasooli, M., Ghofrani, S., Fatemizadeh, E.: Farsi license plate detection based on element analysis and characters recognition. Int. J. Sig. Process. Image Process. Pattern Recogn. **4**(4), 697–700 (2011)
15. Sauvola, J., Pietikainen, M.: Adaptive document image binarization. Pattern Recogn. **33**(2), 225–236 (2000)
16. Wen, Y., Lu, Y., Yan, J., Zhou, Z., Deneen, K.M., Shi, P.: An algorithm for license plate recognition applied to intelligent transportation system. IEEE Trans. Intell. Transp. Syst. **12**(3), 830–845 (2011)
17. Zheng, L., He, X., Samali, B., Yang, L.T.: An algorithm for accuracy enhancement of license plate recognition. J. Comput. Syst. Sci. **79**(2), 245–255 (2013)

Energy Prediction Model Based on Kernel Partial Least Squares for Energy Harvesting Wireless Sensor Network

Xuecai Bao(✉)

Jiangxi Province Key Laboratory of Water Information Cooperative Sensing and Intelligent Processing,
Nanchang Institute of Technology, 330099 Nanchang, China
lx97821@126.com

Abstract. In energy harvesting wireless sensor network (EH-WSN), many energy harvesting technologies are developed to sustain the long-term operation of wireless sensor network. However, the prediction of harvested energy plays an important role for energy management. In this paper, we focus on energy prediction for EH-WSN. We first analyze the factors of affecting energy harvesting and the characteristic of the solar array. Then, the kernel partial least squares (KPLS) is proposed as the energy prediction model. According to the difference of energy intake for the days, months, season and year, the four energy prediction models are established. By extensive experimental analysis for real solar data in different areas, the proposed prediction model improves prediction accuracy than existing energy prediction algorithms in EH-WSN.

Keywords: Energy prediction · Kernel partial least squares · Energy harvesting wireless sensor network

1 Introduction

Wireless sensor network (WSN) is widely used for information gathering in many sensor fields (such as environmental monitoring, automatic monitoring of agriculture, disaster management and so on. For traditional WSN, since the energy supply based on the battery is limited, most of research problems in traditional WSN are studied based on energy efficient [1]. These algorithms mainly focus on minimizing energy consumption or maximizing network utility for battery-power WSN. However, due to the limited lifetime of sensor node in battery-power WSN, nodes may fail. A failed node may result in a disconnected network which can affect the information transmission in some monitoring application domains.

Recently, many energy harvesting technologies (e.g., solar, thermal, wireless, and piezoelectric) are developed to solve the energy limitation of traditional WSN [2]. The energy harvesting wireless sensor network (EH-WSN) uses the rechargeable power supply instead of using the traditional battery. In order to achieve near-perpetual lifetime of network, the power management is the necessary. The energy prediction is one of the key technologies for effectively power management, which provide accurate

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 138–145, 2017.
DOI: 10.1007/978-981-10-3969-0_16

energy harvested in a period of future time. This prediction will allow exploiting fully the available energy, achieving the maximum network performance, and minimizing the energy waste. Therefore, many works about energy prediction are proposed. In existing researches, Kansal et al. [3] presented an energy prediction algorithm based on Exponentially Weighted Moving-Average (EWMA). The method is suitable to the diurnal cycle in solar energy but at the same time adapt to the seasonal variations. The prediction error of the EWMA is approximately equal to twenty percent. In order to improve the accuracy, Piorno et al. [4] proposed a short-term energy prediction method of solar energy harvesting, namely, weather-conditioned moving average (WCMA). In WCMA, the current and past-days weather condition is taken into account in prediction model. Compared with EWMA model, the WCWA obtain gain of more than 90% in energy utilization. But, the prediction model only predicts the energy of a day or several days. Therefore, Bergonzini et al. [5] proposed an improved WCMA predict method, WCMA-PDR. The WCMA-PDR reduces the average error to less than 9.2% at a minimum energy cost by using a phase displacement regulator (PDR). In [6], a new solar energy prediction based on additive decomposition (SEPAD) model was proposed. In this model, seasonal and daily trends along with Sun's diurnal cycle are both considered. In [7], aiming at one future slot and multiple future slots prediction model with high accuracy and low complexity, a novel weather-aware solar prediction scheme, WC-EWMA, is proposed to meet the requirements of both long-term seasonal and short-term daily solar profiles. The prediction accuracy is better than EWMA and WCMA-PDR prediction method. However, the mean prediction error still has 18.7%.

However, most of existing energy prediction algorithms only achieves the energy prediction at timescales between several hours to several days for the EH-WSN. They do not consider the prediction on future energy harvested for different time period, such as days, months, season and year.

In this paper, we study an efficient energy prediction method with high accuracy in EH-WSN. According to the difference of energy intake for the day, month, season and year, we established four types of energy prediction models to predict the harvested energy based on kernel partial least squares (KPLS) method according to the application demand. In practical deployed EH-WSN, the energy management of EH-WSN is supported by the energy prediction of different time periods, which can utilize fully the harvested energy and maximize the network performance. Therefore, our objective is to seek an effective energy prediction method to prolong the predict time and increase the predict accuracy. Our major contributions are summarized as follows

(1) We illustrated the factors of affecting energy prediction in EH-WSN and defined the input and output variables based on KPLS with multiple parameters. Furthermore, we gave the analysis for real solar data in different areas.
(2) According to different time length of energy prediction. We present different energy prediction method based on KPLS, which improve the prediction accuracy.
(3) The extensive simulation results showed the prediction error delivered by the proposed method is less than the existing energy prediction method and achieves the energy prediction for different time periods.

The rest of the paper is organized as follows. Section 2 presents related problem description of energy prediction for EH-WSN and the principle of KPLS. Section 3 presents the prediction method based on KPLS according to the different prediction interval length in EH-WSN. In Sect. 4, we evaluate the proposed method performance, followed by concluding remarks in Sect. 5.

2 Paper Preparation Problem Description and Kernel Partial Least Squares Model

2.1 Problem Description

In EH-WSN, each sensor node requires forecasting harvested power in a period of future time. Therefore, the application of energy harvesting technologies is required to deal with the variable behavior of the energy sources. By selecting the predictable, non-controllable power sources (such as the solar one) and recording the harvested energy of each time interval, the methods of energy prediction are required to forecast the source availability and estimate the expected energy intake [8]. For example, the Fig. 1 show solar power varies smoothly and fluctuates on different days. The objective of energy prediction is to forecast energy intake in future time periods by the harvested energy data of past time periods. The existing energy prediction methods mainly focus on the prediction of short term, such as several hours or several days. In this paper, we put the time periods divide into four phases, which are respectively days, mouths, season, and year. The different time periods use different prediction model.

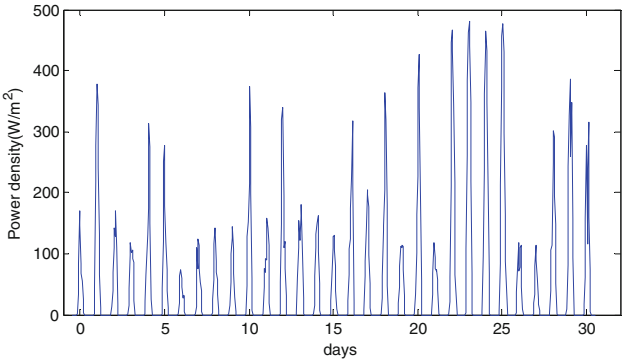


Fig. 1. Variance of solar power on different days

2.2 Kernel Partial Least Squares (KPLS)

Rosipal and Trejo [9] first integrated the kernel function into partial least squares regression model, which have extended the linear PLS model into its nonlinear kernel form. This model overcomes the limitation that the partial least-squares regression can only deal with the linear relationship. Moreover, it can make fully use of the

distribution information of sample space, establish a relationship model between response variables and explanatory variables, and greatly improve the fitting precision and prediction precision. The specific principle of KPLS is as follows.

Assume having a mapping $\Phi : x_i \in \Re^N \rightarrow \Phi(x_i) \in F$—a nonlinear transformation of mapped the input variable $\mathbf{X}$ in feature space $\boldsymbol{F}$. The goal is to establish a linear model in feature space $\boldsymbol{F}$. The Φ denotes an $(n \times S)$ matrix obtained by $\Phi(x_i)$. To overcome the limitation of processing the linear problem for PLS, the dimensional of feature space depends on the nonlinear transformation $\Phi(.)$. The specific calculation steps are as follows [9].

Setp 1: Randomly initialize $\boldsymbol{u}$.
Step 2: $\boldsymbol{t} = \boldsymbol{Ku}$, $\boldsymbol{t} = \boldsymbol{t}/\|\boldsymbol{t}\|$, $\boldsymbol{K}_{ij} = \boldsymbol{K}(x_i, x_j)$ where $\boldsymbol{K}$ represents the $(n \times n)$ kernel Gram matrix of the cross dot products between all input data points $\{\Phi(x_i)\}_{i=1}^{n}$, that is, $\boldsymbol{K} = \Phi\Phi^T$, $K_{ij} = K(x_i, x_j)$ where $K(.,.)$ is a selected kernel function.
Step 3: $\boldsymbol{c} = \boldsymbol{Y}^T\boldsymbol{t}$.
Step 4: $\boldsymbol{u} = \boldsymbol{Yc}$, $\boldsymbol{u} = \boldsymbol{u}/\|\boldsymbol{u}\|$
Step 5: Repeat step 2 to 5 until convergence
Step 6: Calculating the residual matrix of $\boldsymbol{K}$ and $\boldsymbol{Y}$. $\boldsymbol{K} = (\boldsymbol{I} - \boldsymbol{tt}^{\mathbf{T}})\ \boldsymbol{K}(\boldsymbol{I} - \boldsymbol{tt}^{\mathbf{T}})$, $\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{tt}^{\mathbf{T}}\boldsymbol{Y}$.
Step 7: Repeat Step 2–Step 7, until reach the required number of principle component.
Step 8: Calculating the regression coefficient,

$$\boldsymbol{a} = \Phi^T U(T^T KU)^{-1} T^T Y$$

Step 9: The prediction formula is

$$\hat{Y} = \Phi_t a = K_t U(T^T KU)^{-1} T^T Y$$

where the K is obtained by Step 2 and $K_t = \{K_t - \frac{1}{n} I_t I_n^T K\} K \{\mathrm{I} - \frac{1}{n} I_t I_n^T\}$

3 Establishing Energy Prediction Model Based on KPLS

In this section, we present our energy prediction model to predict the harvested energy of senor node in a period of future time. Our proposed models include days, months, season, and year prediction models, respectively. According to the model of KPLS in Sect. 2, each energy prediction model is established by past harvested energy data. First, we define the input variables and output variables of energy prediction model. For the input variables, we select the solar radiation, average wind speed, ambient temperature as input variable. The total solar radiation includes T timeslots in prediction time period. The output variables refer to the output power of each timeslots in total prediction time period. In procedure of energy prediction, the prediction includes two aspects. The first aspect mainly considers the demand of energy prediction for different time period, such as days, months, season, and year. The second aspect reflects the energy prediction according to the prediction demand. Our objective is

suitable for the energy prediction from short term to long term. The description of the prediction algorithm based on KPLS model is as follows.

Algorithm 1: Energy prediction algorithm

Input: The specific prediction interval length L, input variables $\boldsymbol{X}$.

Output: Output variable $\hat{Y}$.

1. Input the value of L.
2. Establishing Φ_{it} and $\boldsymbol{K}$ based on the training of KPLS model and the value of L
3. Calculating regression coefficient based on KPLS model
   $\boldsymbol{a_i} = \Phi_{it}{}^T U(T^T KU)^{-1} T^T Y$, where $1 \le i \le 5$
4. Select corresponding prediction model according to prediction interval length L.
5. Energy_Pred(L, a_i, Φ_{it})
6. Procedure Energy_Pred (L, a_i, Φ_{it})
7. If L='days'
8. Then $\hat{Y} = \Phi_{2t} a_2$
9. If L='month'
   Then $\hat{Y} = \Phi_{3t} a_3$
10. If L='season'
    Then $\hat{Y} = \Phi_{4t} a_4$
11. If L='year'
    Then $\hat{Y} = \Phi_{5t} a_5$

For above energy prediction algorithm, we first require to input prediction interval length L. Then according to the value of T, the corresponding variables are calculated, such as Φ_{it} and K(line 2). In line 3, the regression coefficient of KPLS a_{i} is calculated, where i denotes the type of prediction interval length. For example, when i = 1, the a_{i} is regression coefficient of predicting the energy value in several days. In line 4 and line 5, the procedure of energy prediction is called to calculate the final value of energy prediction. The specific procedure of energy prediction is described in line 6–11.

4 Performance Evaluation

In this section, we evaluate the performance of the proposed energy prediction algorithm based on KPLS model. All simulations were based on the public solar database [10]. Solar cells are technically named photovoltaic cells because they change light (or in Greek, photo) into electricity. PV modules come in sizes from 10 W to 300 W. We utilize the model of solar cells in [10]. The specific description is as follow: Systems that are oriented generally south and on a 4/12 pitched roof (18.5 degree tilt) or steeper produce at least 95% of the electricity of an optimally oriented system. Solar radiation is approximately 1,000 W/m^2. Mono-crystalline (single crystal) and multi-crystalline solar panels change about 15–18% of the incident sunlight into electricity. The inverter is usually about 90% efficient in turning DC current to AC current. Other factors such

as line losses and dirt on the array typically cause another 10% decrease in performance [10]. Photovoltaic (PV) cells use the same technology that is used to make computer chips and other solid state electronic components. We compared the performance of the proposed algorithm with the classic scheme EWMA [3] and WCMA [4], based on real solar data [10]. For all experiments, the L denotes prediction interval length and each prediction interval length divide into 24 time slots. Here, we give the comparison results of energy prediction based on prediction interval length L = 'days' and L = 'months'. The specific results are shown in Figs. 2 and 3.

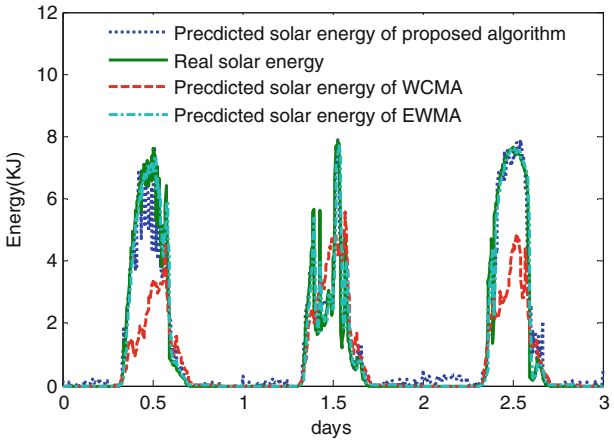


Fig. 2. Comparison of Energy prediction for prediction interval length L = 'days'

From Fig. 2 as we know, the proposed prediction algorithm presents better prediction precision than WCMA and EWMA. The prediction error of EWMA is less than that of WCMA, which is different from the result in literature [4]. By the further analysis, we found that the ambient temperature play an important role in the precision of the energy prediction. In [4], the prediction model doesn't consider these factors of average wind speed, ambient temperature. For Fig. 3, the performance of the proposed algorithm similarly is better than others. But the prediction errors of all energy prediction algorithms are larger than the result in Fig. 2.

In order to further evaluate the performance of the proposed algorithm, the prediction errors for the proposed energy prediction algorithm, EWMA and WCMA for different prediction interval length L are shown Table 1.

From Table 1, the prediction errors of the proposed prediction algorithm are less than other algorithms. Moreover, the prediction errors of all the energy prediction algorithms increase with the increasing of prediction interval length. The results indicated the proposed algorithm present a valid energy prediction.

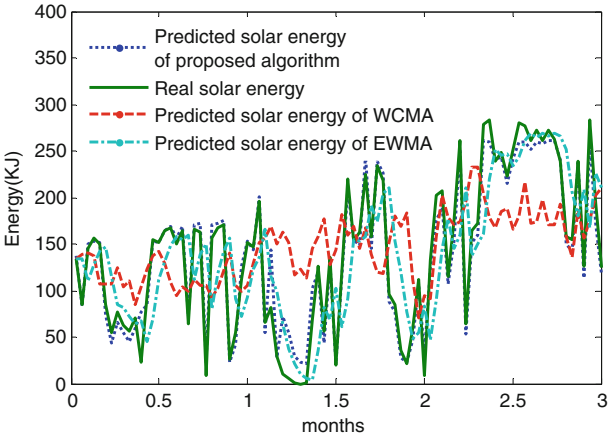


Fig. 3. Comparison of Energy prediction for prediction interval length L = 'months'

Table 1. Comparison of prediction errors

	Days	Months	Season	Year
The proposed algorithm	7.3%	8.2%	13.6%	18.3%
EWMA	10.6%	15.9%	18.1%	25.4%
WCMA	30.2%	40.5%	46.4%	50.2%

5 Conclusion

In this paper, we proposed an energy prediction algorithm based on KPLS for energy harvesting WSN. The energy prediction interval length of the existing energy prediction algorithms mainly focus on the hours and days. Our proposed energy prediction algorithms achieve different prediction interval length, including the days, months, season and year. We first analyze the factors of affecting energy intake in EH-WSN and present the input and output variables. Based on the analysis of KPLS model, the energy prediction algorithm is proposed. The simulations based on real solar data show that our proposed algorithm has less prediction error than existing typical prediction algorithms in different energy prediction interval length.

Acknowledgement. This research is supported by the National Natural Science Foundation of China (Grant No. 61401189), Natural Science Foundation of Jiangxi, China (Grant No. 20161BAB212036), and Natural Science Fund of Nanchang Institute of technology (Grant No. 2014KJ016).

References

1. Anastasi, G., Conti, M., Francesco, M.D., Passarella, A.: Energy conservation in wireless sensor networks: a survey. Ad Hoc Netw. **7**, 537–568 (2009)

2. Ren, X., Liang, W., Xu, W.: Data collection maximization in renewable sensor networks via time-slot scheduling. IEEE Trans. Comput. **64**(7), 1870–1883 (2015)
3. Kansal, A., Hsu, J., Zahedi, S., Srivastava, M.B.: Power management in energy harvesting sensor networks. ACM Trans. Embed. Comput. Syst. (TECS 2007) **6**(4) (2007)
4. Piorno, J., Bergonzini, C., Atienza, D., Rosing, T.: Prediction and management in energy harvested wireless sensor nodes. In: Proceedings of Wireless VITAE 2009, Aalborg, Denmark, pp. 6–10 (2009)
5. Bergonzini, C., Brunelli, D., Benini, L.: Comparison of Energy intake prediction algorithms for systems powered by photovoltaic harvesters. Microelectron. J. **41**(11), 766–777 (2010)
6. Hassan, M., Bermak, A.: Solar harvested energy prediction algorithm for wireless sensors. Qual. Electron. Des. **48**(1), 178–181 (2012)
7. Yang, S., Yang, X., Mccann, J.A., Zhang, T.: Distributed networking in autonomic solar powered wireless sensor networks. IEEE J. Sel. Areas Commun. **31**(12), 750–761 (2013)
8. Cammarano, A., Petrioli, C., Spenza, D.: Pro-energy: a novel energy prediction model for solar and wind energy-harvesting wireless sensor networks. In: IEEE International Conference on Mobile Ad-Hoc and Sensor Systems, pp. 75–83 (2012)
9. Rosipal, R., Trejo, L.J.: Kernel partial least squares regression reproducing kernel Hilbert space. J. Mach. Learn. Res. **2**, 97–123 (2002)
10. http://solardat.uoregon.edu/SelectArchival.html

Deep Convolution Neural Network Recognition Algorithm Based on Maximum Scatter Difference Criterion

Kunlun Li[1(✉)], Xuefei Geng[1], and Weiduan Li[2]

[1] College of Electronic and Information Engineering, Hebei University, Baoding 071000, China
likunlun@hbu.edu.cn
[2] College of Civil Engineering and Architecture, Hebei University, Baoding 071000, China

Abstract. Convolution neural network is a method that can extract features automatically of deep learning. It has a better recognition effect compared with a variety of face recognition algorithms. In view of the problem that the number of face training samples is reduced and the recognition performance is reduced too, the recognition algorithm based on maximum scatter difference criterion is proposed. The maximum scatter difference criterion is introduced to minimize the error when the gradient descent method is used to adjust the weight. And the within-class scatter of the sample should be the minimum and the between-class should be the maximum. Finally, the weights can be more close to the optimal value of the classification and the recognition rate of the system can be improved. A large number of experiments show the effectiveness of the algorithm.

Keywords: Deep learning · Convolution network · Maximum scatter difference criterion · Face recognition

1 Introduction

In recent years, face recognition is one of the hot topics in the field of pattern recognition, image processing, computer vision and so on [1, 2]. And the method of deep learning has gained remarkable success in machine learning tasks such as face recognition [3, 4]. Deep learning network has a strong ability to function, and it has a good effect on complex function representation and classification. Under the premise of inheritance automatic learning feature extraction, convolution neural network as one of the deep learning approaches ensures the spatial structure of the original signal and shares weights to reduce the need to train the parameters, so in many fields have good effect.

LeCun firstly applied the convolution neural network to the field of handwritten character recognition and proposed a gradient descent method. When adjusting the weights, the BP algorithm which is based on the minimum error method should carry on the back propagation [5]. It is worth noting that the model is also used in other feature extraction. However, the existing convolution neural network accuracy in

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 146–153, 2017.
DOI: 10.1007/978-981-10-3969-0_17

dealing with the problem of face recognition is still to be improved, especially when faced with the lack of training samples or the attitude changes of face images. Its convergence speed is slow and the recognition accuracy is not high enough. According to the above problems, deep convolution neural network recognition algorithm based on the maximum scatter difference criterion is proposed in this paper. The ability of classify of the convolution neural network should be improved when the training samples are small and the orientation of the face samples are changeable. This paper introduces the maximum scatter difference criterion, taking into account the error is minimized at the same time to keep the sample within-class scatter minimum and between-class scatter maximum, the weight can be more conducive to the classification of the approximation of the optimal value, the recognition rate of the system can be improved. The experimental results show that a better effect can be achieved when the training sample is insufficient or the human face has the attitude change.

2 Depth Convolution Neural Network

2.1 Convolution Neural Network Architecture

Convolution neural network (CNN) has developed on the basis of traditional neural network and has aroused wide attention on highly efficient recognition [6]. Because the original image can be directly entered into the network, which avoids the pre-processing of the image, hence, it has been widely used.

The core idea of the convolution neural network is that local receptive fields, weight sharing and down sampling were used to optimize the neural network structure and reduce the network of neurons in both the number and weight. At the same time, the pool technology was used to keep the feature having the displacement, scaling, and distortion invariable [7].

Convolution neural network includes forward propagation, back propagation, convolution layer and the down sampling. The processes of convolution kernel down sampling are that a training of the convolution kernel and different combinations of feature map are convoluted, then added bias, and finally got the current layer of the feature map. The sub sampling layer can greatly reduce the dimension, and it has a advantage of translation invariable. The process is shown in the following manner.

$$x_j^l = f(\beta_j^l down(x_j^{l-1}) + b_j^l) \tag{1}$$

Where: $down\left(x_j^{l-1}\right)$ is the $l-1$ layer of the jth feature map sub sampling, β_i^j is multiplicative bias, b_j^i is the additive bias, f (x) is the activation function, x_j^{l-1} is the l layer of the jth characteristic diagram. Under the combination of more than one of the above convolution sampling process constitutes the deep convolution neural network. The typical deep learning neural network structure schematic diagram is shown in Fig. 1.

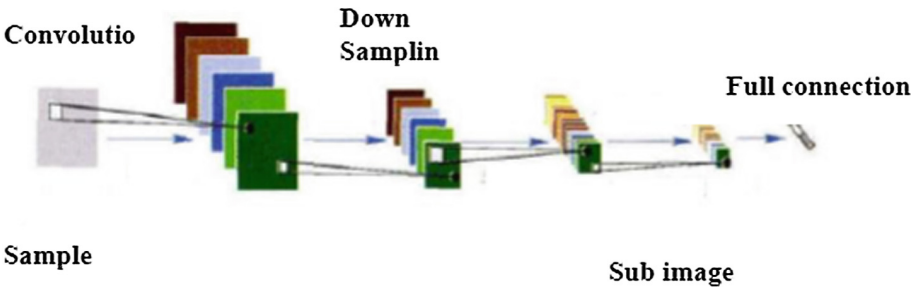


Fig. 1. Sketch map of deep convolution neural network

2.2 Convolution Neural Network Cost Function and BP Algorithm

A sample set consisting of m samples is $\{(x^l), \ldots, (y^l)\}$, they belong to the n categories, $y^{(i)}$ is the corresponding category label of $x^{(i)}$. The basic cost function of the convolution neural network is

$$\begin{aligned} J(W,b) &= R(W,b) \\ &= [\frac{1}{m}\sum_{i=1}^{m} J(W,b;x^{(i)},y^{(i)})] \\ &= [\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left\|h_{W,b}(x^{(i)}) - y^{(i)}\right\|^2] \end{aligned} \tag{2}$$

Formula: W is the weight for each layer of the connection among the parameters; b is the bias term; $h_{W,b}(x^i)$ is predictive value for the final layer of the neural network output. The goal of the training network is to find the minimum value of the function J (W, b) for the parameters W and b, and the objective function is optimized by using the gradient descent method.

$$W_{ij}^{(l)} = W_{ij}^{(l)} - \alpha\frac{\partial}{\partial W_{ij}^{(l)}}J(W,b) \tag{3}$$

$$b_{ij}^{(l)} = b_{ij}^{(l)} - \alpha\frac{\partial}{\partial b_{ij}^{(l)}}J(W,b) \tag{4}$$

α is a learning rate, type (3), (4) can be solved by BP algorithm. When using BP algorithm, we should first spread to calculate the network's final output value of $h_{W,b}(x^i)$. Then the gap is calculated between the output value and actual value, which is δ_i^{nl}. Finally, through the compute residuals for each layer calculate the partial derivative of the (3) and (4).

The formula for calculating the residual error of the last layer of the traditional neural network is

$$\delta_i^{(nl)} = \frac{\partial J_1}{\partial Z_i^{(nl)}} = \frac{\partial}{\partial Z_i^{(nl)}} \frac{1}{2} \left\| h_{W,b}(x^{(i)}) - y^{(i)} \right\|^2 \tag{5}$$

3 Depth Convolution Recognition Algorithm Based on Maximum Scatter Difference Criterion

Maximum scatter difference criterion is put forward based on Fisher criterion, and it is a kind of supervised classification criteria. Its basic purpose is searching for an optimal projection direction to implement the smallest of the within-class scatter and the largest of the between-class scatter. Its classification effect is better [8]. In order to make the algorithm more beneficial to the classification, the energy function based on the scatter of the between-class and the within-class is proposed, which is based on the idea of the maximum scatter difference criterion. S_w is similarity measure function within class which is defined as the sum of scatter between all samples and their class mean values. S_b is similarity measure function between class which is defined as the sum of scatter between the average of all the samples.

$$S_w = \sum_{i=1}^{m} \sum_{j=1}^{n} \left\| h_{w,b}(x^{(i,j)}) - M^{(i)} \right\|^2 \tag{6}$$

$$S_b = \frac{1}{2m} \sum_{i=1}^{m} \sum_{j=1}^{m} \left\| M^{(i)} - M^{(j)} \right\|^2 \tag{7}$$

The mean value of the class i sample is expressed in the formula $M^{(i)}$, i.e.

$$M^{(i)} = \frac{\sum_{j=1}^{n} h_{w,b}(x^{(i,j)})}{n} \tag{8}$$

When using S_w as the cost function to compute the gradient, each iteration step makes the sample predictive value to average predictive value of the categories of samples smaller. When using S_b as the cost function to compute the gradient, each step of the iteration makes the scatter between different classes larger.

In order to make the characteristics of each layer of the deep learning network more favorable to the classification, the energy function model of constraints between class and within class is proposed. It is $J = R + \gamma S_w + \eta S_b$

R is the cost function of neural network. The total cost function is J. Considered the error makes the scatter of within class is the smallest and between classes is the largest. When the weights are adjusted, the weights of each layer can be adjusted to the direction of the classification, so that the sample quantity is small or the face direction is changed, the target of the classification can be more quickly close.

It is most important that using BP algorithm to update the weights is to find the target function of the final output layer of each output unit of the residual error. S_w is a function of measure of similarity within class, and the formula for calculating the residual error of each output unit in the output layer is:

$$\begin{aligned}
\delta_i^{nl} &= \frac{\partial S_w}{\partial Z_i^{(nl)}} = \frac{\partial}{\partial Z_i^{(nl)}} \sum_{i=1}^{m} \sum_{j=1}^{n} \left\| h_{w,b}(x^{(i,j)}) - M^{(i)} \right\|^2 \\
&= 2 \sum_{i=1}^{m} \sum_{j=1}^{n} (h_{w,b}(x^{(i,j)}) - M^{(i)}) \cdot ((\alpha_i^{(nl)})' - (M^{(i)})') \\
&= 2 \sum_{i=1}^{m} \sum_{j=1}^{n} (\alpha_i^{(nl)} - M^{(i)}) \cdot (\alpha_i^{(nl)} \cdot (1 - \alpha_i^{(nl)})) - (nM^{(i)} \cdot (\frac{1}{n} - M^{(i)})) \\
&= 2 \sum_{i=1}^{m} \sum_{j=1}^{n} (\alpha_i^{(nl)} - M^{(i)}) \cdot (\alpha_i^{(nl)} \cdot (1 - \alpha_i^{(nl)})) - \frac{\alpha_i^{(nl)} \cdot (1 - \alpha_i^{(nl)})}{n} \\
&= 2 \sum_{i=1}^{m} \sum_{j=1}^{n} (\alpha_i^{(nl)} - M^{(i)}) \cdot (\alpha_i^{(nl)} \cdot (1 - \alpha_i^{(nl)}))(1 - \frac{1}{n})
\end{aligned} \tag{9}$$

S_b is a function of measure of similarity between classes, and the formula for calculating the residual error of each output unit in the output layer is:

$$\begin{aligned}
\delta_i^{nl} &= \frac{\partial S_b}{\partial Z_i^{(nl)}} = \frac{\partial}{\partial Z_i^{(nl)}} \frac{1}{2m} \sum_{i=1}^{m} \sum_{j=1}^{m} \left\| M^{(i)} - M^{(j)} \right\|^2 \\
&= \frac{1}{m} (M^{(i)} - M^{(j)}) \cdot [(M^{(i)})' - (M^{(j)})'] \\
&= \frac{1}{m} (M^{(i)} - M^{(j)}) \cdot [nM^{(i)} \cdot (\frac{1}{n} - M^{(i)}) - nM^{(j)} \cdot (\frac{1}{n} - M^{(j)})]
\end{aligned} \tag{10}$$

In the model, each sub function can be obtained after the last layer of residual error can be iterated through the BP algorithm and get all the weights.

4 Experimental Results and Analysis

4.1 Experiment Results on ORL Face Database

ORL database is created by the AT&T Laboratory of University of Cambridge. It contains 400 face images of 40 individuals. For each individual, 10 pictures are taken various attitude changes and expression changes. Each image resized to 92 * 112 for PGM format, as shown in Fig. 2.

Because of the small number of samples of ORL face database, it is found that a lot of iterative training needs to be carried out to achieve a better recognition effect. Figure 3 shows the experimental results on NN, CNN, and MCNN in the ORL database.

Fig. 2. Part of the face images in ORL face database

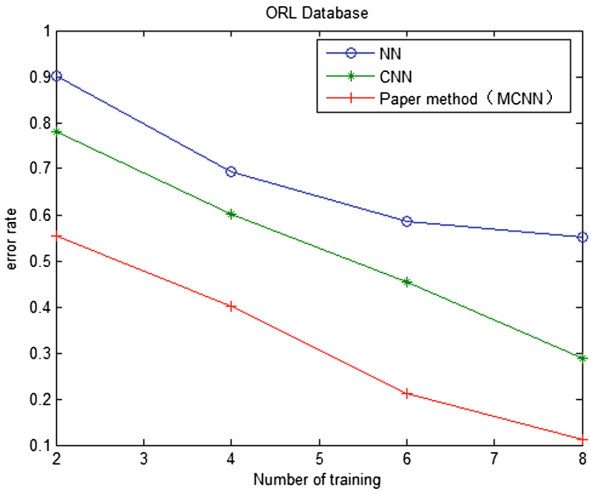


Fig. 3. The error rate of different methods of training samples in ORL database

From the graph, we can see that when the training samples are less, the overall recognition rate of the depth learning algorithm is not very good. However, with the increase of the number of training samples, the recognition effect has been increased to a certain extent. Even when the number of training samples is small, the error rate of MCNN algorithm is still lower than that of the other two methods. And when the number of training samples is less, the advantage is more obvious. This is the MCNN algorithm which added to the class of constraints, so that the classification performance is better.

In this experiment, 9 images of each person were selected as a training sample, and 1 image was used as test samples. The experimental results were as Table 1.

Table 1. The recognition rate of the classical CNN and MCNN in the ORL database

Test times	Training frequency	Method	Error rate	Increase %
1	500	CNN	0.2250	11.37
2	500	MCNN	0.1113	
3	1700	CNN	0.2188	10.80
4	1700	MCNN	0.1108	
5	2500	CNN	0.2105	10.00
6	2500	MCNN	0.1105	

4.2 Experiment Results on IMM Face Database

IMM face database contains 7 women, 33 men and a total of 240 individuals of 40 human face color images. Each of the 6 images contain a human face pose, angle, light and expression. Each image resized to 92 * 112 for PGM format, as shown in Fig. 4.

Fig. 4. Part of the face images in IMM face database

2, 4, 6 pictures were selected from each of the 6 pictures, and the 2 images were tested in the experiment. Experimental results are shown in Table 2. From the table we can see this algorithm recognition rate is higher compared with other methods. That is to say, when the training sample is less, the method is more effective.

Table 2. Experimental results on IMM face database

Sample number experimental method	2	4	6
NN	0.9138	0.8342	0.6160
CNN	0.8503	0.7942	0.4839
MCNN	0.6228	0.5939	0.4208

5 Conclusion

In this paper, a new algorithm for the recognition of neural network based on the maximum scatter difference criterion is proposed. In the iterative adjustment of the weights, it is not only to consider the minimization of error, but also to make the sample to keep the smallest of scatter within the class and the largest of scatter between the classes. Thus the weight can be more quickly approaching the optimal value of the classification. The follow-up study of this paper is to optimize the structure of the neural network, to improve the recognition rate and reduce the complexity of the network.

Acknowledgments. This work is supported by the National Science and Technology Support Program (2013BAK07B00), the Natural Science Foundation of Heibei Province of China under granted (F2013201170), the Educational Commission of Hebei Province of China (ZD2014008) and the National Natural Science Foundation of China (No. 61672205).

References

1. Gui, J., Sun, Z., Jia, W., et al.: Discriminant sparse neighborhood preserving embedding for face recognition. J. Pattern Recogn. **45**, 2884–2897 (2012)
2. Geng, C., Jiang, X.D.: Fully automatic face recognition framework based on local and global features. J Mach. Vis. Appl. **19**(3), 549–571 (2013)
3. Hinton, G.E., Osindero, S.: A fast learning algorithm for deep belief nets. J. Neural Comput. **18**(7), 1527–1554 (2006)
4. Sun, Z., Xue, L., Xu, Y., et al.: Review on the research of deep learning. J. Comput. Appl. Res. **29**(8), 2806–2810 (2012)
5. Lecun, Y.L., Bottou, L., Bengio, Y., et al.: Gradient-based learning applied to document recognition. J. Proc. IEEE **86**(11), 2278–2324 (1998)
6. Tang, P., Wang, H.: The depth of the parallel cross convolution neural network model. J. Chin. J. Image Graph. (2016)
7. http://blog.csdn.net/nan355655600/article/details/17717589
8. Zheng, N., Qi, L., Guan, L.: Generalized multiple maximum scatter difference feature extraction using QR decomposition. J. Vis. Commun. Image Represent. **25**(6), 1460–1471 (2014)

Energy Efficient Routing Algorithm Using Software Defining Network for WSNs via Unequal Clustering

Hang Yu[1], Zhiping Jia[1](✉), Lei Ju[1], Chunguang Liu[2], and Xianzhong Ding[1]

[1] School of Computer Science and Technology, University of Shandong, Jinan, China
jzp@sdu.edu.cn
[2] School of Environmental Science and Engineering, Shandong University, Jinan 250100, China

Abstract. Clustering and multi-hop routing algorithms substantially prolong the lifetime of wireless sensor networks (WSNs). However, in multi-hop routing, the nodes near the sink are burdened with heavy relay traffic and tend to exhaust their energy very quickly. Thus, the energy hole arises. In this paper, we propose an SDN-based Unequal Clustering Routing (SDUCR) protocol for WSNs, which consists of a Centralized Unequal Clustering algorithm (CUCA) and a Connected Graph Based Minimum Energy Consumption (CGMEC) multi-hop routing algorithm. We use the centralized intelligence of Software Defined Network (SDN) to implement clustering and routing. Our CUCA is used to partition the network into clusters of unequal size based on residual energy and degree of sensor nodes. The CGMEC algorithm constructs a routing tree among cluster heads which ensures the connectivity among nodes and balances the communication cost of all nodes. Simulation results show that our SDUCR protocol balances the energy consumption among sensor nodes and achieves an obvious improvement on the network lifetime.

Keywords: Wireless sensor networks · Energy hole problem · Software Defined Network · Unequal Clustering · Multi-hop routing · Routing algorithm

1 Introduction

WSNs are characterized by limited power, computation ability and memory constraint. As the energy is non-rechargeable, the energy should be managed carefully. In order to achieve high energy efficiency and assure long network lifetime, sensor nodes can be organized hierarchically by clustering [1]. Previous research [2] has shown that multi-hop communication is more energy efficient than direct transmission because of the characteristics of wireless channel. However, the "energy hole" problem arises in multi-hop forwarding model because in this model nodes near to the sink are burdened with more data transmission task which make them die earlier than nodes away from the sink.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 154–163, 2017.
DOI: 10.1007/978-981-10-3969-0_18

Clustering and multi-hop routing can reduce energy consumption obviously. Therefore, many energy-aware clustering and routing algorithms have been proposed [3]. However, these algorithms run under the distributed structure which need amount of information exchange.

Software Defined Networking (SDN) is an emerging network paradigm that decouples data forwarding path and the control path. SDN was developed to facilitate innovation and enable simple programmatic control of the network datapath, which allows network administrators to manage network services through abstraction of lower level functionality. Recently, some researchers apply SDN in WSNs [4, 5]. Most of these researches design the architecture for WSNs based on SDN.

In this paper, a new protocol named SDN-based Unequal Clustering Routing protocol (SDUCR) proposed which use the centralized intelligence to implement clustering and routing. Utilizing the global view of the network, we complete the process of clustering and construct routing path in the Controller which provides the optimum solution for clustering and routing.

The remainder of this paper is organized as follows. Section 2 describes the related work. Section 3 describes the network model and the energy consumption model. Section 4 presents our CUCA clustering algorithm and CGMEC multi-hop routing algorithm in detail. Section 5 presents our experimental results with discussions. Finally, Sect. 6 offers the concluding remarks.

2 Related Work

In the past few years, many clustering protocols have been proposed to prolong the lifetime of WSNs. A hybrid energy-efficient distributed (HEED) clustering algorithm is proposed in [7], which chooses cluster head base on node's residual energy. Sensors that are not covered by any cluster heads double their probability of becoming a cluster-head. This procedure iterates until all sensors are covered by at least one cluster-head. Hence, it needs many times of iterations and incurs high overhead.

In EEUC [8], the authors propose an Energy-Efficient Unequal Clustering mechanism for periodical data gathering applications in wireless sensor networks. EEUC is a distributed competitive algorithm, where cluster heads are elected by localized competition. The node's competition range decreases as its distance to the base station decreasing. In UCRA [9], authors use vote mechanism to choose cluster head (WCA), the rest of the nodes choose the best cluster head to join according to fitness or other mechanisms. This procedure iterates until all nodes are covered by at least one cluster-head. For inter-clustering phase, UCRA uses Minimum Energy Consumption (MEC) routing algorithm which makes the energy more efficient. In CAUCR [10], authors use an Optimized Weighted Unequal-Clustering Algorithm (OWUCA) which is also vote-based algorithm to execute the clustering phase. In RME routing phase, the cluster head has known the cluster heads around itself according to the clustering phase's information. RME starts routing construction from the cluster heads close to the sink, and finishes in certain steps.

3 Network Model

We consider a square sensing field and N nodes uniformly distributed within the field. The sink node in WSNs collects data from sensor nodes periodically. We use the system model just like [4] which introduces SDN architecture into WSNs. The Controller runs on the sink and sensors transmit data packets following the rules in flow table. The Controller identifies the cluster head with unequal aggregating size using the information gathering from topology discovery phase and builds the routing tree between the cluster heads. Cluster members send their data to the cluster head. The cluster head receive and aggregate these data, then relay them to the sink via routing tree.

We make the following assumptions and the underlying network model (just like previous researches [8–10]):

1. There is a sink in the network; sensors and the sink are all stationary after deployment.
2. All nodes are homogeneous and have the same capabilities. A unique identifier (ID) is assigned to each node.
3. Nodes can use power control to vary the amount of transmission power, which depends on the distance to the receiver.
4. Each node knows its own location through GPRS or RSSI localization [11].

We use a standard energy consumption model as described in [6]. The energy consumption for transmission of a k-bit packet over distance d is:

$$E_{Tx}(k,d) = \begin{cases} ke_{tx} + k\varepsilon_{fs}d^2 & d \le d_0 \\ ke_{tx} + k\varepsilon_{mp}d^4 & d > d_0 \end{cases} \tag{1}$$

where, e_{tx} depends on factors such as the digital coding, modulation, filtering, and spreading of the signal. The distance d_0 is the threshold defined as $\sqrt{\varepsilon_{fs}/\varepsilon_{mp}}$. ε_{fs} and ε_{mp} are the power amplification factors. In addition, to receive this message, the radio energy consumption is:

$$E_{Rx}(k) = k \cdot e_{rx} \tag{2}$$

where, e_{rx} is the receiver electronics energy for each bit. The total aggregation energy is:

$$E_{ag}(m,k) = k \cdot m \cdot e_{ag} \tag{3}$$

where m is the number of cluster members and e_{ag} is the data aggregation energy per bit, k is the size of aggregation data.

4 Proposed Routing Framework

In this section, we describe our routing protocol in detail. Our protocol contains four phase: topology discovery, cluster building, Data transmission, re-clustering and rerouting.

4.1 Topology Discovery

In topology discovery phase, the topology manager module in the WISE-Visor builds a consistent view of the current network status. Therefore, it requires to collect local topology information generated by sensor nodes. The Topology Discovery (TD) layer of sensor node maintains its current neighbors with transmission range d_0. It will deliver their neighbors' information to the WISE-Visor, then the WISE-Visor constructs the network topology graph [4].

For the reason of reducing control overheads, the controller runs TD phase only when the initialization of the network and the residual energy of one of the current cluster heads is less than the energy threshold.

4.2 Unequal Clustering Protocol

We propose Centralized Unequal Clustering Algorithm (CUCA) to complete the process of clustering. The Controller holds the topology graph containing nodes geography location information and the residual energy of each node. The Controller calculates d_{max} and d_{min}, then, calculates the competition radius based on the formula [8]:

$$R_i = (1 - c \times \frac{d_{max} - d(v_i, sink)}{d_{max} - d_{min}})R_{max} \tag{4}$$

R_i is the competition radius of node i. d_{max} and d_{min} are calculated based on nodes geography location information. $d(v_i, sink)$ is the distance between node i and sink. R_{max} is the maximum competition radius. c is a constant coefficient between 0 and 1.

Topology Reconstruction. The Controller utilizes the competition radius of each node and nodes geography location information to reconstruct topology graph which means neighbor list of each node will be changed.

Cluster Head Selection. In this phase, the Controller uses CUCA to select the optimum nodes to be cluster head. The node which has more residual energy will have more chance to be cluster head. CUCA chooses the node whose energy ratio in among it's neighbors is biggest to be cluster head, then its neighbors become member nodes.

The node set "*node*" contains all the sensor nodes in the sensor field. Each node sums the energy of all its neighbors in $rNb(i)$ up as $SE(i)$.

$$\mathrm{SE(i)} = \sum_{j \in rNb(i)} E_j + E_i \tag{5}$$

where, $rNb(i)$ is the neighbor list of node i after topology reconstruction. E_j is the residual energy of node j. $SE(i)$ is the total energy of node i and its neighbors. Then each node calculates its own priority to be cluster head.

$$\mathrm{Pro(i)} = \sum_{j \in rNb(i)} \frac{E_i}{SE(j)} \tag{6}$$

where Pro(i) is the priority of node i to be cluster head where the higher priority means the higher probability to be cluster head. We apply a greedy algorithm to determine the cluster heads candidates.

After the calculation of *Pro*(*i*), the Controller sorts the node with the value of *Pro* (*i*) from largest to smallest indicated as set *Chpr*. The node set *CH* represents the nodes which are selected to be cluster head. The Controller selects the node with the maximum value in *Chpr* to be cluster head. The neighbors of the cluster head node which is in the list before reconstruction are marked as the member node. Delete the node in *Chpr* which is marked as member node. Each node in sensor node set "*node*" has a *ChToJoin* domain which records the cluster head candidates that the node will join in.

Cluster Formation. To balance the energy distribution, more sensors should subscribe to a high-energy cluster head. If the node's *ChToJoin* domain just has one element, the element is the cluster head in which the node will attend. If the domain has multiple elements, then it uses the following formula to determine which cluster head to join in:

$$\text{fitness(i)} = \beta \frac{d(i,j)}{d_{maxtoCH}} + (1-\beta)\frac{E_{max} - E_i}{E_{max}} \tag{7}$$

where, *d*(*i, j*) is the distance between node j and cluster head i, *dmaxtoCH* is the maximum distance between a member node i and CHs that cover it. β represents impact factor to determine which factor makes greater impact. The node chooses cluster head i with biggest fitness as its cluster head. Each node in *CH* maintains a *Cmb* domain indicating the cluster member of the cluster head. The cluster head constructs *Cmb* according to the *ChToJoin* domain in each node.

Cluster Head Notification. The Controller generates cluster head notification packet (i.e. *CH_NTF*). The *CH_NTF* packet contains the member information. The sink transmits the *CH_NTF* packet to the corresponding cluster heads. When the cluster head receives the *CH_NTF* packet, it creates corresponding flow table entry in its WISE flow table. The cluster head checks the member information and generates member notification packet (i.e. *Mb_NTF)* that notifies its member nodes. For reducing the congestion, the cluster head generates the time-slot schedule for cluster members based on TDMA [12] and sends it to the cluster members.

4.3 The Multi-hop Routing Mechanism

Routing Tree Construction. In this phase, clusters have been formed. Cluster heads transmit their data to the sink via multi-hop communication. We propose Connected Graph Based Minimum Energy Consumption (CGMEC) multi-hop routing algorithm for inter-cluster communication. The Controller calculates the distance from cluster heads to sink (*DisToSink*) and builds connected graph about cluster heads using *TD_MAX* Let *CH_Nb* represent the neighbors set of cluster heads, *NextHop* represents the next hop of cluster head to sink and *EnToSink* represents the energy consumption to sink. At the beginning, all cluster heads in *TD_MAX* range around sink directly

communicate to sink, set *NextHop* to the ID of sink, calculate energy consumption to sink (*EnToSink*) by:

$$\mathrm{EnToSink(i)} = \mathrm{k}e_{tx} + \begin{cases} k\varepsilon_{fs}d(i)^2 & d(i) \le d_0 \\ k\varepsilon_{mp}d(i)^4 & d(i) > d_0 \end{cases} \tag{8}$$

Then, calculate the *NextHop* of each cluster head. Check node's *DisToSink* and the value of nodes in its *CH_Nb*, and insert the node whose *DisToSink* is smaller than in its *NextHop*. We consider the sink as the source point, perform the breadth first search (*BFS*) of graph theory, and calculate *EnToSink* of the element in *NextHop* for all cluster heads layer by layer. The *EnToSink* is calculated by:

$$\mathrm{EnToSink(i)} = \mathrm{EnToSink(j)} + ke_{rx}(j) + ke_{tx}(i) + \begin{cases} k\varepsilon_{fs}d(i,j)^2 & d(i,j) \le d_0 \\ k\varepsilon_{mp}d(i,j)^4 & d(i,j) > d_0 \end{cases} \tag{9}$$

The Controller updates the *NextHop* of each node based on minimal energy consumption to sink for the purpose of energy conservation. So, the *NextHop* has only one element whose *EnToSink* is minimum. Then, the routing tree has been constructed. The detail of CGMEC algorithm is given in Algorithm 1.

Algorithm 1. Routing Tree Construction

```
Let G={Connected graph constructed based on CH_Nb};
bfs={Set of the result of BFS(G) };
for all x ∈ node do
   if x.DisToSink < TD_MAX then
        Calculate x.EnToSink by (8);
        x.NextHop = sink;
   else
      for all y ∈ x.CH_Nb do
         if y.DisToSink < x.DisToSink then
            x.NextHop = x.NextHop U y;
         end if
      end for
    end if
 end for
 for all x ∈ bfs do
    for all y ∈ x.NextHop do
      Calculate x.EnToSink by (9);
    end for
    Choose the node z which has the min value in x.EnToSink;
    x.NextHop = z;
    x.EnToSink = x.EnToSink(z);
 end for
```

Flow Table Entry Issued. After routing tree construction, the Controller generates packet that contains flow table entries and transmits to the corresponding cluster heads. The cluster head receives the packet and updates the WISE flow table.

Cluster Head Rotation and Rerouting. For the purpose of energy conservation, the Controller monitors the residual energy of each senor node in the network. If any cluster head falls bellow the threshold value ($E_{th} = E_{avg}/2$, E_{avg} is the average of residual energy of the nodes in cluster), it initiates cluster head rotation or rerouting. The Controller selects the node in the cluster whose E_{th} is larger than $E_{avg}/2$ and distance to current cluster head is minimum to be new cluster head. The Controller decides *NextHop* of this node and the nodes whose *NextHop* are old cluster head according to *EnToSink* calculated based on 9. Then, the Controller generates and transmits corresponding flow table entries to the corresponding nodes and transmits cluster member ID of the new cluster head to it. Cluster heads receive the packets and update their flow table.

5 Simulation Processes

In this section, we evaluate the performance of our SDUCR mechanism via simulations. Simulation are carried out in the network simulator OMNET++. Only one controller used in the simulations. The simulation parameters are given in Table 1.

Table 1. Simulation parameters

Parameter	Value
Network coverage	(0,0) (400,400)
Sink location	(200,400)
N	200
Initial Energy	1 J
Transmitter circuity, e_{tx}	50 nJ/bit
Receiver circuitry, e_{rx}	50 nJ/bit
ε_{fs}	10 pJ/bit/m^2
ε_{mp}	0.0013 pJ/bit/m4
Aggregation energy per bit, e_{ag}	5 nJ/bit
d_0	87 m
Control packet size	400 bit
Data packet size	4000 bit

We focus on the overall performance of HEED [7], EEUC [8], UCRA [9], CAUCR [10] and SDUCR. The performances of the protocols were compared based on the network lifetime which is expressed as the round of the first node die.

We investigate how SDUCR prolongs the network lifetime. We define the network lifetime as the number of rounds until the first node dies. For each simulation, 200 sensor nodes were deployed in a square of 400 * 400 m^2, uses 20 different location

sets and ran 100 times to get the average value. The simulated networks use the energy model mentioned in Sect. 3.

In Fig. 1, we compare the network lifetime for the four different algorithms (c = 0.4, N = 200). Figure 1 shows SDUCR prolongs the network lifetime more than the other three. Because HEED is based on equal clustering it does not consider the problem of unbalanced energy consumption among cluster heads. EEUC uses an unequal clustering algorithm based on competition radius, so it generates a longer lifetime than HEED. UCRA uses vote mechanism in clustering phase, it can choose

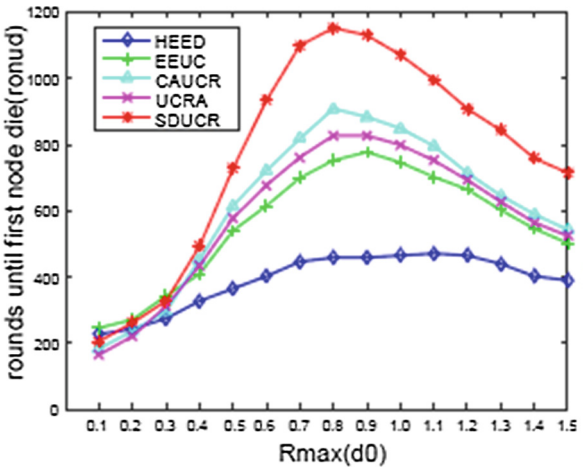


Fig. 1. Network lifetime with different *Rmax*

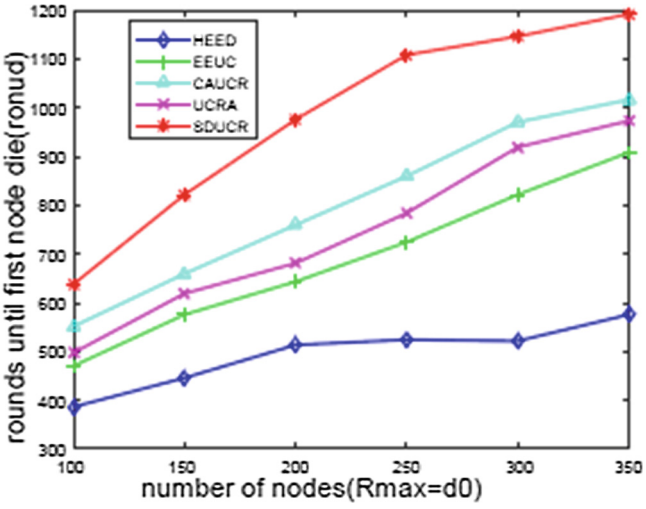


Fig. 2. Network lifetime with different numbers of nodes

more suitable cluster head than EEUC. CAUCR reduces the control overheads in inter-cluster communication phase, thus, it is more energy efficient than UCRA. Our proposed protocol reduces the overall control overheads and generates more suitable cluster and routing path by SDN intelligence and prolongs the network lifetime.

Figure 2 shows the impact of different numbers of nodes on the network lifetime about different protocols. The number of nodes doesn't have much impact on HEED because HEED chooses the cluster head randomly. With the increase of nodes, the network lifetime increases in EEUC, UCRA, CAUCR and SDUCR.

In SDUCR, the centralized clustering and routing tree construction successfully balance the energy consumption and prolong the network lifetime.

6 Conclusion

In this paper, we firstly introduce CUCA to balance the energy consumption among cluster heads and reduce the control overhead by SDN architecture which rotates the cluster heads only if the residual energy of one of the current cluster heads less than the energy threshold instead of rotating every round. Cluster heads closer to the sink can maintain some energy for inter-cluster data forwarding because they have smaller sizes. Secondly, we design CGMEC multi-hop routing algorithm for inter-cluster data communication with minimum energy consumption. Simulation results show that, the SDUCR protocol significantly prolongs the network life cycle.

Acknowledgments. This research is sponsored by the State Key Program of National Natural Science Foundation of China No. 61533011, Shandong Provincial Natural Science Foundation under Grant No. ZR2015FM001, the Fundamental Research Funds of Shandong University No. 2015JC030.

References

1. AI Karaki, J.N., Kamal, A.E.: Routing techniques in wireless sensor networks: a survey. IEEE Wirel. Commun. **11**(6), 6–28 (2004)
2. Mhatre, V., Rosenberg, C.: Design guidelines for wireless sensor networks: communication, clustering and aggregation. Ad Hoc Netw. **2**(1), 45–63 (2004)
3. Singh, S.K., Singh, M.P., Singh, D.K.: A survey of energy-efficient hierarchical cluster-based routing in wireless sensor networks. Int. J. Adv. Netw. Appl. (IJANA) **2**(2), 570–580 (2010)
4. Galluccio, L., Milardo, S., Morabito, G., Palazzo, S.: SDN-WISE: design, prototyping and experimentation of a stateful SDN solution for wireless sensor networks. In: IEEE Infocom 2015, April 2015
5. Costanzo, S., Galluccio, L., Morabito, G., Palazzo, S.: Software defined wireless networks: unbridling SDNs. In: 2012 European Workshop on Software Defined Networking (EWSDN). IEEE (2012)
6. Heinzelman, W.B., Chandrakasan, A., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. IEEE Trans. Wirel. Commun. **1**(4), 660–670 (2002)

7. Younis, O., Fahmy, S.: HEED: a hybrid, energy-efficient distributed clustering approach for ad hoc sensor networks. IEEE Trans. Mob. Comput. **3**(4), 366–379 (2004)
8. Li, C.F., Chen, G.H., Ye, M., Wu, J.: An energy-efficient unequal clustering mechanism for wireless sensor networks. In: IEEE International Conference on Mobile Ad Hoc and Sensor Systems Conference, Washington, D.C., pp. 598–604 (2005)
9. Zhang, R., Ju, L., Jia, Z., Li, X.: Energy efficient routing algorithm for WSNs via unequal clustering. In: High Performance Computing and Communication, pp. 1226–1231 (2012)
10. Zheng, L., Jia, Z., Zhang, R., et al.: Context-aware routing algorithm for WSNs based on unequal clustering. In: 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 1307–1314. IEEE (2013)
11. Jia, F.-l., Li, F., Zhang, R.-h.: RSSI localization based on core in WSN. Comput. Eng. Appl. **44**(30), 118–120 (2008)
12. Cionca, V., Newe, T., Dadarlat, V.: TDMA protocol requirements for wireless sensor networks. In: Proceedings of the IEEE Second International Conference on Sensor Technologies and Applications, pp. 30–35, August 2008

An Energy Efficient and Secure Data Aggregation Method for WSNs Based on Dynamic Set

Jinsheng Zhu and Zhiping Jia(✉)

School of Computer Science and Technology,
University of Shandong, Jinan, China
zhujs_sdu@sina.com, jzp@sdu.edu.cn

Abstract. As the era of big data comes, numerous research works have been directed to focus on the massive, multi-source data acquisition, transmission, storage and management. The wireless sensor network is an important way and means to get "metadata" for big data. However, restricted energy resources and poor security have always been the bottleneck of wireless sensor network. In this paper, based on the method of software and hardware co-design, we introduce non-volatile memory (NVM) into memory system and propose an algorithm DAEE, managing the NVM dynamically to reduce energy cost and meanwhile adopting the security model of dynamic set theory to improve the data security. Experimental results show that the proposed method effectively guarantees the data security, reduces the network data flow and the whole network energy consumption, providing an efficient way for data processing in wireless sensor network.

Keywords: WSNs · NVM · Dynamic set · Software-hardware co-design

1 Introduction

Cloud computing, internet of things, social networking and other emerging services have maked the types and scale of human society data growing at an unprecedented rate. Thus data has become a very important factor in production. Big data [1, 4] has attracted wide attention, and more and more studies focus on the data acquisition, storage, transmission and management.

The wireless sensor network [5, 6] is a hot research field which combines various topics including sensing, wireless communication and embedded computing. Wireless sensor networks can sense, collect and measure various information in real time by collaboration of many integrated micro-sensors. Then the information is transmitted wirelessly and aggregated through multi-hop ad-hoc networks, which connects the physical world, computing world and human society dynamically. It is critical to ensure the stability and efficiency of the data-centric wireless sensor network, making the data stored and transformed safely and effectively. Besides, the wireless sensor node is an important component of wireless sensor network. Owing to the unique characters of wireless sensor network, such as it being deployed widely and applicable to open

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 164–172, 2017.
DOI: 10.1007/978-981-10-3969-0_19

environment, the wireless sensor nodes usually have less energy resources, storage capacity and computing power compared with other computing devices.

In this paper, based on the method of software and hardware co-design, we introduce non-volatile memory (NVM) into the wireless sensor node to extend the memory capacity and propose a software-controlled management scheme that adaptively switches off and on the NVM device to reduce the energy cost, and meanwhile adopt the security model of dynamic set theory to improve the data security in storing and transmitting.

2 Collaborative Design of Software and Hardware

Various application scenarios of wireless sensor networks require wireless sensor nodes to be equipped with the characteristics of cheap, low power consumption, small size and short-range wireless communication, which limits the nodes when it comes to storage, energy consumption and computing power, etc.

2.1 Extending Memory Capacity with NVM

The Big Data era brings opportunities for innovation of data storage technology. With the rapid development of NVM technology, the bandwidth of NVM becomes much higher with decreasing latency. It begins to shake the monopoly position of traditional main memory that is based on volatile dynamic random access memory (DRAM). And increasing capacity and lower price make non-volatile memory (NVM) to become more and more suitable for being one part of main memory system.

The non-volatile memory used in the hardware design is phase change memory (PCM). Compared with the traditional dynamic random access memory (DRAM), PCM has higher density, lower energy consumption and comparable read speed, as shown in Table 1. In this paper, an extra PCM memory chip is added to the wireless sensor node for extending the memory capacity, as Fig. 1 illustrated. Meanwhile we propose a management scheme, which can adaptively switch off and on the PCM device, to dynamically use the space of PCM and reduce the energy cost. With the extended PCM memory capacity, the forwarding packets can be temporarily stored and processed in transmitting nodes, which can reduce the network congestion and achieve the balance of network energy.

Table 1. PCM and DRAM

	PCM	DRAM
Density power	2-4X100-500 mW/GB	1X ∼W/GB
Idle	≪0.1 W	∼W/GB
Refresh	No	Yes
Write latency	150 ns	∼10 ns
Read latency	55 ns	∼10 ns

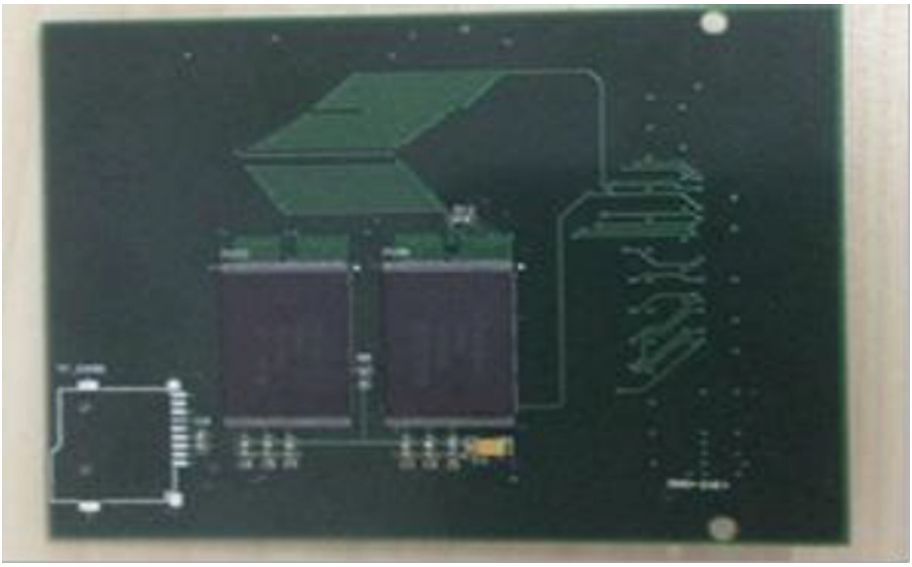


Fig. 1. Extended PCM memory board.

2.2 The Design of Secure Storage Based on Dynamic Set Theory

In wireless sensor network, if the data stored in the non-volatile memory (NVM), such as the PCM used in our hardware design, an unavoidable problem is the security of stored data. Owing to the non-volatility characteristics and wear-leveling mechanism of NVM, which make it possible to generate residual data so that effective information can be obtained from these residual data by data recovery techniques, and the leakage of private and confidential data is an unavoidable result. Data encryption technology is the most common and effective technique to ensure the security of data storage and transmission. However, traditional data encryption mechanisms can't work well in the embedded environment which has limited resources and lower computing power. In this paper, we adopt the dynamic set theory to ensure the security of data storage and transmission.

Set theory is one of the basic theoretical tools used in modern mathematic, information science and system science. For classical sets, they have the following three important characteristics – accuracy, boundary certainty and stability. Zadeh proposed fuzzy set by replacing 'boundary certainty' with 'boundary uncertainty' [7]. Pawlak proposed rough set theory by introducing vagueness instead of accuracy in classical set [8]. Shi proposed packet sets theory called P-sets, which are composed of interior P-sets XF and outer P-sets XF and have dynamic characteristics [14].

In this paper, based on the P-sets theory, we treat the data stored and transmitted in network as a set. The data are dynamically changed based on the data transmission path

and recovered at the sink node. Take the network topology depicted in Fig. 2 for example, there are four sets: $ID = \{ID_1, ID_2, ID_3, ID_4, ID_5\}$ (the set of node identity), $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}$ (the set of node attribute value), $F = \{f_1, f_2, f_3, f_4, f_5\}$ (the set of function rules of each node), $F^{-1} = \{f_1^{-1}, f_2^{-1}, f_3^{-1}, f_4^{-1}, f_5^{-1}\}$ (the set of corresponding inverse function rules of each node).

There are the following constraints:

(1) ID_i is the unique identifier of the i_{st} node; α_i is the exclusive attribute of the i_{st} node; f_i is the only function rule of the i_{st} node, and its unique inverse function rule belongs to the set F^{-1}.
(2) Sink node records the *ID* of all nodes and the corresponding attributes and inverse function rules.

From Fig. 2, we can see that Node 4 and Node 5 have data X and Y respectively. The data need to go through Node 3 and Node 1 to arrive at the sink node and Node 3 is regarded as an intermediate node which stores data in temporarily. The process is as follows:

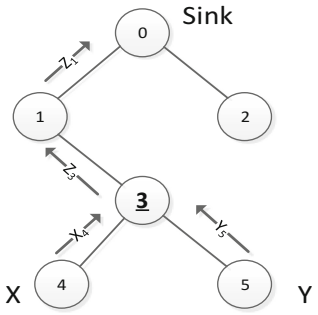


Fig. 2. The evolution process of data during transmission

(1) Data X is processed by Node 4 and becomes X_4, which is sent to Node 3:

$$X_4 = (f_4(X, \alpha_4), ID_4) \tag{1}$$

(2) Data Y is processed by Node 5 and becomes Y_5, which is sent to Node 3:

$$Y_5 = (f_5(Y, \alpha_5), ID_5) \tag{2}$$

(3) Data X_4 and Y_5 are combined at the Node 3 and becomes Z_3, which are sent to Node 1:

$$Z_3 = (f_3(X_4 + Y_5, \alpha_3), ID_3) \tag{3}$$

(4) Data Z_3 is processed by Node 1 and becomes Z_1, which is sent to sink node:

$$Z_1 = (f_3(Z_3, \alpha_1), ID_1) \tag{4}$$

5) Sink node analyzes data Z_1 by using F^{-1} set and restoring the original data X and Y layer by layer.

The sink node maintains the information table of nodes in the whole network, as shown in Table 2.

Table 2. Information table of nodes

ID	f^{-1}	α
ID_1	f_1^{-1}	α_1
:	:	:
:	:	:

3 Algorithm

The energy cost of wireless sensor network mainly comes from the cost of data receiving, sending, collecting, computing and storing, especially the data sending. When the distance of data transmission is constant, the energy cost depends on the length and frequency of data transmission. In this paper, we propose a dynamic adaptive energy efficient algorithm (DAEE), aiming to reduce the frequency and redundancy of data transmission.

Considering that the data packets in wireless sensor network usually have the characteristics of small-size and high-frequency, DAEE algorithm dynamically uses the extended NVM memory space to temporarily store the data needed to be transmitted in a period of time and meanwhile integrates the related data to reduce data redundancy, as a result achieving the goal of reducing energy cost.

The algorithm maintains two variables, Ga and Ge, which indicate the number of packets collected by node itself and received from other nodes, respectively. For indicating the different impact on energy cost caused by the above two types of packets, we add two extra impact factors, μ and §, as illustrated in Algorithm 1.

In *Algorithm 1,* variable R is used to record the amount of data transmitted during r time. If R is larger than the threshold φ, it means that the frequency of data transmission is much higher. So the extended NVM memory space is waken up and used to store the data and reduce the data redundancy. Otherwise the state of the extended NVM memory space is set to be idle for reducing storage energy cost.

Algorithm 1. DAEE Algorithm

```
① Maintain the time variable τ
② Ga=Ga+1 //when data is collected by node itself, Or,
  Ge=Ge+1 //when data is received from other nodes
③R=(μGa+§Ge) /τ;
④If R>ϕ // Waking expanded NVM memory space
    Pcm=1, Ge=0, Ga=0,τ=0;
⑤Else //Making expanded memory space sleep
    Pcm=0;
⑥EndIf
⑦If Pcm=0, Sending the data collected or received di-
  rectly;
⑧Else  Storing the data collected or received into the e
  xpanded storage space; W = 1;
⑨EndIf
⑩While W
⑪ Making the data stored in the expanded memory space in
       tegrated and sending it; Maintaining W flag;
⑫ EndWhile
⑬ Goto ①
```

4 Simulation and Analysis

In our simulation environment, 200 sensor nodes are deployed statically in random, and the data collected by all other nodes converges at the only sink node. There are some assumptions about the sensor node. Firstly, the transmission radius of each node is constant and same. Secondly, the frequency of collecting data varies among nodes. Thirdly, the energy cost of data collecting, receiving and sending is the same in all nodes. Lastly, the energy cost of waking and accessing the extended NVM memory space is same in all nodes. Owing to the extremely lower static power of NVM, we ignore the energy cost of NVM when its state is idle.

In this experiment, DAEE algorithm employs a proactive table-driven routing method to maintain the state of routers. Each node forwards data based on the local routing table. Owing to the fact that the location of node is stable, router maintenance is needed only when the node's energy is used up.

Firstly we propose an experiment about surviving periods of network, and at the same time we study the impact of different threshold φ values (φ1 > φ2.). The corresponding experimental result illustrated in Fig. 3 indicates that our DAEE delays the appearing time of disabled sensor node and prolongs the surviving periods of network, compared with the traditional wireless sensor network algorithm LEACH [9] and DSDV [10]. With the appearance of disabled sensor node, the speed of other nodes becoming disabled in DAEE is faster than that in LEACH, which is caused by DAEE not considering the comparison of remaining energy among nodes.

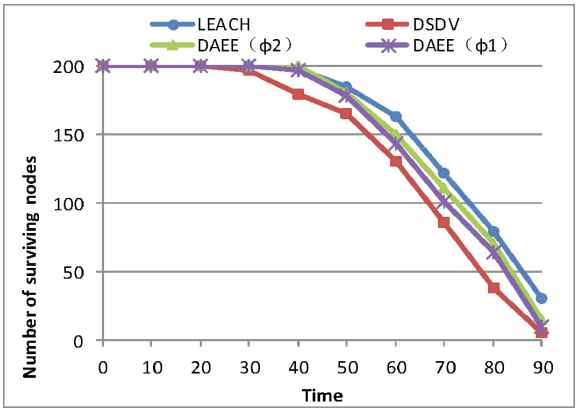


Fig. 3. Surviving periods of network

In DAEE algorithm, when the frequency of data transmission(R) is bigger than the predefined threshold (φ), the extended NVM memory space will be woken up and used to store the data in temporarily, which can reduce the blocking probability of networks. As Fig. 4 illustrated, the ratio of packet arriving when the threshold value is set to φ1 is relatively higher than that when the threshold value is φ2. The reason is that the higher the threshold value, the less the number of nodes in which extended memory space is waken up. Figure 5 illustrates that it delays the time of packet arriving at the sink node to store data in extended NVM memory space. The more nodes with waken-up NVM space, the greater the impact on the delay.

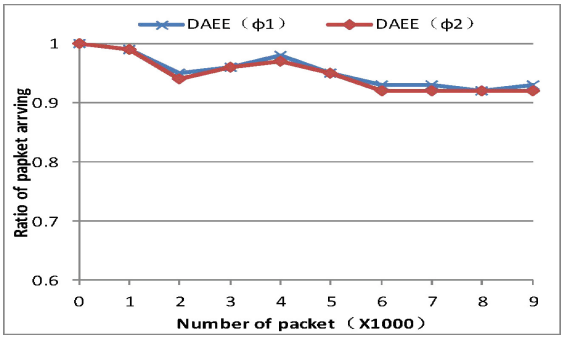


Fig. 4. Ratio of packet arriving in different threshold

The threshold value (φ) determines the number of nodes in which NVM memory space is waken up. Figure 6 illustrates the number of waken-up nodes in different threshold values. To obtain a reasonable delay, the threshold value should be to make the percent of waken-up nodes accounting for all nodes should be about 20.

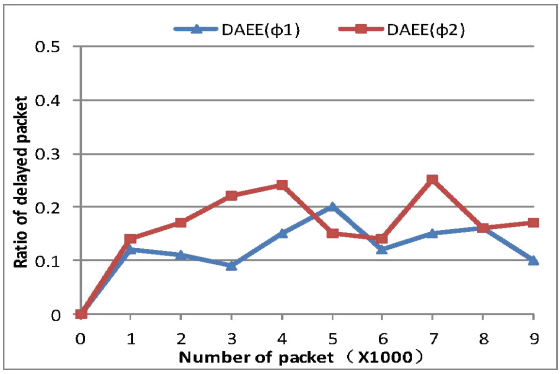


Fig. 5. Ratio of delayed packets with different thresholds

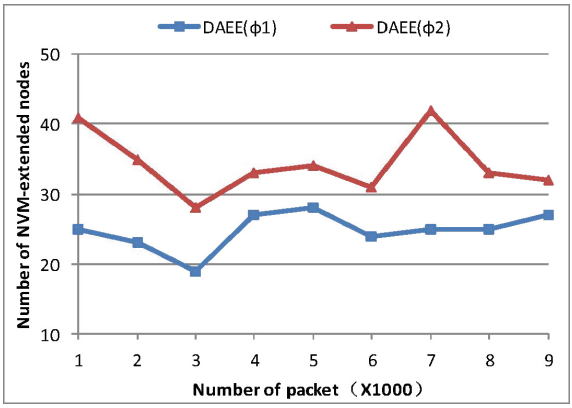


Fig. 6. The waken-up nodes in different threshold

In our simulation environment, the unique attribute (*a)* of each node is a random value ranging between 0 and 65536. Considering the limited computing performance and energy of node, the complexity of function F is $O(n)$. During the process of data transmission, the data will be encrypted every once arriving at a node, which constructs a chain structure of encryption. Because the attribute values and computation functions of each node are different, the data stored and transmitted in networks have better security.

5 Conclusion

In this paper, based on the method of software and hardware co-design, we introduce the NVM of lower static power into the wireless sensor node to extend the memory capacity, which can improve the ratio of packet arriving and prolong the surviving

periods of networks. Meanwhile we propose a mechanism of chain encryption based on the dynamic set theory to guarantee the security of data in networks.

Taking the inner-node residual energy and inter-node relationship of energy consumption into account is our future research direction.

References

1. Manyika, J., Chui, M., Brown, B., et al.: Big data: the next frontier for innovation, competition, and productivity. Analytics (2011)
2. Lohr, S.: The age of big data. New York Times (2012)
3. Mayer-Schönberger, V., Cukier, K.: Big Data: A Revolution that Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt, Boston (2013)
4. Chen, H., Chiang, R.H.L., Storey, V.C.: Business intelligence and analytics: from big data to big impact. MIS Q. **36**(4), 1165–1188 (2012)
5. Akyildiz, L.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. Commun. Mag. IEEE **40**(8), 102–114 (2002)
6. Lotf, J.J., Ghazani, S.H.H.N.: Overview on routing protocols in wireless sensor networks. In: 2010 2nd International Conference on Computer Engineering and Technology (ICCET), vol. 3, pp. 610–614. IEEE, Piscataway (2010)
7. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
8. Pawlak, Z.: Roughsets. Int. J. Comput. Inf. Sci. **11**, 341–356 (1982)
9. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor network. In: IEEE Proceedings of the Hawaii International Conference on System Science, Washington, pp. 3005–3014 (2000)
10. Royer, E.M., Toh, C.K.: A review of current routing protocols for ad-hoc mobile wireless networks. IEEE Pers. Commun. Mag. 46–55 (1999)
11. Liu, F., Chen, Z.G., Liu, Y.P., Xiao, N.: Development and prospect of solid-state storage technology. Comput. Soc. Newslett. China 15–20 (2012)
12. Lu, Y.Y., Shu, J.: Survey on flash-based storage systems. J. Comput. Res. Dev. 49–59 (2013)
13. Wong, H.S.P., Raoux, S., Kim, S.B., et al.: Phase change memory. Proc. IEEE 2201–2227 (2010)
14. Shi, K.q.: P-sets. J. Shandong Univ. (Nat. Sci.) 77–84 (2008)
15. Shi, K.q.: Function P-sets. J. Shandong Univ. (Nat. Sci.) 62–69 (2011)
16. Fan, C.: The Study of Dynamic Information and Dynamic Information Law Characteristic. Shandong University, Shandong (2013)

A Novel Quality Detection Approach for Non-mark Printing Image

Qiong Zhang[1], Bin Li[2(✉)], Minfen Shen[3], and Haihong Shen[4]

[1] Shantou University Medical College, Shantou, Guangdong, China
[2] Shantou Institute of Ultrasonic Instruments Co., Ltd., Shantou, Guangdong, China
lb@siui.com
[3] Shantou Polytechnic, Shantou, Guangdong, China
[4] Department of Electronic Engineering, Shantou University, Shantou, Guangdong, China

Abstract. In printing business, a lot of printing products have no apparent marks for registration, which cause the difficulty of printing image quality auto-detection. Aiming to this problem, a novel quality detection approach for non-mark printing image is proposed in this paper. The proposed approach mainly consists of the region feature based registration region selection and fast shape-based image matching method and an improved difference matching method to detect the printing defects. The proposed approach is realized by the well-known machine vision software HALCON. The experiment results show that the proposed approach can detect the printing defects efficiently with high accuracy, fast speed and strong robustness.

Keywords: Printing image · Defects detection · Registration region · Non-mark printing image · HALCON

1 Introduction

In printing business, the quality of the products is evaluated by the degree of recovering the original pattern, the higher the better. However, in fact, for the imperfection of the printing technique and unavoidable factors, defects often appear on the surface of printing products. The traditional quality inspection is realized by manual work, which costs a great amount of time, manpower and resource. In recent years, the machine vision technique has found application in the printing inspection. The emergence of this new technique effectively solves the problem of time-consuming and poor work efficiency of the conventional manual way [1–3].

In the process of the quality detection based on machine vision technique, image registration is a step of great importance directly deciding the detecting speed and performance. Usually, the printing product has a particular mark, such as solid block with certain color, circular ring, square frame, color bar and so on. These marks can be utilized in the registration process. However, there are still a lot of printing products without apparent marks, such as cigarette brand [1]. Therefore, how to select appropriate registration regions is significant for non-mark printing products. In this paper, a novel

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 173–180, 2017.
DOI: 10.1007/978-981-10-3969-0_20

quality detection approach for non-mark printing image is proposed, which is able to select registration region automatically, match speedily and detect defects efficiently.

2 The Proposed Approach

The detection procedure involved two stages, detection preparation and detection, as shown in Fig. 1. Firstly, the reference image is acquired and used to generate the detection template. Secondly, the registration region based on the region features is selected and the registration template is generated for matching. Thirdly, the fast shape-based image matching method is carried out between the acquired and reference image. Then an improved difference matching method is proposed to detect the printing defects.

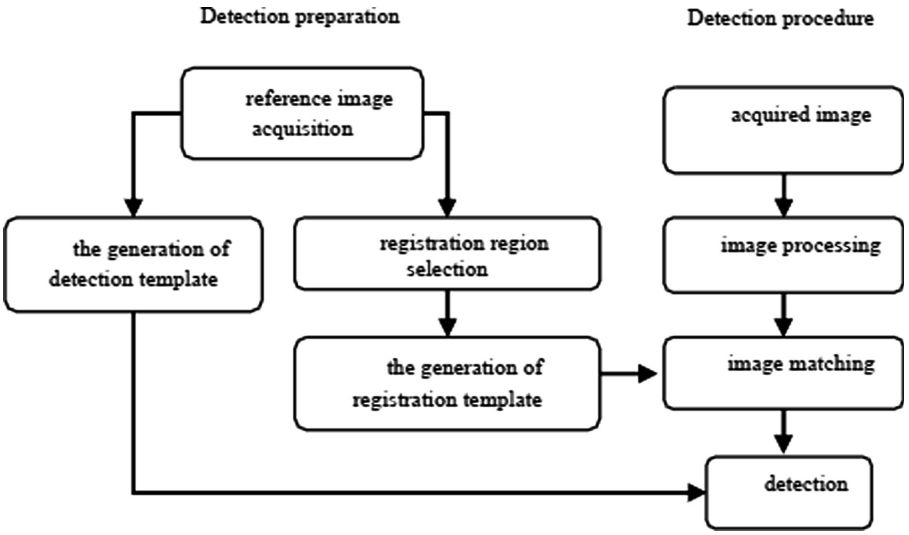


Fig. 1. The detection procedure

2.1 Image Matching Algorithm

Image registration is a process building a relationship between the acquired image and reference image via a certain registration algorithm. It is actually a similarity based optimal searching problem according to a certain matching criteria [4]. The shape-based image matching algorithm was used in the paper, which mainly included three steps, i.e. feature space, similarity measure and search strategy [3, 5].

Feature Space and Similarity Measure Feature space is the feature extracted from the image and used for matching. The correlation characteristic of every registration test depends on the similarity measure. Firstly, the edge gray images of reference and acquired images are calculated and then direction vectors of the edges, which are the feature space, are calculated. After every registration process, the sum of the scalar

products of the direction vectors of every pixel in the reference image and the corresponding pixel in the acquired image are calculated as matching score, which is called similarity measure.

The similarity measure s is calculated by

$$s = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{d'}_{\mathrm{i}}{}^{\mathrm{T}} \boldsymbol{e}_{\mathrm{q}+\mathrm{p'}} = \frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{t'}_i \boldsymbol{v}_{\mathrm{r}+\mathrm{r'_i},\mathrm{c}+\mathrm{c'}} + \boldsymbol{u'}_i \boldsymbol{w}_{\mathrm{r}+\mathrm{r'_i},\mathrm{c}+\mathrm{c'}}\right) \tag{1}$$

where $p_{\mathrm{i}} = (r_{\mathrm{i}}, c_{\mathrm{i}})^{\mathrm{T}}$ is the point coordinate of the reference image, $\boldsymbol{d}_{\mathrm{i}} = (\boldsymbol{t}_{\mathrm{i}}, \boldsymbol{u}_{\mathrm{i}})^{\mathrm{T}}$ is the direction vector related to the point, $q = (r, c)^{\mathrm{T}}$ is the corresponding point coordinate in acquired image which can be found by affine transformation of the reference image, $\boldsymbol{e}_{\mathrm{r,c}} = (\boldsymbol{v}_{\mathrm{r,c}}, \boldsymbol{w}_{\mathrm{r,c}})^{\mathrm{T}}$ is the direction vector related to the point, and $i = 1, \cdots, n$.

When the similarity measure s reaches the user defined threshold $s_{\min}$, it is regarded as a successfully matching example.

Search Strategy In this paper, hierarchical search using graphic pyramid data structure shown in Fig. 2 was proposed as the search strategy. It mainly consisted of five steps:

a. Calculate the graphic pyramid with an appropriate number of levels for both reference and acquired image;
b. Carry out the matching process in the highest layer of the pyramid and search the example matched with the reference image;
c. Map the above result to next layer of the pyramid and define the region around the matching area as new searching region;
d. Carry out a new matching process in the new searching region and map the result to next lower layer;
e. Repeat the search process until to the lowest layer of the pyramid.

Because the search region is small in every layer using graphic pyramid hierarchical search, it is efficient to realize the search process and reduce the amount of calculation.

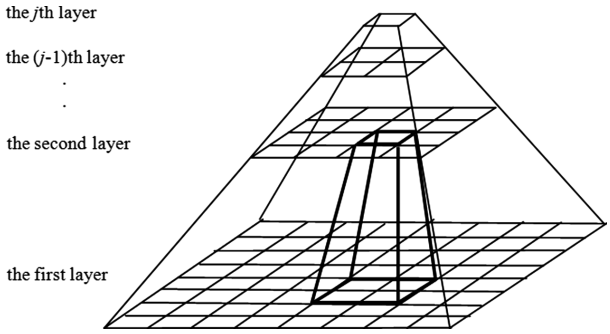
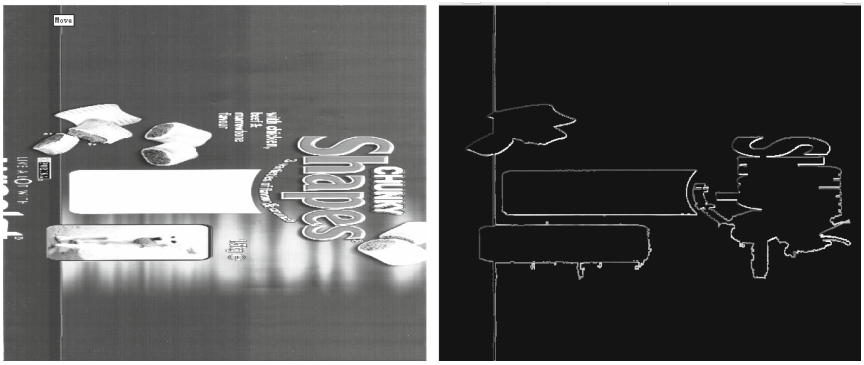


Fig. 2. Graphic pyramid [6]

2.2 The Selection of Registration Region

The shape-based image matching algorithm utilizes the image edge to registration. If the registration region is too small, the edge obtained will be too little, it will reduce the registration precision, even cause false results. Otherwise, there will be too much edge which will reduce the speed and precision. In order to increase the registration speed and precision, the region including the outer edges of biggish shape instead of the entire shape is used as registration region in the paper, as the white area shown in Fig. 3(b).



(a) the gray reference image (b) the selected region of registration(white area)

Fig. 3. The selection of registration region

In this paper, the common used region features, such as area, rectangle descriptor and circular descriptor [6] are used. The detail strategy is shown as below:

a. Detect the edge of the image, delimit and fill the connected region to obtain the solid circular regions.
b. Calculate the area of all the circular regions and select the circular regions with areas not smaller than the threshold.
c. Calculate the rectangle descriptor and circular descriptor of the selected regions and select the region with areas not smaller than the threshold.
d. If there is no satisfied region, reduce the previously set threshold and do step b and c again.
e. To obtain the region including only the outer edge of the selected region.

2.3 Defect Detection

Considering the advantage of difference image matching, such as simple implement and real time, it is adopted to detect the defects in the paper. The difference image matching is calculated according to

$$I_{abs}(r,c) = |f(r,c) - h(r,c)| \tag{2}$$

where $f(c,r)$ is the acquired image, $h(c,r)$ is the reference image, $I(c,r)$ denotes the difference result.

As the gray value of pixels can't be minus, the absolute value of the result is used. In order to obtain a complete defect image, the minus difference needs to be shown. The complete defect image can be obtained by

$$g(r,c) = \begin{cases} b_1 & I_{abs} \geq g_{thred} \\ b_0 & I_{abs} < g_{thred} \end{cases} \tag{3}$$

where g_{thred} is the threshold, $b_1 = 1$ (white), $b_0 = 0$ (black). The value bigger than g_{thred} is regarded as defect.

However, artifacts will exist in the defect image using the difference matching algorithm directly. The artifacts appear after the affine transformation, for the affine transformation is transformed in pixels integer times, while the offset of the image usually is not an integer. Assume the registration precision reaches 100%. If the offset of the acquired image is not an integer, for example, (189.4, 789.8), after affine transformation, it will round up and round down to (189, 790). And the offset deviation (0.4, −0.1) will produce and therefore the artifacts appear after the difference matching process. Obviously, no matter how precise the registration process, it is impossible to eliminate the artifacts. Since the offset deviation is quite small, the artifacts between the qualified and reference printing images will occur only on the edge.

To solve the problem of artifact, an improved difference matching algorithm based on soft and hard thresholds is proposed in the paper. g_{thred} denotes a global hard threshold. $g_{\text{soft}}(r,c)$ denotes the soft threshold which are the gray values after dilating the edge gray image. The image constructed by $g_{\text{soft}}(r,c)$ is called difference image. Combine the soft and hard thresholds, the new threshold $g'_{\text{thred}}(r,c)$ for each pixel, as shown

$$g'_{\text{thred}}(r,c) = \begin{cases} g_{\text{thread}}(r,c), & g_{\text{thread}}(r,c) \geq g_{\text{soft}}(r,c) \\ g_{\text{soft}}(r,c), & g_{\text{thread}}(r,c) < g_{\text{soft}}(r,c) \end{cases} \tag{4}$$

It is effective to eliminate the artifacts and detect the accurate defect image using $g'_{\text{thred}}(r,c)$.

3 Results

3.1 The Construction of the Detection Platform

The quality detection system of printing image based on machine vision is required to be able to acquire, recognize the printing pattern, detect different types of defects and give a report about the type and frequency of the defects. In this paper, a driven roller was used to mimic the workings of the printing conveyor belt. The pending products

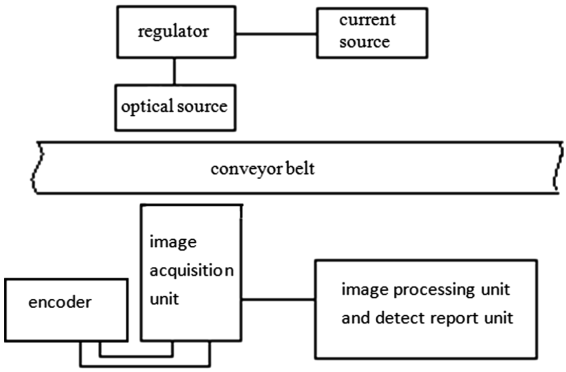


Fig. 4. The construction of the detection platform

were pasted on the roller for detection. The detection platform primarily consisted of optical source, image acquisition unit, rotary encoder, image processing and detect report unit, as shown in Fig. 4. The image acquisition unit was composed by CCD color digital linear camera (Dalsa-PC30-04K80, digital acquisition card (Dalsa-X64 Xcelera-CL PX4 Dual) and camera lens (Myutron FV5026L-F).

To meet the requirement of printing quality detection, LED white strip source were used as optical source. In the detection process, CCD color digital linear camera was

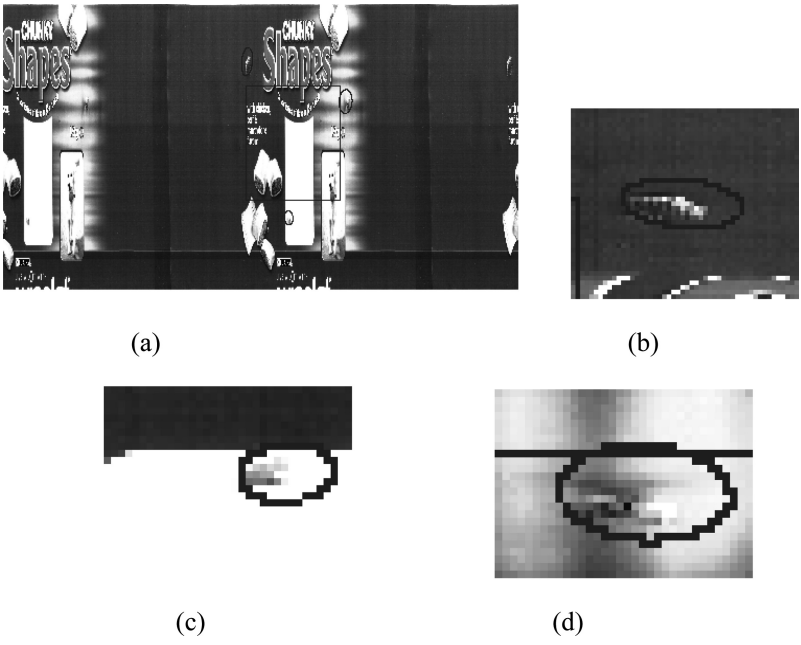


Fig. 5. The detected result with the proposed approach. (a) printing image with defects marked with ellipses. (b) (c) and (d) are the magnification image of the detected result detected image.

working under the way of linear scanning and acquired the image when rotary encoder triggered. Then the acquired image was transmitted to software HALCON which is a powerful machine vision software that provides a comprehensive vision processing base, including all the standard and senior image processing methods, such as fuzzy analysis, pattern match, 3D correction and so on [7].

3.2 The Detected Result

Figure 5(a) is a non-marking printing image scanned with 600 dpi and gray model. It contains three defects. After acquiring the printing image, registration template was searched in the acquired image. The black rectangle frame represents the matching region. Then the defects were detected by the proposed detection algorithm. Figure 5 (b), (c) and (d) are the magnification image of the detected result which marked all the defects correctly. As shown in Fig. 5, (b) represents surface mark. (c) represents a small ink spot. (d) represents an ink stain.

(a)

(b) (c)

Fig. 6. The detected result with small point defects (a) printing image with small point defects. (b) is the magnification image of the detected result detected image. (c) is the zoom in image of defects.

Figure 6(a) is a non-marking printing image scanned with 600 dpi and gray model. It contains four small point defects. After acquiring the printing image, registration template was searched in the acquired image. Then the defects were detected by the proposed detection algorithm. Figure 6(b) is the magnification image of the detected result. And Fig. 6(c) is the zoom in result of the defects area. As shown in the result, all the small defects were detected correctly.

4 Conclusion

In this paper, a novel quality detection approach for non-mark printing image is proposed including the registration region auto-selection, rapid registration and quality detection schemes. The well-known machine vision softeware HALCON is used for realization. The experiment results show the efficiency of the proposed approach. It is believed that the machine vision technology will find more and more application in printing detection with the development of production automation and capacity. Comparing to the methods for particular marks, the proposed approach is much more adaptable. It can used in not only printing detection for various marks, but also other auto-detection fields.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (No. 61302049), Science and Technology planning Project of Guangdong Province (No. 2015B020233018, No. cgzhzd1105, No. 2012B050300024), Science and Technology Planning Project of Shantou and Open Fund of Guangdong Provincial Key Laboratory of Digital Signal and Image Processing Techniques.

References

1. Shang, H.C.: The Study on Algorithm and System Implementation of Printing Image Online Detection. Huazhong University of Science and Technology (2008)
2. Li, C.P., Fan, Y.B., Hu, Q.C.: 3 recognition methods and analysis of PCB mark point based on HALCON. J. Foshan Univ. (Nat. Sci. Ed.) **28**(2), 29–33 (2010)
3. Carsten, S., Markus, U., Christian, W.: Machine Vision Algorithms and Applications. Wiley-VCH, Weinheim (2008)
4. van Beusekom, J., Shafait, F., Breuel, T.M.: Image-matching for revision detection in printed historical documents. In: Hamprecht, F.A., Schnörr, C., Jähne, B. (eds.) Pattern Recognition, DAGM 2007. LNCS, vol. 4713, pp. 507–516. Springer, Heidelberg (2007)
5. Ding, J.H., Lu, Y., Huang, W., Qin, M.: A background subtraction method for defect detection of printed image. Appl. Mech. Mater. **462**, 421–427 (2014)
6. Milan, S., Vaclav, H., Roger, B.: Image Processing, Analysis, and Machine Vision, 3rd edn. Tsinghua University Press, Beijing (2005)
7. Jin, C.: Research on Printing Image Quality Detection Technology Based on HALCON. Central South University (2013)

Passive Packet Reordering Measurement on Terrestrial-Based and Satellite-Based Internet

Zhengguo Xu(✉) and Hui Zheng

National Key Laboratory of Science and Technology on Blind Signals Processing, 610041 Chengdu, China
zhengguo_xu@163.com, hui_zheng08@163.com

Abstract. With the increasing deployment of various medium, the transmission performance has become a key issue of the Internet research. An important impact on the performance is the packet reordering, which is well-known in packet transmission. First, this paper provides a brief review of the existing packet reordering metrics. We also explore the flexibility of existing passive metrics. As a new rising transmission medium, the satellite-based Internet covers widespread areas nowadays, and its transmission performance is susceptible to the packet reordering. We carried out the experiments in practice network environment, and compare the results of the terrestrial-based and satellite-based Internet. Experimental results show the reordering ratio on satellite-based Internet is less than that of terrestrial-based Internet. Finally, we present a novel perspective to explain the reordering phenomenon in TCP/IP flows.

Keywords: TCP/IP · Network traffic · Passive metric · Packet reordering · Terrestrial-based Internet · Satellite-based Internet

1 Introduction

As the key role-played in the architecture of Internet, TCP/IP is designed to provide the best-effort service, various bandwidth-sharing algorithms are implemented for elastic traffic in TCP/IP stack. As a result, the routing packets are delivered without effective guaranty of their ordering, validity or timeliness through the intermediate nodes. An important impact on the performance of transmission is the packet reordering, which is also a common, inevitable, and not pathological behavior in the Internet [1, 2, 14]. The main reason is the local parallelism within routers and switches, and the multipath routing for packet flows. Besides the inherent factors, the packet loss, timeout, retransmissions on link layer, differential service, route fluttering, forwarding lulls may be also the causes of reordering [15].

Because of the content recovery from packets, the receiver needs to reassemble the packets in order, especially for connection-oriented protocols such as TCP. Even for some connectionless applications, it also need to keep packets to render services based UDP, like voice over IP, video on demand and IP videoconferencing. To alleviate the

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 181–190, 2017.
DOI: 10.1007/978-981-10-3969-0_21

affects of packet reordering and improve the performance of transmission, a few metrics are proposed to identify and measure the out-of-order flows [4]. According to whether injecting packets to the network or not, we categorize the metrics into active and passive groups. Active methods employ a test bed through the network on one end or both, then check the order of packets transmitting between sender and receiver. On the contrary, passive methods only take advantage of receiving flows, and make the reordered packet awareness based on the continuity of sequence number of packets. Despite the result of active measurement is precise, some scenarios remain unsolved, and should be tackled in passive way.

(1) **Free of disturbing normal traffic.** The purpose of packet reordering measurement is to reflect the perturbation of network traffic, but probing traffic is possible to disturb real features of existing traffic, and change the pattern from original. However, in passive way, the patterns of network traffic are derived from traffic sampling, it is not harmful to the existing traffic.
(2) **Online processing.** Some passive metrics can be implemented while the traffic passing by. Thus, it is a real-time measurement of packet reordering. Further, the reordering helps to adjust the flow scheduling, adapt the dynamic routing path, and improve the transmission performance promptly.
(3) **Widespread applicability.** Without any configuration of a test bed on specific ends [5], the passive measurement only makes use of sampling traffic at a few observation points in routing paths. It is flexible to be expanded and applied in large-scale network.

Different from present researches, our work gives a comparison of packet reordering in the terrestrial-base and satellite-based network. There are three main contributions in this paper. First we provide the comparison of empirical results between two different networks. Second, we discover a phenomenon of packet reordering in satellite-based network against intuition. Finally, according to the behavior of packet reordering, we offer a way to distinguish the transmission medium, which is in favor of QoS (quality of service) improvement.

The rest of this paper is organized as follows. In Sect. 2, we take an overview of packet reordering and its application. In Sect. 3, we give the difficulties of passive packet reordering measurement, and describe the principle of reordering metrics. In Sect. 4, we show our results of the practical network, and discuss the different reordering performances between terrestrial-based and satellite-based network. In Sect. 5, we conclude our work and point out the considerations of future work.

2 Background

2.1 The Principle of Packet Ordering

In this paper, we mainly focus on the packets encapsulated in TCP/UDP over IP. For measuring the packet reordering passively, we take advantage of the fields related to the order, such as the identification of IP fragments (IPID), and sequence number (SEQ) or acknowledgment number (ACK) of TCP segments. The IPID sequence can

be used in the packet reordering measurement of both TCP and UDP flows, while the SEQ and ACK sequence can only be used to measure the reordering of TCP flows.

(1) **IPID.** To guarantee the uniqueness, a counter is implemented to labeling IP packets. The counter is set to a positive number at the initialization. Commonly and theoretically, when an IP packet is about to send out, the current value of the counter is assigned to the identification field of IP packet, and the counter increases by one [6]. However, the way of the counter working depends on operation systems in fact. Many of them are implemented as a global counter in commercial operating system, including various versions of Windows, Linux version 2.2 and earlier. The value of IPID is maintained with a unique counter in the host. Others are implemented as a per-flow counter, as a random number or a constant [7].

It is apparent to obtain nothing from a random or constant sequence. Using a global counter also has some problems. Because it may be confused by multiple flows. That means if the host is sending out two flows simultaneously, the packets of IPID k and $k + 1$ are separated into two flows. As the packet reordering is defined by flow, the discontinuous points do not pertain to the reordering. So the value of IPID implemented per flow is satisfied the measurement requirement best, but it is not common in the captured flows. Thus, we still use the value of IPID of a global counter in our experiments. Another trivial problem of IPID counting is the number will roll over if it reaches the upper bound, but it is easy to be adjusted.

(2) **SEQ and ACK.** When there are TCP flows in traffic sampling, it is better to take use of the numbering system of TCP instead of IPID, because the counter is defined by flow. There are SEQ and ACK fields in a TCP segment header, the two fields aim to keep the track of the segments transmitted or received, and the value of them is a byte number, not a segment number. The numbering is independent in forward and backward direction, and started at a random number in the range of 0 to $2^{32} - 1$.

During the establishment, transmission, termination, and abortion of TCP flows, each segment consumes one sequence number when it carries one byte. So in practice, the $SEQ[k+1]$ of the next segment equals to $SEQ[k] + L_k]$, where L_k is the length of TCP payload in the kth packet. In addition, the communication of TCP is full duplex, the ACK number of the kth packet is cumulative, which means that the arrived bytes in a TCP flow are unbroken and in order till the $ACK[k]$. In other words, if the packet of $SEQ_{sender}[k']$ responds to the packet of $ACK_{receiver}[k]$, $SEQ_{sender}[k'] = ACK_{receiver}[k]$.

2.2 The Difficulties of Passive Reordering Measurement

We have address the difficulty of using IPID as passive reordering metric briefly. But It is more complex for TCP numbering First, the connection-oriented TCP suffers in various network environment, such as:

(1) **Unnecessary segment retransmission.** All TCP segments are numbered by SEQ and ACK. Once there is a gap in arriving segments, segment retransmission will be triggered by duplicate ACKs, even if the out-of-order packet will arrive a little later. These spurious loss of packets leads to the congestion collapse possibly [3], and increase the difficulty of passive reordering measurement.
(2) **Reducing congestion window inaccurately.** To control flows, TCP use a sliding window. The size of the window is determined by the smaller of receiver window and congestion window. When the arriving packets miss their orders, it often implies congestion occurs in the routing path. The congestion window should be shrunk, even closed to avoid heavier congestion. As the result, the utilization of bandwidth and the performance of transmission both decrease.
(3) **Loss of self-clocking.** Besides confirming the integrity of receiving data, ACK is also used to measure the round-trip time (RTT) and retransmission time-out (RTO), which are important parameters for the ends to be aware of the dynamic routing. TCP sets up a clock when it sends out a data segment, and calculates the delay in round-trip by the receipt of self-clocking, then estimate RTT and RTO respectively. Thus, packet reordering may lead to the loss of self-clocking, and arouse bursty traffic and transient network congestion [3].

Because of the interaction of the controlling mechanisms and packet reordering, the passive metrics may be confused in forecasting packet ordering. However, the no decreasing property of SEQs and ACKs can still be used to measure the reordering in packet sequence. Moreover, as the SEQ and ACK is related to the direction of TCP flows, there are forward path reordering and backward path reordering. For passive measurements, it is obscure to distinguish the directions of sampling traffic. So in this paper, we do not consider the direction of packet reordering.

3 Passive Metrics of Packet Reordering

Several packet reordering metrics are proposed in corresponding application field in recent years. We categorize them into three groups based on their usage and principle.

3.1 Native Reordering Metric

The inequality of $s[k+1] < s[k]$ is the basic and native rule to identify the out-of-order packet in sequence. In RFC 4737 [4], reordering packet ratio (RPR) use this rule to compute the degree of packet reordering. RPR forecasts the next expected sequence number *ne* by default step one, which is suitable for the active measurement. We improve the method to be compatible with passive metrics, and the major difference is the rule of *ne* updating. RPR updates when the condition $s[i] \geq ne$ is satisfied, but the problem is it is low robustness against the network perturbation. For example, if the sequence number of arriving packets is $(1, 99, 3, 4, 5, \ldots)$, where the 99th packet seems to be too early to arrive, the follow-up $(3, 4, 5, \ldots)$ would cause a false reordering ratio drastically. However, the packets may be all in order, except for the one affected by the disturbance or bit error, especially in wireless environment.

Instead, our modified algorithm identifies the reordering only by the previous packet, and *ne* is updated to $s[i] + step$ for every arrived packet, where *step* equals one in IPID sequence, or the length of TCP payload in SEQ sequence. It can prevent the diffusion of burst error in the sequence. And the loss and timeout of packets also are taken in consideration of the modified algorithm to fit the practice situations. Note that ACK sequence cannot be calculated by this metric, because the next expected number cannot be forecasted in ACK sequence.

Another metric grouped in this type is named reordering-free runs (RFR) [4]. It is defined based on a count of consecutive packets in order. RFR does not only quantize the reordering, but also measure the fluctuation of reordering-free runs in a sequence, which can be applied in network evaluation. In our experiments, we advance this metric in the same way to overcome the problem as we mentioned in RPR.

3.2 Reordering Extent Metric

The remained methods proposed in RFC 4737 are all based on lateness. The key point of the metrics is the definition of reordering extent *rex*. Reordering Late time offset (RLTO), reordering byte offset (RBO), reordering gap/gap time (RG), and n-reordering (NR) are either based on or derived from the *rex*. Here we describe these metrics briefly, the details can be reference to [4].

Reordering extent only concerns the continuity of the receiving sequence, and it is independent of *ne*. This simplifies the computation for all kinds of sequences mentioned previously. The main idea is to find out the maximum distance between a reordered packet and the earliest arrived packet which contains a larger sequence number. If a packet is in order, its *rex* is undefined. Formally, consider a sequence of packets $(1, 2, \ldots, N)$, and a distance function $d_i(j) = i - j$ between the *i*th and *j*th packet, *rex* is defined as:

$$rex_i := \max d_i(j) \tag{1}$$

The merits of this metric is it combines the frequency of reordering in time and the gap of reordering in space. It helps to estimate the cache memory for order restoring. One bug of computing rex_i is that, in some cases, the rex_i tends to be overestimated because of too early packet arrivals. This bug can be fixed by setting a threshold, such as the upper bound of extent, late time or byte offset. If rex_i is out of the bound, the *i*th packet should be discarded, and the order of packet $i, i+1, \ldots$ should be also reassigned.

Based on the rex_i, several metrics are derived. RLTO calculates the interval of packet i and $i - rex_i$. RBO sums up the payloads of packet j where $s[j] \geq s[i]$ and $i - rex_i \leq j < i$. RG computes the gap between two discontinuities k and k', where $k = i - rex_i, k' = j - rex_j, i < j$. NR can be deemed as a variant and special case of basic reordering extent metric. The *i*th packet is called n-reordering if and only if $\forall_i - n \leq j < i, \exists s[j] > s[i]$. Notice that the definition of reordering is bias, different from the reordering extent [8]. For instance, if $s[i-1] < s[i], s[i+1] < s[i], s[i-1] > s[i+1]$, then only packet i could be defined as n-reordering.

In general, this group of reordering metrics can reveal precise and diverse estimations of the reordering performance in different kinds of flows. They are in favor of predicting the boundary of memory or delay, particularly for jitter-sensitive applications. For the purpose of our work in this paper, the major advantage of these metrics is their feasibility of passive measurement.

3.3 Reordering Density Metric

There are two metrics, reordered density (RD) and reordered buffer-occupancy density (RBD). They are both presented by Jayasumana et al. [9, 10, 16] recently. The metrics performs the reordering degree of packets one by one, and shows a probability density distribution of reordering in sequence. Both the metrics are based on the reordering extent as well, though, there are called displacement or buffer-occupancy, and differs from the previous definition.

RD captures displacements of packets from their original positions, Consider a sequence of packet $(1, 2, 4, 3, 5)$ without any loss or duplicate, and the index sequence of the receiver is assigned to $(1, 2, 3, 4, 5)$. Then the displacement of each packet is the deviant of the reordered sequence number against the index sequence, it is $(0, 0, -1, 1, 0)$ for this case. The reordering density is the frequency of the displacements. The requirement of the index sequence implies its active usage. When a packet is lost and repeated, its index is not assigned, in other words, the aim of index sequence is to restore the order of arrived packets. But in passive way, it is hard to accomplish the aim.

RBD defines the displacement of each arrived packet in another way. It defines an occupancy of a virtual buffer. For the same example of $(1, 2, 4, 3, 5)$, the expected number of packet 2 is 3, but packet 4 is reordered, when packet 3 arrives, the occupancy of virtual buffer is 1. RBD can help to make sense of cache buffer planning. Subject to the practice limitation, a threshold can be used to filter the time-out or duplicate packets. However, the problem of this metric is the same as we discussed previously, if the former packet is lost, the sequence number of expected packet cannot be predicted. If we reset the counter after a lost and unpredictable point, the measurement can still work in someways.

4 Experiments and Results

In order to observe the reordering behavior and compare the differences of packet reordering inherent in wired and wireless network, we choose two kinds of representative backbone networks to accomplish our experiments. The one is the fiber network, the another is satellite-based network. The traffic collectors are set near local hosts. Figure 1 illustrates the setting of our experiments. Furthermore, to protect the privacy of users, all of the collected traffic are anonymized, only a few related fields of protocol header are extracted from raw data.

In terrestrial-based Internet, we capture daily traffics in an OC-48 link of CERNET, and we access into the satellite-based Internet through an iPSTAR terminal. The

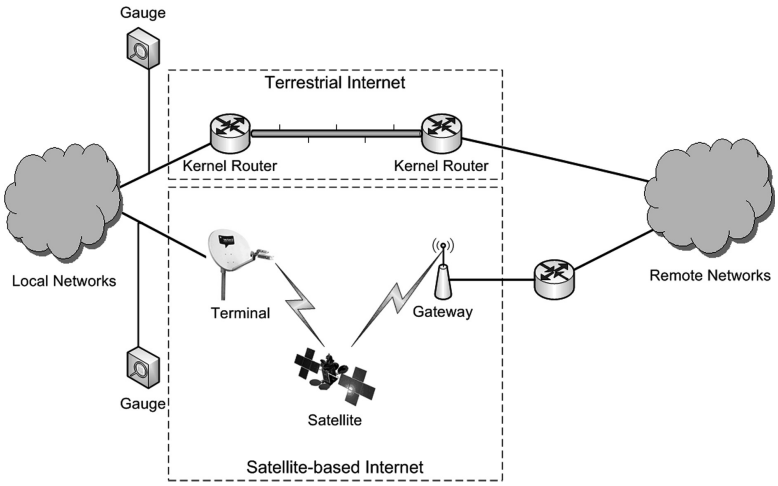


Fig. 1. The illustration of experiment settings

bandwidth of OC-48 is 2.5Gpbs, while for each iPSTAR terminal, the bandwidth of downlink is 8Mpbs, and uplink is 4Mbps. The transmission speed of iPSTAR is commensurate with ADSL in wired network.

There are some substantial differences between terrestrial-based and satellite-based Internet, which may influence the reordering behavior of transmission. Because of the high delay and error rate in wireless environment, the rules of transmission are different from the original design of TCP/IP. For example, there is no onboard processing on iPSTAR satellites to relay the communication of two remote hosts on the ground. All routing operations are processed at the gateway on the earth. It can raise the transmission rate to the best of its ability. Due to the high latency, whose typical value is 250–280 ms in round-trip, and error-prone of satellite links, several improvements are dedicated to enhance TCP performance. TCP acceleration acts as a proxy to handle the connections between terminals and satellites. Three approaches are implemented for transparent and seamless TCP transmission, including TCP spoofing, TCP splitting and Web caching.

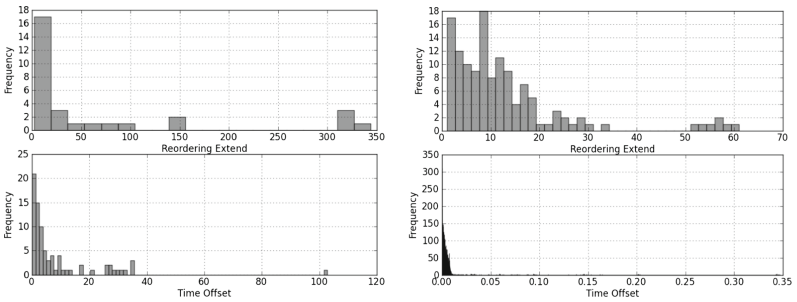
In our experiments, we use the SEQ sequence of TCP to measure the reordering ratio of transmission. Table 1 gives an overview of reordering flows in iPSTAR and OC-48 links by RPR metric. The percentage of reordered flows and the average RPR are much lower in iPSTAR links than OC-48 links. Intuitionally, high error rate and mobility would cause more retransmissions and multiple path routings in wireless environment. But in fact, the practice results show that OC-48 links suffer more

Table 1. An overview of packet reordering in iPSTAR and OC-48 links

Link type	Total	Reordered	Reordered ratio	Average RPR
iPSTAR	3579	61	1.70%	39.70%
OC-48	11642	464	3.99%	60.94%

heavily. The reason lies in the high speed of parallel forwarding in fiber network. Several empirical studies also confirmed this phenomenon [5, 13] in different speed of wired links, and our results of wired links are consistent with theirs.

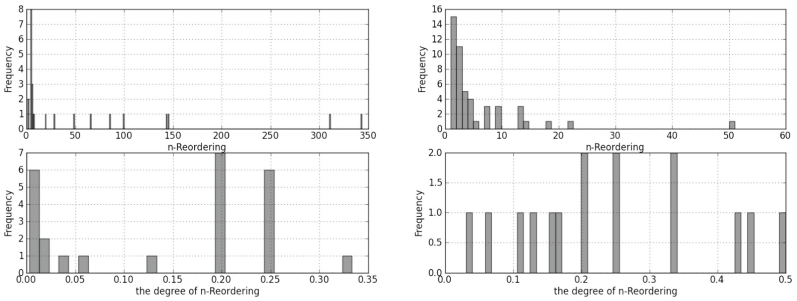
To understand the reordered sequence more precisely, we use the metrics based on reordering extent to measure the flows over iPSTAR and OC-48 links. We still take use of the SEQ sequence. Figure 2 shows the results of reordering extent metric and the derived metric RLTO. It is clear to find out the reordering extent and its late time offset of OC-48 are smaller than that of iPSTAR. Considered the results of Table 1 and Fig. 2, the conclusion is, there is high reordering ratio in OC-48 link, though, small reordering extent occurs in each flow, and for iPSTAR vice versa. In other words, flows over satellite link do not tend to be reordered by the affection of error-prone and high latency, but once the flow is reordered, the reordering extent may be affected severely.



(a) iPSTAR (b) OC-48

Fig. 2. The reordering extent and time offset of iPSTAR and OC-48

Figure 3 illustrates the results of n-reordering measurements. The peak of NR over iPSTAR is 4, that means with a high probability the reordered packet would arrive after 4 packets on the average. Thus, if the threshold of DUP-ACK in an implementation of

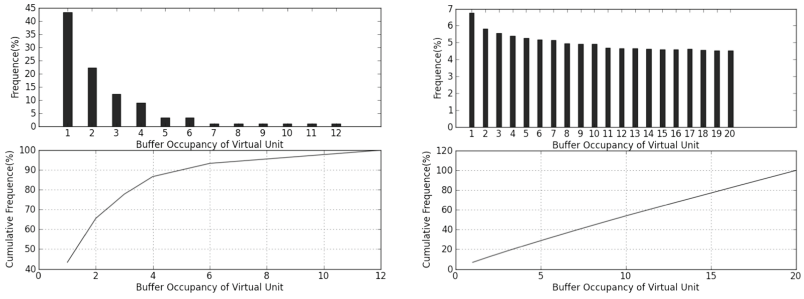


(a) iPSTAR (b) OC-48

Fig. 3. The n-reordering of iPSTAR and OC-48

TCP is reset to 4 instead of 3, it reduces a plenty of retransmissions caused by 3 duplicate ACKs. While the peak is 1 over OC-48, that indicates frequent occurrence of packet reordering because of the parallel forwarding.

For the third group of reordering metrics, we use RBD to measure the SEQ sequences of TCP in OC-48 and iPSTAR links. Because RD is not suitable for passive measurement, RBD is considered only. Figure 4 shows the results. We set a threshold of buffer occupancy for OC-48 and iPSTAR respectively, 20 for OC-48 and 12 for iPSTAR. There is a significant difference in the two types of links. The peak of buffer occupancy is not conspicuous in the flows of OC-48. In fact, if we set a larger threshold, a smooth line of buffer occupancy will stretch all the same. By contrast, it is clear in Fig. 4(b) that the cumulative buffer occupancy, which nearly holds 90 percents, is less than 5. The reason is many connections in OC-48 contain duplicate 5-tuples of IP, due to our definition of flows, some different connections are mixed, but the order of sequences in different connections are irrelative, so there is a untrue large buffer occupancy in OC-48 link. For the iPSTAR, there is a valuable deduction. If the buffer of the terminal is limited to 6 cache units, it can endure most of the reordering packets, and restore them without any retransmission.



(a) iPSTAR (b) OC-48

Fig. 4. Reordered buffer-occupancy density of iPSTAR and OC-48

5 Conclusion

In this paper we take an overview of the passive reordering metrics, and measure the packet reordering of TCP/IP flows in terrestrial-based and satellite-based links: OC-48 and iPSTAR respectively. Based on the survey of the reordering metrics proposed in recent years, we address the usage of passive metrics, and discuss the applicability of proposed reordering metrics, some of them are not suitable for passive measurements, the others should be adjusted. Then we use the passive metrics to accomplish several experiments in practical environment. The results of reordering ratio in OC-48 and iPSTAR are against the intuition. Though error-prone and high latency may raise the probability of retransmission in wireless links, but the reordering ratio of iPSTAR is less than that of fiber links. Besides the comparison of the reordering ratio between

terrestrial-based and satellite-based Internet, the reordered extent can provide more efficient and flexible estimation of TCP parameters to improve its performance, and this will be our future work.

References

1. Paxson, V.: End-to-end internet packet dynamics. ACM SIGCOMM Comput. Commun. Rev. **27**(4), 139–152 (1997)
2. Bennett, J.C.R., Partridge, C., Shectman, N.: Packet reordering is not pathological network behavior. IEEE/ACM Trans. Netw. **7**(6), 789–798 (1999)
3. Floyd, S., Fall, K.: Promoting the use of end-to-end congestion control in the internet. IEEE/ACM Trans. Netw. **7**(4), 458–472 (1999)
4. Morton, A., Ciavattone, L., Ramachandran, G., Shalunov, S., Perser, J.: Packet reordering metrics (rfc4737) (2006)
5. Gharai, L., Perkins, C., Lehman, T.: Packet reordering, high speed networks and transport protocol performance. In: Proceedings 13th International Conference on Computer Communications and Networks, pp. 73–78 (2004)
6. Chen, W., Huang, Y., Ribeiro, B.F., Suh, K., Zhang, H.: Exploiting the IPID field to infer network path and end-system characteristics. In: International Workshop on Passive and Active Network Measurement, pp. 108–120 (2005)
7. Bellovin, S.: A technique for counting NATted hosts. In: Proceedings of ACM Internet Measurement Workshop (IMW), pp. 267–272 (2002)
8. Piratla, N.M., Jayasumana, A.P.: Metrics for packet reordering a comparative analysis. Int. J. Commun Syst **21**(1), 99–113 (2008)
9. Jayasumana, A.P., Piratla, N.M., Banka, T., Bare, A.A., Whitner, R.: Improved packet reordering metrics (rfc 5236) (2008)
10. Piratla, N.M., Jayasumana, A.P., Bare, A.A.: Reorder density (RD): a formal, comprehensive metric for packet reordering. In: International Conference on Research in Networking, pp. 78–89 (2005)
11. iPSTAR. http://www.ipstar.com
12. Iyer, T., Boreli, R., Sarwar, G., Dwertmann, C.: Data acceleration and reduction technology. Satellite and Space Communications (2009)
13. Jaiswal, S., Iannaccone, G., Diot, C., et al.: Measurement and classification of out-of-sequence packets in a tier-1 IP backbone. IEEE/ACM Trans. Netw. **15**(1), 54–66 (2007)
14. Govindarajan, J., Vibhurani, N., Kousalya, G.: An analysis on TCP packet reordering problem in mobile ad-hoc network. Indian J. Sci. Technol. **8**(16), 1 (2015)
15. Leung, K.C., Lai, C., Li, V.O.K., et al.: A packet-reordering solution to wireless losses in transmission control protocol. Wirel. Netw. **19**(7), 1577–1593 (2013)
16. Narasiodeyar, R.M., Jayasumana, A.P.: Improvement in packet-reordering with limited re-sequencing buffers: an analysis. In: 2013 IEEE 38th Conference on Local Computer Networks (LCN), pp. 416–424 (2013)

Research on the Description Method of the Atomic Services in Extensible Network Service Model

Jie Ren(✉) and Jun Shen

School of Computer Science and Engineering,
Southeast University, Nanjing 211189, China
renjie_seu@l26.com

Abstract. The thesis firstly analyzed the disadvantages of the weak service extensibility in traditional network service model, and summarize the status quo of the research on network service extension. Secondly, it introduced the fundamental principles of Extensible Network Service Model (ENSM), more importantly, it focused on the ENSM's kernel about the definition and conceptual model of atomic services. The description method of atomic service is studied in detail, and two specific descriptions of atomic services were enumerated as examples. The paper finally conducted two experiments about data transmission from a new best-effort service model constructed by the dynamic combination of atomic services to a traditional best-effort service model. On the basis of experiments' results, the correctness and feasibility of the atomic services' description method has been demonstrated.

Keywords: Atomic service combination · The Extensible Network Service Model · Network service model · Network architecture · The description method of network service

1 Introduction

With the gradual development of network application requirements, the extensibility of network services has become the key to computer network research. The best-effort service model in the traditional IP network, the integrated service model which can guarantee end-to-end QoS, and the differentiated service model which can provide better service quality assurance and support real-time application, they are exploring the way of service delivery and the ability of service extension in the traditional network architecture. Traditional network lacks a unified service delivery model. It extends the service by continually deploying new protocols and making adjustments to the original hierarchy. This approach is gradually showing a series of drawbacks, just as the network system will be more and more complex, the network protocol will be functional redundancy, the cycle of developing and deploying protocol will be longer, the network system efficiency will be lower [1].

There has been a large number of studies about the extensibility of network services in the field of computer network at home and abroad. A. Lazar (1997) proposed a

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 191–199, 2017.
DOI: 10.1007/978-981-10-3969-0_22

programmable network model which can customize network services according to the needs of application [2]. It attempted to change the static model in the traditional network which can only provide a fixed service for all applications. Braden et al. (2002) proposed a non-hierarchical role-based network architecture [3] hoping to solve the difficult problem of expanding new services in traditional network. The SILO structure [4] (2007) was proposed in the plan FIND which defines the smallest functional building blocks to provide independent network services. In addition to, the recursive network architecture RNA [5] proposed by Joseph D. Touch and the service unit based network architecture [6] proposed by Zeng Jiazhi are related researches around the next generation of extensible network services. These studies are trying to break down the concept of the inherent level in traditional system to flexibly organize network services, but they did not give a unified network service model and their service expansion capacity is also inadequate.

Extensible Network Service Model (ENSM) explores and improves the traditional network service model from the perspective of service extensibility. It decouples the control logic and the business logic, based on abstracting and decomposing the basic functional unit of the traditional hierarchical protocol then dynamically combining them on demand. It can provide and extend network services by programming to implement dynamic composition of atomic services. The atomic services' dynamic composition, as one of the key elements, is the foundation of the ENSM system to provide network services and maintain service extensibility. So the definition and description of atomic services is the core of the Extensible Network Service Model.

The thesis firstly introduced the fundamental principles of Extensible Network Service Model, then it focused on the ENSM's kernel about the definition and conceptual model of atomic services. Secondly, the description method of atomic service is studied in detail, and two specific descriptions of atomic services were enumerated as examples. Finally, it demonstrated the correctness and feasibility of the description method by two experiments.

2 The Fundamental Principles of ENSM

The ENSM which is a new network service model decouples the control logic and the business logic. Firstly, the common mechanisms and processing modes of traditional protocol implementation are abstracted and summarized. It breaks down the network services in the traditional protocol into sub-services with smaller granularity, which can complete an independent and no longer distinguishable service function. These sub-services are called atomic services. Secondly, according to user' needs, the ENSM dynamically combines multiple independent atomic services into service instance. Run the service instance to provide the required service. Especially, atomic services and service instances belong to two different levels. The atomic services provide service to service instances, while the service instances provide service to user by dynamically combining the atomic services. The ENSM implements the dynamic extensibility of network services by dynamic extension of the set of atomic services and the dynamic combination of atomic services. Figure 1 shows the evolutionary process from traditional network service model to Extensible Network Service Model.

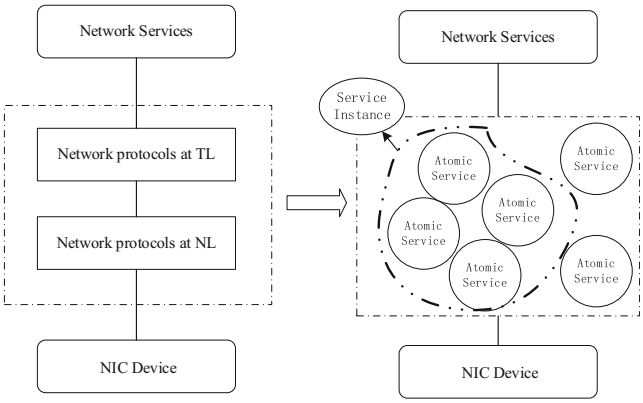


Fig. 1. The evolution from traditional network service model to ENSM

Formula 1 expresses the form of ENSM.

$$\text{xSery} = (\text{AS},\ \text{SR}) \tag{1}$$

xSery expresses the ENSM, AS expresses the atomic service set of ENSM, and SR expresses the set of relationship between atomic services. Every network service is equal to an instance of xSery, and the SR is equal to the control logic of the service instance. As to the same AS, there are many different combinations of atomic services, so that the ENSM can provide many different kinds of services. The ENSM can dynamically add new atomic services to AS to implement the extensibility of network services.

3 The Definition, Conceptual Model and Key Characteristics of Atomic Service

As the kernel of the ENSM, the atomic services provide the most basic functional services. The atomic service is a functional unit which can provide indivisible minimum network service. They are independent of each other and support parallel execution.

Based on the fundamental principle of the ENSM, this paper starts with the basic elements of atomic service. Each atomic service needs to have three basic elements: meta-operation (MO), behavior or constraint operation (BCO), attribute information (AI). So the conceptual model of atomic service is expressed by the triplet method as Formula 2.

$$\text{AS} = (\text{AI},\ \text{MO},\ \text{BCO}) \tag{2}$$

AI expresses the relevant information of atomic service like name, path, and so on. MO expresses the core function of atomic service which is an indivisible minimum

network service provided by itself. In the process of implementing MO, some internal auxiliary operations (behavior operation, BO) and related binding operations (constraint operation, CO) like self-management operation, message-exchange operation. Of the three elements, only MO can provide service, while BO and CO can indirectly determine the relationship between atomic services. And AI is mainly used to distinguish between atomic services.

Based on the definition and conceptual model of the atomic service, atomic service has three obvious characteristics: minimum corpuscular property, parallelizability, and independence.

(1) Minimum Corpuscular Property. Each atomic service provides an indivisible minimum network service. So compared with the protocols embedded with many complex functions as a service provider in the traditional network service model, the minimum corpuscular property of atomic services will guarantee atomic services can more flexibly provide more corpuscular network services. So that the ENSM with atomic services will overcome the difficult problem about protocol functional redundancy in the traditional network service model.
(2) Parallelizability. Based on the fundamental principle of the ENSM, the ENSM dynamically combines multiple independent atomic services into service instance. Service instance is the provider of network service. An atomic service must support the running of multiple service instances at the same time. This is the parallelizability of atomic service. The characteristic can meet the needs of the extension of the network services and guarantee the basic performance of the network system.
(3) Independence. Atomic services come from network protocols, but they are not limited by the level of traditional network protocols. They are independent of each other, and there is no fixed affiliation. No matter which layer the network service it needs to provide belongs to, the ENSM will dynamically combine atomic services needed according to a specific combination mechanism. Completely breaking the traditional network-level restrictions, it makes the provision and extension of network services are more flexible.

In summary, the atomic services' characteristics of minimum corpuscular property, parallelizability, and independence provide the most powerful support for the ENSM, which has a highly efficient, flexible, and non-redundant extensibility of network services.

4 The Description Method of Atomic Service

The conceptual model of atomic service will be described in an approach with two modules: information description and function description. The description method must be highly recognizable on information and non-redundant on function. Information description of the atomic service includes attribute information such as atomic service class, path, and parameter type of meta-operation. And function description includes three types of operation such as meta-operation, behavior operation and constraint operation.

(1) Information Description. It needs to describe the atomic service's own attribute information such as name, storage path, category and functional overview. To facilitate the control engine system (the system to support the running of ENSM) to load and combine atomic services, the information description needs to record parameter type used in meta-operation. The types of parameter can be some simple system types, such as int, double, bool, string or complex user-defined data types.
(2) Function Description. It needs to describe three types of operation: the meta-operation (MO) which provides external network services as the nature of atomic service, the internal behavior operation (BO) which assists in the running of meta-operation, and the constraint operation (CO) which implements self-management and message-exchange. For example, as to the Construct DataPacket service, its MO is constructing a data packet in the specified format. When this data packet is being constructed, it needs populating the fields and checksum. The two operations are its BO. While its initialization and message-exchange with the SendDataPacket service are its CO. BO and CO do not provide network service outward.

Especially, In order to ensure the independence of atomic service, the description of the relationship between atomic services will be involved in the combination of service instances, and will not be considered at the atomic service level.

With reference to the concept of object-oriented programming (OOP), atomic service can be regarded as a service object which comes from the instantiation of the corresponding atomic service class. For each atomic service object, its data member is equal to its information description, and its member function is equal to its function description. Based on the three characteristics of OOP, such as encapsulation, inheritance and polymorphism, many atomic service classes can inherit from a unified atomic service base class.

According to the atomic service description method given above, two specific atomic services are described in detail as examples. They will be used in the experiments in Chap. 5.

(1) ConstructDataPacket Service

Define the service that this atomic service needs to provide: constructs a complete TCP/IP packet according to the fixed TCP/IP packet format, then print the packet on the screen.

The information description of the service: ID number is 6, name is "Construct DataPacket", storage path is "ConstructDataPacket.dll" in the local path, the parameter types are "dataPacket" and "psdTcp_header" as user-defined data types (used to store the TCP/IP packet constructed). The function description of the service: ① Initialization, ② Receive the command message, ③ Calculate the checksum, ④ Populate the fields, ⑤ Construct the packet, ⑥ Print the packet on the screen, ⑦ Return message, ⑧ Clear.

Among them, MO is ⑤, BO is ③④⑥, CO is ①②⑦⑧.

(2) SendDataPacket Service

Define the service that this atomic service needs to provide: send the TCP/IP packet constructed to the NIC equipment, then print the result of sending.

The information description of the service: ID number is 7, name is "SendDataPacket", storage path is "SendDataPacket.dll" in the local path, no parameter type. The function description of the service: ① Initialization, ② Receive the command message, ③ Populate the fields, ④ Send the packet, ⑤ Print the result, ⑥ Return message, ⑦ Clear.

Among them, MO is ④, BO is ③⑤, CO is ①②⑥⑦.

5 Experiment and Analysis

In order to demonstrate the correctness and feasibility of the description method, many atomic services have been programmed according to the description method and software architecture given above. Structuring the traditional TCP/IP network best-effort service with these services, implement the basic network function of data transmission through dynamic combination of atomic services rather than traditional network protocol.

The experiment needs to break away from protocol-encapsulation in traditional network model. So the paper decided to use WinPcap technology [7, 8] to conduct transmission and capture of data packet.

There had conducted two experiments.

(1) The transmission and reception of data packets between service instance nodes

Loading service instance combined by atomic services at two nodes on the network, the transmission and reception of data packet between two nodes were completed through the ENSM.

(2) The transmission and reception of data packets between service instance node and traditional network node

At one node, loading the service instance combined by ConstructDataPacket Service and SendDataPacket Service, a complete TCP/IP packet was constructed and sent to the traditional network. At the other node, capture the data packet via Wireshark then match the packet constructed and the packet captured.

The results of the two experiments are analyzed below.

5.1 The Transmission and Reception of Data Packets Between Service Instance Nodes

Some atomic services were programming implemented according to the description mothed above, such as Initialization Service, Checksum Service, Construct Service, Populate Service, Split Service, Send Service, Capture Service, and other Message Services. Respectively, run the control engine system at two computers (in the experiment, named Host F and Host R) connected with network cable. Among them,

Host F loaded Initialization Service, Checksum Service, Construct Service, Populate Service, and Send Service, combining into a service instance. It successfully constructed a data packet with the content "this is a greeting message" and sent it to Host R. Host R loaded Split Service, Capture Service, Parse Service, and Print Service, combining into a service instance. It captured the data packet, parsed it, and printed the corresponding content on the screen.

Figures 2 and 3 show the result of the experiment. Host R successfully constructed the data packet constructed by Host F, demonstrating that the service instance nodes combined by atomic services implemented by the description method above in the ENSM are interworking.

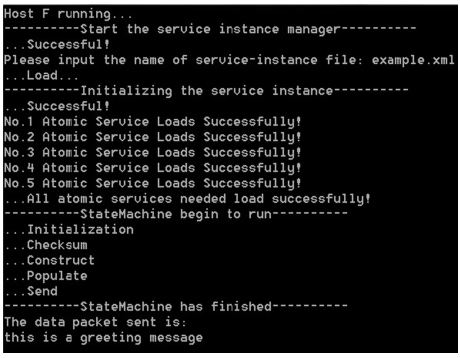


Fig. 2. Host F successfully sent the data packet constructed. And host R successfully constructed the data packet

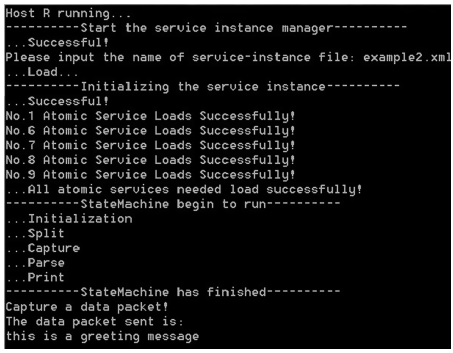


Fig. 3. Host F successfully sent the data packet constructed. And host R successfully constructed the data packet

5.2 The Transmission and Reception of Data Packets Between Service Instance Node and Traditional Network Node

Using WinPcap technology can directly send the original data packet to the NIC device, getting rid of the encapsulation and population of network protocols. Taking advantage of this feature, this experiment sent a complete TCP/IP packet constructed by ConstructDataPacket Service to the traditional network, in order to demonstrate that the service instance node combined by atomic services implemented by the description method above and the traditional network node are interworking.

The ConstructDataPacket Service constructed a complete TCP/IP packet structure according to the fixed TCP/IP packet format. The data packet was sent to the NIC device directly using WinPcap technology. When sent successfully (as Fig. 4), Host R captured the data packet with Wireshark and found the packet's data was correct after compared (as Fig. 5).

As can be clearly seen from Figs. 4 and 5, the IP Source Address of the packet sent is "192.168.1.9", and the IP Destination Address is "192.168.1.24", matching the source address and destination address of the captured packet. The content of data

```
TCP Data Packet has been sent!
Showing on...
The Total Length of Data Packet:17664
MAC Source Address:00:1a:4d:70:a3:89
MAC Destination Address:00:11:22:33:44:55
IP Source Address:192.168.1.9
IP Destination Address:192.168.1.24
TCP Source Port:1000
TCP Destination Port:88
TCP Data:Construce a Tcp Data Packet!
```

Fig. 4. Host F successfully sent the data packet constructed.

packet constructed is "Construct a TCP Data Packet!", matching the content of the captured packet.

The results of the two experiments are combined to demonstrate the correctness and feasibility of the description method of atomic service. And the service instances can indeed provide the corresponding network services through dynamic combination of the atomic services implemented as the description method above.

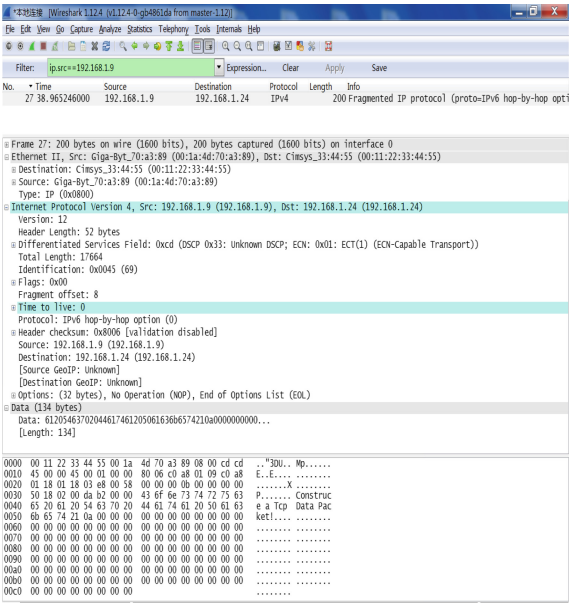


Fig. 5. The format of the packet Host R captured.

6 Summary

With the requirements of network application increasingly diverse and complex, the ability to provide network services has become increasingly demanding. It is more and more evident that the traditional network service model has many drawbacks such as

protocol redundancy, development difficulties, and more difficulty to extend services. The Extensible Network Service Model introduced in the thesis can indeed resolve the conflict. Furthermore, the core of the ENSM is the research on the description method of the atomic services.

In this paper, the definition of atomic service and triplet expression of its conceptual model are given, and the description method of atomic service is described in detail. The data transmission experiment is completed by constructing the best-effort service with atomic service dynamic combination, according to the description method. The results of two experiments demonstrates the correctness and feasibility of the atomic services' description method. And the service instances can indeed provide the corresponding network services through dynamic combination of the atomic services implemented as the description method.

In the future, according to the description method of the combination with information description and function description, the network protocols and network functions will be summarized, classified, and abstracted, then many atomic services will be implemented. So that the ENSM can assist in getting rid of the shackles of hierarchy structure, enhancing the extensibility of network services. The drawback of difficulty in the extension of network services will be resolved from the root, and the ENSM can provide more possibilities for further development of next generation extensible networks.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (Grant No. 61370206).

References

1. Clark, D.D., Wroclawski, J., Sollins, K.R., et al.: Tussle in cyberspace: defining tomorrow's internet. IEEE/ACM Trans. Netw. **13**(3), 462–475 (2005)
2. Lazar, A.A.: Programming telecommunication networks. IEEE Network: Mag. Global Internetworking **11**(5), 8–18 (1997)
3. Braden, R., Faber, T., Handley, M.: From protocol stack to protocol heap: role-based architecture. ACM SIGCOMM Comput. Commun. Rev. **33**(1), 17–22 (2003)
4. Dutta, R., Rouskas, G.N., Baldine, I., et al.: The SILO architecture for services integration, control, and optimization for the future internet. In: IFIP TC5/WG5.3 Forth IFIP/IEEE International Conference on Information Technology for Balanced Automation Systems in Manufacture and Transportation: Advanced Network Enterprises, Virtual Organizations, Balanced Automation, and Systems Integration, pp. 1899–1904. Kluwer, B.V. (2007)
5. Touch, J.D., Pingali, V.K.: The RNA metaprotocol, pp. 1–6 (2008)
6. Zeng, J.Z., Jie, X.U., Yue, W.U., et al.: Service unit based network architecture and its micro-communication element system. Acta Electronica Sin. **32**(5), 745–749 (2004)
7. Quinn, B.: Windows Sockets Network Programming. Addison Wesley, Boston (1996)
8. Wang, Z., Zhang, D.: Research on WinPcap capture IPv6 packet method. In: International Conference on Computer Sciences and Applications, pp. 94–97 (2013)

The Risk Assessment for Unmanned Vehicle Using Bayesian Network

Dapeng Li(✉), Ting Liu, Tingting Cao, Pingke Deng, Ling-chuan Zeng, and Yi Qu

Academy of Opto-Electronics, Beijing, China
lidapeng@aoe.ac.cn

Abstract. The unmanned vehicle shows great potential in national economic. The risk level of unmanned vehicle has direct impact on the development of the unmanned transport industry. The Bayes net for the risk of unmanned vehicle are created. The parameters of net are analyzed, and the quantitative computational method for the evaluation is given. A case study on the typical scene is introduced. The simulation proved the feasibility of the method.

Keywords: Transport security · Risk assessment · Bayesian network · Unmanned vehicle

1 Introduction

Nowadays, with the development of driverless techniques, Volvo and Uber created the unmanned taxi which fused the multi-sensors (GNSS, non-GNSS) and integrated the safety systems and self-control models. Will this unmanned car that arrive on its own to pick up passengers and safely drive them and drop them off at their destination? It's a complicated question and some penetrating evaluation work is needed. To avoid the insecurity affairs caused by GNSS failure, the integrity concepts of GNSS were proposed in the past, such as the GPS integrity [1] and the GALILEO integrity [2] etc. However, the integrity concepts designed for navigation system, like GNSS, do not work for the unmanned vehicle (UV). For example, an unmanned car using the multi-sensor navigation goes along the road. When one of the non-GNSS sensors or the control system provides the wrong signal, the accident may be happened, while, the GNSS integrity is OK. Moreover, which parts of the system responsible to the accident is unclear. This brings a big challenge to the application of UV in the future. Although some literatures about the assessment of UV risk are proposed [4–6], they focus on some specific aspects. Therefore, it's necessary to study one method of the assessment of UV which synthesized many factors into one framework.

A quantitative evaluation method based on the Bayes net for the UV risk is introduced in this paper. First, the Bayes net of the UV risk is created by causality. Second, the models to compute the parameters of Bayes net are presented. Subsequently, the joint probability of events about the UV risk and the assessment algorithm are deduced. Third, a typical example is demonstrated and the simulation results are discussed. Finally, a conclusion is drawn.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 200–207, 2017.
DOI: 10.1007/978-981-10-3969-0_23

2 Methodology

2.1 Bayes Net for the Risk Assessment of UV

The Bayes net for evaluation of the UV risk is created as follows (Fig. 1):

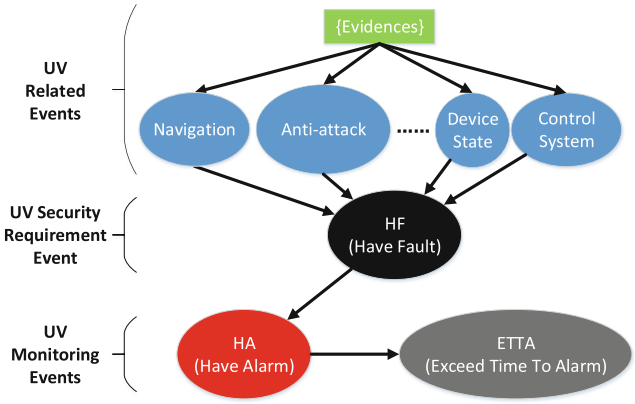


Fig. 1. The directed acyclic graph for the risk assessment of UV.

Observing from up to down, the set named {Evidences} provides the information necessary to compute the prior probability of events connected. Taken as input, the {Evidences} can be acquired from online estimation, manufacturers or other third-parts. Based on the prior probability and the conditional probability deduced by causality, some joint probability of event could be obtained. Finally, the probability of UV risk is computed, which could be used to evaluate the security level of UV.

The causality events are classified as UV Related Events, UV Requirement Event and UV Monitoring Events. The event and type of each node is specified in Table 1.

2.2 Parameters of Events and Risk Assessment

The parameters in the set of {Evidences} for Bayes net could be obtained by online estimation, manufacturers or other third-parts, which could also be used as the qualified tags of security.

- UV Related Events

To simplify, the navigation source errors are projected to some dimensions (e.g. east, up, north and time) and assumed independent. Then, the methodology of over-bounding is applied to describe the error distribution with an over-bounding Gaussian distribution [1]. The navigation errors could be expressed as follows,

Table 1. The causality events

Causality events category	Event name	Descriptions
UV related events	{Evidences}	It's the set of prior information about the UV, such as, the statistical characters of the navigation error, the capability of anti-attack, the reliability of device and the validity of control instruction, etc. In brief, this set provides parameters to describe the probability models of following events
	Navigation (NV)	The NV event is binary. NV = True means the result of navigation wouldn't exceed an allowable level called x Alarm Limit (xAL, x could be some dimension of the space-time information). NV = False is the opposite
	Anti-attack (AA)	The AA event is binary. AA = True means the anti-attack ability is valid, which proves the protection measures satisfy the requirement. AA = False is the opposite
	Device State (DS)	The DS event is binary. DS = True means the UV device is functioning normally, which indicates the device's function and performance satisfy the UV requirement. DS = False is the opposite
	Driver Operation (CS)	The CS event is binary. CS = True means the operation of control system on the UV is correct. CS = False is the opposite
	……	Other events affect the security of UV
UV requirement event	Have Fault (HF)	The HF event is binary. HF = True means something wrong occurs to the UV which do not meet the security requirement. HF = False is the opposite
UV monitoring events	Have Alarm (HA)	The HA event is binary. HA = True means the UV monitoring unit works and gives alarm. HA = False is the opposite
	Exceed Time To Alarm (ETTA)	The ETTA event is binary. ETTA = True means the alarm is not received within the specified TTA (Time To Alarm). ETTA = False is the opposite

$$\begin{cases} \tilde{x}_E = \sum_{i=1}^{Num} w_{E,i}\tilde{x}_{E,i} \\ \tilde{x}_N = \sum_{i=1}^{Num} w_{N,i}\tilde{x}_{N,i} \\ \tilde{x}_U = \sum_{i=1}^{Num} w_{U,i}\tilde{x}_{U,i} \\ \tilde{x}_T = \sum_{i=1}^{Num} w_{T,i}\tilde{x}_{T,i} \end{cases} \tag{1}$$

$$\begin{cases} \tilde{x}_E = \sum_{i=1}^{Num} w_{E,i}\tilde{x}_{E,i} \\ \tilde{x}_N = \sum_{i=1}^{Num} w_{N,i}\tilde{x}_{N,i} \\ \tilde{x}_U = \sum_{i=1}^{Num} w_{U,i}\tilde{x}_{U,i} \\ \tilde{x}_T = \sum_{i=1}^{Num} w_{T,i}\tilde{x}_{T,i} \end{cases} \tag{2}$$

Where $\tilde{x}_{E,i} \sim N(\mu_{E,i}, \sigma_{E,i}^2)$ is the estimation of the i-th navigation source error in the east direction and. $w_{E,i}$ is the i-th weight determined by the specific data fusion method of navigation and $\sum_{i=1}^{N} w_{E,i} = 1$. Because most of the data fusion method of navigation is linear, the Eq. (1) in the form of best linear unbiased estimation (BLUE) is reasonable for most cases. Therefore, $\tilde{x}_E \sim N\left(\mu_E, \sigma_E^2\right)$. The estimations in the north, up and time dimensions are similar. Thus, we have the failure probability of NV event as follows,

$$\begin{aligned} &P_{NV\_failure} = P(NV = F) = P(\tilde{x}_E > EAL) \cup P(\tilde{x}_N > NAL) \cup P(\tilde{x}_U > UAL) \cup P(\tilde{x}_T > TAL) \\ &= 1 - erf\left(\frac{EAL}{\sqrt{2}\cdot\sigma_E}\right) + 1 - erf\left(\frac{NAL}{\sqrt{2}\cdot\sigma_N}\right) + 1 - erf\left(\frac{UAL}{\sqrt{2}\cdot\sigma_U}\right) + 1 - erf\left(\frac{TAL}{\sqrt{2}\cdot\sigma_T}\right) \\ &P(NV = T) = 1 - P(NV = F) \\ &\mu_E = \sum_{i=1}^{Num} w_{E,i}\mu_{E,i},\ \mu_N = \sum_{i=1}^{Num} w_{N,i}\mu_{N,i},\ \mu_U = \sum_{i=1}^{Num} w_{U,i}\mu_{U,i},\ \mu_T = \sum_{i=1}^{Num} w_{T,i}\mu_{T,i} \\ &\sigma_E^2 = \sum_{i=1}^{Num} w_{E,i}^2\sigma_{E,i}^2,\ \sigma_N^2 = \sum_{i=1}^{Num} w_{N,i}^2\sigma_{N,i}^2,\ \sigma_U^2 = \sum_{i=1}^{Num} w_{U,i}^2\sigma_{U,i}^2,\ \sigma_T^2 = \sum_{i=1}^{Num} w_{T,i}^2\sigma_{T,i}^2 \end{aligned} \tag{3}$$

Where Num is the number of available navigation sources, F = False, T = True, *xAL* is the alarm limit in the x-dimension determined by the UV requirement, $erf(u) = \frac{2}{\sqrt{\pi}}\int_0^u e^{-x^2}dx$.

Parameters of other events, such as AA, DS, CS, are provided by manufacturers, other third-parts, or experience. The probability distributions are follows (Table 2),

Table 2. The probability of events in the UV related events

Event	$P(NV)$	$P(AA)$	……	$P(DS)$	$P(CS)$
F	$P_{NV\_failure}$	$P_{AA\_failure}$	……	$P_{DS\_failure}$	$P_{CS\_failure}$
T	1-$P_{NV\_failure}$	1-$P_{AA\_failure}$	……	1-$P_{DS\_failure}$	1-$P_{CS\_failure}$

Note that the $P(AA)$, …, $P(DS)$, $P(CS)$ could be some functions of environment.

- UV Requirement Event

The main problem of UV risk evaluation is how to determine what can be considered unsafe. This depends on the requirements of different standards of transportation. Therefore, we define the UV failure event (Have fault, HF) happened as

$$P(HF = T) = P(NV = F) \cup P(AA = F) \cup P(\ldots = F) \cup P(DS = F) \cup P(CS = F) \tag{4}$$

The conditional probability can be obtained in Table 3.

Table 3. Conditional probability table of $\mathrm{P}(HF|NV, AA, \ldots, DS, CS)$

$P(NV)$	$P(AA)$	……	$P(DS)$	$P(CS)$	HF = F	HF = T
F	F	……	F	F	0	1
T	F	……	F	F	0	1
……	……	……	……	……	0	1
T	T	……	T	T	1	0

Assuming the events are independent, we have

$$\begin{aligned}
\mathrm{P}(HF = F) &= \mathrm{P}(NV = T, AA = T, \ldots, DS = T, CS = T) \\
&= \mathrm{P}(NV = T)\mathrm{P}(AA = T)\ldots\mathrm{P}(DS = T)\mathrm{P}(CS = T) \\
&= (1 - P_{NV_failure})(1 - P_{AA_failure})\ldots(1 - P_{DS_failure})(1 - P_{CS_failure}) \\
\mathrm{P}(HF = T) &= 1 - \mathrm{P}(HF = F)
\end{aligned} \tag{5}$$

- UV Monitoring Event

The UV monitoring unit protects the UV against the failure. The probability of False Alarm (FA), Missing Detection (MD) are used to define the performance of the UV monitoring. The conditional probability of event (Have alarm, HA) can be obtained in Table 4.

Table 4. Conditional probability table of $\mathrm{P}(HA|HF)$

HF	HA = F	HA = T
F	1-P_{FA}	P_{FA}
T	P_{MD}	1-P_{MD}

The alarm must be received within a given period of time and with a given probability. The probability of the event that delivering alarm exceed TTA (Exceed Time To Alarm, ETTA) is defined as P_{ETTA}. The conditional probability can be obtained in Table 5.

Table 5. Conditional probability table of P($ETTA|HA$)

HA	ETTA = F	ETTA = T
F	1	0
T	1-P_{ETTA}	P_{ETTA}

- UV Risk Assessment

  The joint distribution is

$$\begin{aligned} \mathrm{P}(HF, HA, ETTA) &= \mathrm{P}(HF)\mathrm{P}(HA|HF)\mathrm{P}(ETTA|HA, HF) \\ &= \mathrm{P}(HF)\mathrm{P}(HA|HF)\mathrm{P}(ETTA|HA) \end{aligned} \tag{6}$$

  The UV risk probability is

$$P_{Risk} = \mathrm{P}(HF = T, HA = F, ETTA = F) \cup \mathrm{P}(HF = T, HA = T, ETTA = T) \tag{7}$$

  By comparing with the failure rate levels defined in the requirement of UV, the P_{Risk} could be used to assess the security of the UV.

3 Simulation

A typical dynamic scene is used to demonstrate the proposed method. A car equipped with GNSS/INS going through a city canyon is simulated as follows (Fig. 2).

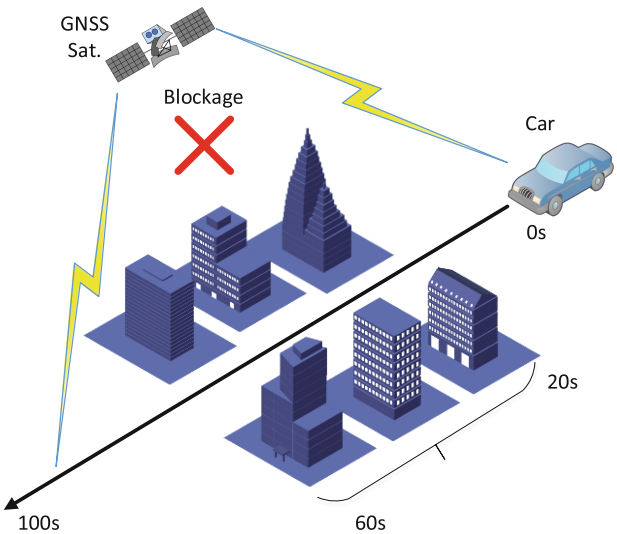


Fig. 2. The dynamic simulation scene of city canyon.

The simulation time is 100 s totally. The blockage of the GNSS navigation signals begins at 20 s and end at 60 s. Assuming no attack (no jamming or hacker), no device failure and no wrong control operation, the prior probability $P_{AA\_failure} = 0$, $P_{DS\_failure} = 0, P_{CS\_failure} = 0$. Set $P_{FA} = 1.0e-8, P_{MD} = 5e-9, P_{ETTA} = 4.6e-11$. The poisoning errors of GNSS and INS follow $N(0, 10)$ and $N(0, 2.5)$ respectively, and the timing error follows $N(0, 1)$. The number of Visible Satellite (VS) is 10 initially. The alarm limit EAL = 5, NAL = 5, UAL = 5, TAL = 1. To simplify, the weights of the estimation in the data fusion of navigation are calculated by the BLUE,

$$w_{GNSS} = \frac{\sigma_{INS}^2}{\tilde{\sigma}_{GNSS}^2 + \sigma_{INS}^2}, w_{INS} = \frac{\tilde{\sigma}_{GNSS}^2}{\tilde{\sigma}_{GNSS}^2 + \sigma_{INS}^2} \tag{8}$$

Where $\tilde{\sigma}_{GNSS}^2$ is the online estimation of GNSS error variance and σ_{INS}^2 is the INS error variance provided by manufacturer. The simulation results are shown in Fig. 3.

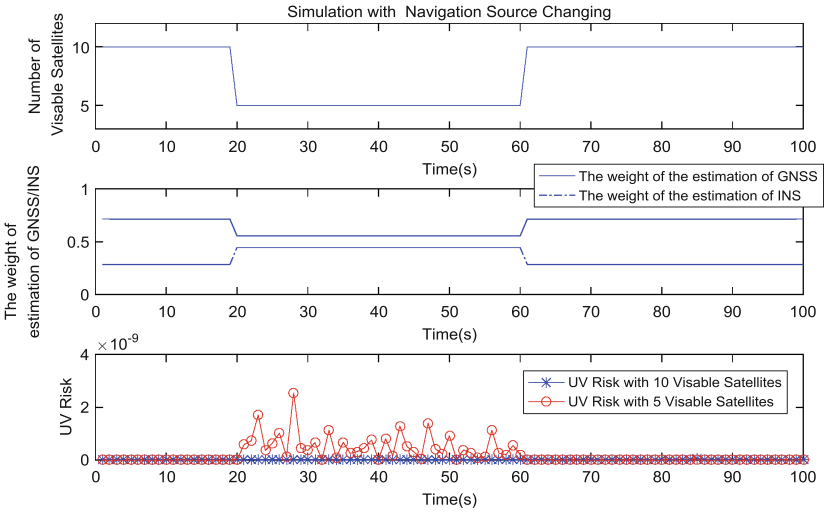


Fig. 3. The simulation results.

As can be seen, during the GNSS signal blockage, the number of VS is decreased from 10 to 5, and the estimation variance has doubled for the loss of navigation source. The weights are adjusted adaptively according to the BLUE in the data fusion step. The UV risk rises up when the navigation sources reduced. It could provide some dynamic simulation results to the UV designer in the aim to getting a deep view of the UV system and finding the key factors of UV security.

4 Conclusion

The article proposed a flexibility framework synthesized most factors of UV to assess the risk level. According to the characteristics of the UV, we design a Bayesian network for risk evaluation and analyze the network parameters. Use the Bayesian network theory to evaluate the risk level of UV. An example is given, and the simulation results proved the feasibility of the method. More factors or events related to the UV security would be considered in further study.

References

1. Binjammaz, T., et al.: GPS integrity monitoring for an intelligent transport system. In: 10th Positioning Navigation and Communication (WPNC 2013), pp. 1–6 (2013)
2. Oehler, V.: The galileo integrity concept. In: 17th International Technical Meeting of the ION GNSS (2004)
3. Roturier, B.: The SBAS Integrity Concept Standardised by ICAO. Application to EGNOS
4. Aalmoes, R.: A conceptual third party risk model for personal and unmanned aerial vehicles. In: 2015 International Conference Unmanned Aircraft Systems (ICUAS) (2015)
5. Wahlström, J.: Risk assessment of vehicle cornering events in GNSS data driven insurance telematics. In: 2014 IEEE 17th International Conference Intelligent Transportation Systems (ITSC) (2014)
6. Knight, J.: An essay on unmanned aerial systems insurance and risk assessment. In: 2014 IEEE/ASME 10th International Conference Mechatronic and Embedded Systems and Applications (MESA) (2014)

Delay-Constrained Least-Energy-Consumption Multicast Routing Based on Heuristic Genetic Algorithm in Unreliable Wireless Networks

Ting Lu(✉), Shan Chang(✉), and Guohua Liu

School of Computer Science and Technology,
Donghua University, Shanghai, China
{luting, changshan}@dhu.edu.cn

Abstract. Delay-constrained least-energy-consumption multicast tree construction is an important problem in wireless ad hoc networks and sensor networks to support multimedia applications such as audio and video. In the past few years, delay-constrained least-cost multicast tree construction had received much attention. However, these algorithms in wired networks cannot be directly used in wireless networks, because energy consumption are not considered in protocol design. In this paper, we focus on the problem of delay-constrained least-energy-consumption multicast routing in unreliable wireless multi-hop networks. Link error rate is considered in the process of multicast tree construction. We proposed a heuristic genetic algorithm to solve the problem. Simulations are performed to demonstrate the effectiveness and efficiency of the proposed algorithm.

Keywords: Link error rate · Energy · Delay · Wireless multi-hop networks · Genetic algorithm

1 Introduction

Ad hoc networks and sensor networks are important wireless multi-hop networks because they have a wide range of potential applications in military and civil areas. Multicasting is one of key services in resource-constrained wireless networks to support multimedia applications such as audio and video. Through multicasting, source node can send the same information to a group of destinations concurrently in an efficient way. To support delay sensitive applications, wireless multi-hop networks must ensure that end-to-end delay is smaller than the applications' requirement. Since nodes in wireless multi-hop networks are usually powered by batteries with limited capacity, routing protocols must minimize the total energy consumption to prolong network lifetime. Thus, delay-constrained least-energy-consumption multicast routing is an important problem in wireless multi-hop networks.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 208–217, 2017.
DOI: 10.1007/978-981-10-3969-0_24

In the past few years, delay-constrained least-cost multicast routing problem has received much attentions in wired networks [1–3]. However, the proposed protocols for the problem cannot be used in wireless multi-hop networks directly, because the energy consumption are not considered in protocol design. We had earlier studied the problem of delay-constrained energy-efficient multicast routing in wireless multi-hop networks [4]. However, we didn't take link unreliability into consideration and simply assumed that the link error probability is 0. The problem of constructing delay-constrained least-energy-consumption multicast tree is known to be NP-hard. Conventional algorithms is difficult to solve this problem.

Genetic algorithm (GA) is one class of evolutionary algorithm (EA), which mimics the process of natural selection. GA uses techniques inspired by natural evolution, such as inheritance, crossover, mutation and selection. GA can generate optimal solutions for complex problems which are very difficult to solve by conventional techniques. In this paper, we focus on the delay-constrained least-energy-consumption multicasting (DCLECM) problem in unreliable wireless multi-hop networks. We proposed a heuristic GA for the DCLECM problem taking the link error rate into consideration.

The rest of this paper is organized as follows. In Sect. 2, we give the network model and describe the problem formulation. Then, the proposed GA is proposed in Sect. 3. The performance of the proposed algorithm is evaluated in Sect. 4. Finally, the conclusion is given in Sect. 5.

2 Network Model and Problem Formulation

2.1 System Model

A network is modeled as a graph $G(V, E)$, where V is a set of network nodes and E is a set of links. Link $(i, j) \in E$ means that node j is within the transmission range of node i. Each link $(i, j) \in E$ is associated with a transmission delay $delay_{i,j}$, an energy cost $E_{i,j}$, and a packet error probability $error_{i,j}$. The minimum energy needed to transmit a unit packet from node i to node j is $E_{i,j} = k_1(dis_{i,j})^{\alpha} + k_2$. $dis_{i,j}$ is the distance between node i and node j. α is the path loss exponent. Typically, α is 2 for short distance and 4 for larger distance k_1 and k_2 are two constants. k_1 depends on the property of the antenna. k_2 is the overheads of electronics and digital processing which is independent of distance In order to ensure reliable data transmission on wireless links, acknowledgement scheme is required. As described in [5], there are two schemes for packet acknowledgement: (1) positive acknowledgement (ACK) based; (2) negative acknowledgement (NAK) based. In ACK-based scheme, if child node i receives the transmitted packet successfully from parent node, node i sends an acknowledgement to parent node. Because all the child nodes must send an acknowledgement to parent node for each transmitted packet, this would lead to the waste of bandwidth and collision of acknowledgement packets. In addition, parent node has to record which nodes have not received the transmitted packet successfully. In NAK-based scheme, only the nodes

which have not receive the packet need to send negative acknowledgement to parent node. If one of child nodes doesn't receive the transmitted packet successfully, parent node will retransmit the packet. Parent node doesn't need to record the set of nodes which have not received the packet successfully. If parent node receives a negative acknowledgement, it uses the transmission power which is sufficient to reach the farthest child node to retransmit the data packet. Obviously, NAK-based scheme is more efficient than ACK-based scheme in terms of resource utilization. In this paper, NAK-based scheme is used.

The retransmission number for a reliable packet transmission from node i to node j is $num_{i,j} = \frac{1}{1-error_{i,j}}$. The details of modulation schemes are not studied in this paper. Thus, we assume that the binary phase-shift keying (BPSK) is employed as the modulation scheme. The bit-error-rate err^b of a wireless link is

$$err^b = 0.5 \cdot erfc(\sqrt{\frac{P_r}{noise \cdot f}}), \tag{1}$$

where $erfc(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$, P_r is the received power level, *noise* is the noise spectral density (noise power per Hz), and f is the transmission bit-rate. We assume that the received signal power at node j is inversely proportional to $(dis_{i,j})^{\alpha}$. Thus, the received power at node j can be replaced by $P_t/(dis_{i,j})^{\alpha}$. In this paper, we set $\alpha = 2$. The packet error probability $error_{i,j} = 1 - (1 - err^b)^{len}$, where *len* is the packet length.

Let $s \in V$ denote the multicast source node, $D \subseteq V - \{s\}$ denote a set of multicast destinations, and $T(s, D)$ denote the multicast tree rooted at s and spanning all nodes in D. The delay of a reliable packet transmission from multicast source s to a destination node $t \in D$, denoted as $delay(p_T(s,t))$, is computed by

$$delay(p_T(s,t)) = \sum\nolimits_{(i,j) \in p_T(s,t)} delay_{i,j} \cdot num_{i,j}, \tag{2}$$

where $p_T(s,t)$ is the path from source node s to the destination t along the multicast tree T, and $delay_{i,j}$ is the transmission delay for a unit packet on link (i,j).

Given a multicast tree T, let N_i denote a set of children nodes of node $i \in T$ (node i is not leaf node) and $N_i(1)$ be the farthest children node of node i. We compute the energy cost on node i to transmit a unit packet successfully from node i to a set of children nodes N_i by Algorithm 1. In Algorithm 1, $|N_i|$ is the number of nodes in set N_i. $N_i(j).not\_rcv\_flag$ represents the probability that node j has not received the packet transmitted from node i correctly.

Algorithm 1

Input: node i, a set of children nodes N_i

Output: energy cost $cost_i$ on node i

Let $N_i(1)$ be the farthest children node of node i;

$cost_i = 0$;

for $j \leftarrow 1$ to $|N_i|$

$N_i(j).not\_rcv\_flag = 1$;

$num_{i,j} = \frac{1}{1-error_{i,j}}$;

end for

for $j \leftarrow 1$ to $|N_i|$

$cost_i = cost_i + E_{i,N_i(1)} \cdot num_{i,j} \cdot N_i(j).not\_rcv\_flag$;

for $k \leftarrow j+1$ to $|N_i|$

$N_i(k).not\_rcv\_flag = N_i(k).not\_rcv\_flag \cdot (error_{i,k})^{num_{i,j}}$;

end for

end for

return $cost_i$

We define the energy cost for a unit packet reliable transmission along a multicast tree T as

$$cost(T) = \sum_{i \in T} cost_i. \tag{3}$$

Note that the energy cost on leaf node is 0.

2.2 Problem Definition

Given a network $G(V, E)$, a source node $s \in V$, a set of destinations $D \subseteq V - \{s\}$, and a positive delay constraint δ, the delay-constrained least-energy-cost multicast tree is to find the minimum energy cost multicast tree $T(s, D)$ such that $delay(p_T(s, t)) \leq \delta$, $\forall t \in D$.

3 Algorithm Design

3.1 Theory of the Genetic Algorithm

The proposed GA is shown in Algorithm 2. In Algorithm 2, N_p is the population size, N_g is the iteration times which is set by the system.

Algorithm 2
Input: node i, a set of children nodes N_i Output: energy cost $cost_i$ on node i
Step.1 selecting coding scheme for multicast tree.
Step.2 population initialization: generate N_p chromosomes.
Step.3 evaluate the chromosomes with fitness function.
Step.4 select two chromosomes as the parents for crossover operation.
Step.5 crossover the two selected chromosomes to generate an offspring.
Step.6 generate a random number $0 < \tau < 1$. If $\tau \leq pro_m$, mutate the offspring generated in Step 5.
Step.7 repeat Step 4 to Step 6 until N_p offspring are produced.
Step.8 select the best N_p chromosomes from the new generated offspring and current population, and copy them into the next generation.
Step.9 repeat Step 3 to Step 8 N_g times.

3.2 Coding

How to encode a multicast tree into a chromosome is one of important steps for efficient GA design. "Good" encoding scheme have low complexity of encoding/decoding operations between multicast trees and chromosomes. Researchers [6] indicated that using the multicast tree itself as the chromosome is an efficient encoding scheme, because encoding/decoding operations are omitted. In this paper, we use the multicast tree as chromosome directly as shown in Fig. 1.

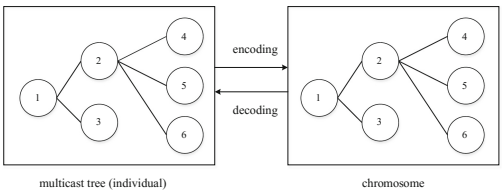


Fig. 1. Encoding scheme.

3.3 Initial Population

The random depth-first search (RDFS) algorithm [7] is used to generate the initial population. First, RDFS algorithm begins at the source node *s* and marked the source

node as currently visited node. In each step, RDFS randomly selects an unvisited adjacent node to currently visited node. The selected node is denoted as currently visited node. RDFS terminates until all the destination nodes have been visited. RDFS are employed N_p times to generate N_p multicast trees. N_p is the population size.

3.4 Fitness Function

Fitness function is used to evaluate the chromosome performance. "Good chromosome" (delay is bounded and energy cost is low) has bigger fitness value than "bad chromosome". Thus, fitness function is defined as follows:

$$f(T) = \frac{\omega}{cost(T)} \cdot \prod_{t \in D} \Phi(delay(p_T(s,t)) - \delta) \tag{4}$$

where ω is a positive real constant, and $\Phi(\cdot)$ is a penalty function as follows:

$$\Phi(Z) = \begin{cases} 1, Z \leq 0 \\ \eta, Z > 0 \end{cases} \tag{5}$$

$0 < \eta < 1$ indicates the penalty degree. In our experiments, we select $\eta = 0.5$.

3.5 Selection

Two parent chromosomes used for crossover are selected by wheel selection scheme. The selection probability for a parent T_i is

$$p(T_i) = \frac{f(T_i)}{\sum_{j=1}^{N_p} f(T_j)} \tag{6}$$

3.6 Crossover Scheme

Crossover is an important genetic operator which operates on two chromosomes at a time and generates an offspring by combining both chromosomes' features. To a great extent, the performance of GA depends on the performance of the crossover operator. We proposed an improved crossover operator [5] to generate new chromosome. According to the roulette wheel selection scheme [8], two chromosomes are selected as the parents to generate an offspring. The selection probability is

$$pro(T_i) = \frac{f(T_i)}{\sum_{j=1}^{N_p} f(T_j)}. \tag{7}$$

The bigger the fitness value, the greater the probability of being selected. The chromosome with bigger fitness value has more "good" feature of the optimal solution. Crossover operator retains the common links between two parents into offspring. The

common links are more likely to represent the "good" feature of the optimal solution. The common links, the source node and the destination nodes are in some separate sub-trees. Then, crossover operator connects these sub-trees into a multicast tree as follows. First, we define the path quality from node i to node j as

$$pathqua(i,j) = \sum\nolimits_{(k,l)\in p(i,j)} \frac{1}{\varepsilon \cdot \frac{dis_{k,l}}{\max\{dis_{k,l}\}} + (1-\varepsilon)\cdot \frac{error_{k,l}}{\max\{error_{k,l}\}}}, \tag{8}$$

Where $p(i,j)$ is a path from node i to node j, and $0<\varepsilon<1$ is a tunable constant. Crossover operator randomly selects two sub-trees and connects them with the least-delay path or the path which has the largest path quality. If one parent chromosome satisfies the delay constraint, the sub-trees are connected by the largest path quality path. Otherwise, the sub-trees are connected by the least delay path. The new generated sub-tree is among the separate sub-trees for the next selection. Crossover operator repeats the sub-trees connection process when a new multicast tree is constructed.

3.7 Mutation Scheme

Mutation is an important genetic operator which produces spontaneous random changes in various chromosomes. In GAs, mutation serves the crucial role of either (1) replacing the lost genes (lost feature of optimal solution) during selection process or (2) providing the genes that were not present in the initial population. When a new offspring is generated by crossover operator, mutation is performed according to the mutation probability pro_{mu}. Mutation operator randomly selects a sub-set of nodes in the new generated offspring and removes the links that are incident into the selected nodes. Thus, some separate sub-trees are generated. Mutation operator re-connects the sub-trees with the least-delay path or the largest path quality path. The re-connection process is the same as described in crossover scheme.

4 Performance Evaluation

4.1 Analysis of Convergence

The proposed GA has the following characteristics: (1) crossover probability is 1; (2) mutation probability $0<pro_{mu}<1$; (3) selection model is elitist scheme. From the Theorem 2.7 in [9], we can conclude that the proposed GA can converge to the global optimal multicast tree.

4.2 Simulation Setting

The proposed GA is implemented in MS VC++ on personal computer with i5 Dual-core 2.8 GHz CPU. For each network instance, the source node and the destination nodes are randomly selected. We take 200 separate runs to get the average value

of each result. The distance and delay of each link and are uniformly distributed in $[10, 200]$ and $[0, 50]$, respectively. The delay constraint δ is uniformly distributed in $[70, 700]$. the noise of each link is uniformly distributed in [0, 0.5]. The transmission power of each node is 20 mW. The transmission bit rate f is 8 bits/s. Packet length len is 8 bits. Mutation probability pro_m is 0.05. Population size N_p and iteration times N_g are 15 and 10, respectively.

4.3 Results

In this section, we verify the convergence ability and convergence speed of the proposed GA. Three metrics are used:

(1) Normalized success ratio (NSR): If the multicast tree constructed by the algorithm satisfies the delay constraint, the routing request is considered as successful one. Success ratio $SR = \frac{num_{suc}}{num_{total}}$, where num_{suc} is the number of successful routing requests and num_{total} is the total number of routing requests. NSR is defined as SR_1/SR_{ref}, where SR_1 and SR_{given} are the NSR of Algorithm 1 and the reference algorithm respectively.
(2) Normalized energy cost (NEC): NEC is defined as EC_1/EC_{ref}, where EC_1 and EC_{ref} are the energy cost of Algorithm 1 and the reference algorithm respectively.
(3) Running time. SR and NEC verify the convergence ability while running time verify the convergence speed. The proposed GA is compared with energy-efficient QoS GA (EEQGA) [4] and least-delay multicast tree (LDT). If all wireless links are reliable, LDT has the highest SR because the source node and the destinations are connected by the least delay path. Thus, LDT is selected as the reference algorithm.

The NSR of the proposed GA is compared with that of EEQGA in Fig. 2. From Fig. 2, we can see that the NSR of proposed GA is bigger that of EEQGA. This is because the proposed GA takes link reliability into consideration, which results in high NSR. In Fig. 2, NSR of EEQGA is 1. This is because EEQGA has the same SR of LDT, if link reliability is not considered.

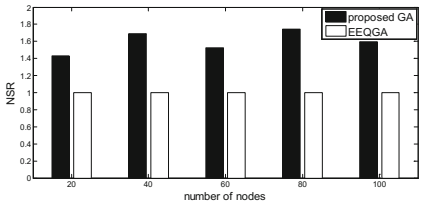


Fig. 2. Comparison of NSR.

The comparison of NEC is shown in Fig. 3. From Fig. 3, we can see that the NEC of the proposed GA is much smaller than that of EEQGA. This is because that the proposed GA takes the link reliability into consideration, while EEQGA and LDT

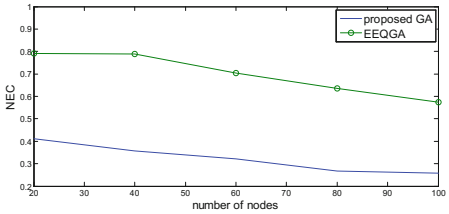


Fig. 3. Comparison of NEC.

suggested that all the links are reliable. In Fig. 3, the NEC of EEQGA is smaller than 1. This is because that EEQGA considers the energy cost for transmission in routing decision, while LDT selects route only based on transmission delay.

The comparison of running time is shown in Fig. 4. From Fig. 4, we can see that the running time of EEQGA and the proposed GA is much smaller than that of LDT. In addition, the running time of the LDT increases slowly with the network size, while the running time grows significantly with the network size.

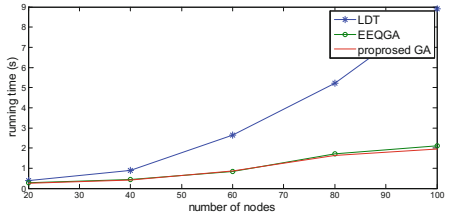


Fig. 4. Comparison of NSR.

5 Conclusion

In this paper, we proposed a heuristic GA for delay-constrained least-energy-consumption multicast routing in unreliable wireless multi-hop networks. The proposed GA considers the link error rate in the process of multicast tree construction. Simulation results demonstrated that the proposed GA can significantly improve the success ratio and energy cost. In addition, the running time of the proposed GA is fairly desirable. The work in this paper only focus on source-based multicast tree. Our future work will study the distributed implementation for the proposed algorithm.

Acknowledgments. This work is supported by National Natural Science Foundation of China (Grant No. 61402101, 61300199), Shanghai Municipal Natural Science Foundation (Grant No. 14ZR1400900), Fundamental Research Funds for the Central Universities (Grant No. 2232014D3-42, 2232014D3-21).

References

1. Forsati, R., Haghighat, A.T., Mahdavi, M.: Harmony search based algorithms for bandwidth-delay-constrained least-cost multicast routing. Comput. Commun. **31**(10), 2505–2519 (2008)
2. Xue, G.L., Zhang, W.Y., Tang, J., Thulasiraman, K.: Polynomial time approximation algorithms for multi-constrained QoS routing. IEEE/ACM Trans. Netw. **16**(3), 656–669 (2008)
3. Zhang, L., Cai, L.B., Li, M., Wang, F.H.: A method for least-cost QoS multicast routing based on genetic simulated annealing algorithm. Comput. Commun. **32**(1), 105–110 (2009)
4. Lu, T., Zhu, J.: Genetic algorithm for energy-efficient QoS multicast routing. IEEE Commun. Lett. **17**(1), 31–34 (2012)
5. Banerjee, S., Misra, A., Yeo, J., Agrawala, A.: Energy-efficient broadcast and multicast trees for reliable wireless communications. In: Proceedings of the IEEE WCNC (2003)
6. Wang, Z., Shi, B., Zhao, E.: Bandwidth-delay-constrained least-cost multicast routing based on heuristic genetic algorithm. Comput. Commun. **24**(7–8), 685–692 (2001)
7. Wikipedia, Depth-first search. https://en.wikipedia.org/wiki/Depth-first_search
8. Lipowski, A., Lipowska, D.: Roulette-wheel selection via stochastic acceptance. Physica A **391**(6), 2193–2196 (2012)
9. Guoliang, C., Xufa, W., Zhenquan, Z., Dongsheng, W.: Genetic Algorithm and Its Application. People's Posts and Telecommunications Press, Beijing (1996)

A Coarse to Fine Object Proposal Framework for Autonomous Driving Object Detection Using Binocular Image

Xiaolong Liu(✉), Wanzeng Cai, Zhengfa Liang, and Yiliu Feng

College of Computer, National University of Defense Technology, Changsha, China
medivhliu@163.com, leonzfa@163.com, zengzeng2016@sina.com, 329436764@qq.com

Abstract. The now widely used object proposal methods for object detection commonly get fulfilling results on the dataset, which is captured in simple scenes. But the performance degraded when it comes to complicate real traffic scene. In our paper, a coarse to fine object proposal generating framework is proposed for autonomous driving object detection, provides a better object proposal solution in complex circumstances. By adding several low level geometrical features, which can be efficiently computed from binocular images, we recalculate scores for the candidate bounding boxes generated by coarse region proposal approaches with a Bayesian probability model. Our proposal generation approach is validated on the challenging KITTI benchmark, achieving state-of-art object proposal performance for pedestrian, car and cyclist.

Keywords: Object proposal · Object detection · Stereo vision · Bayesian probability model · Coarse to fine framework

1 Introduction

Autonomous driving is attaching more and more attention from both industry and the research community. With the expensive sensors, such as LIDIR, radar and high-precision GPS, self-driving cars are already ready to go. However, a low-cost and more intelligent self-driving system is urgent needed. Autonomous vehicle with cameras is a competitive alternative solution. This paper aims at refining the result computed from coarse approaches with several low-level geometrical features, and finally generating high precision object proposal for object detection in the scene of autonomous driving exploiting binocular image.

Traditional object detection approaches use sliding window paradigm to exhaustive search the potential bounding boxes, and then a classifier followed, such as SVM, random forest. However, the sliding window algorithm produces 10^6–10^7 boxes for each VGA resolution image to make sure different scales of objects are included. Millions of bounding boxes make the computation overloaded; a few seconds are need per image to finish the detecting procedure. For reducing computation, plenty of region proposal methods have been put forward, such as Edge Boxes, Selective Search, Bing,

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 218–227, 2017.
DOI: 10.1007/978-981-10-3969-0_25

etc. These measures only produce thousands of bounding boxes to accelerate the classify process, but the unsatisfied recall of bounding boxes makes it hard to achieve a high accuracy detection result, especially in the autonomous driving benchmark KITTI [1], which images are captured in complex street scene.

The development of an effective and computationally efficient region proposal mechanism is still an open problem. Through observing the intermediate result of Edge Boxes [2] and Selective Search [3] on KITTI benchmark, we found that before Non-maximum suppression and final cut-off, the bounding boxes almost covered all desired objects in image, the score degrades recall. In this paper, we take advantage of several low level geometrical features which can be efficiently computed from the binocular images to fine tune the score generated by those coarse region proposal measures. Using Bayesian probability model, a coarse to fine object proposal framework for object detection under complicated street scene is proposed.

By adding the features of height to road, aspect ratio and the actual area of bounding box, we put the original score together to create a Bayesian probability model, calculating the probability of very bounding box containing an interested object to recalculate scores for candidates. We evaluate our approach on the challenging KITTI detection benchmark. Extensive experiments show that the final proposed object proposals achieve state-of-art recall across all categories under various occlusion and truncation levels.

2 Related Work

The goal of generating object proposals is to create a relatively small set of candidate bounding boxes that cover the interested objects in the image. Most detectors use complex and expensive features to classify bounding boxes, so the fewer number and higher proposal recall is the critical path to reduce computation and achieve better accuracy. A lot of related literatures have been published during the past years. Most object proposal approaches can be classified into two categories, one depend on grouping and the other one based on winding scoring.

Selective Search [3] is one of the most well-known grouping proposal methods. It is broadly used in the top performance object detection algorithms, including R-CNN [4] and Fast R-CNN [5]. Selective Search elaborately combined the strength of both exhaustive search and segmentation. Given an input image, it firstly segments image into pieces of small regions through superpixel segmentation [6]. Then different color spaces are utilized to calculate the color similarity, texture similarity, size similarity and fit similarity between two adjacent segmentations. In the end, four kinds of similarity are linearly combined to measure the similarity degree between pieces, and highest similarity degree regions are greedily merged to generate candidate bounding boxes. Other grouping proposal approaches utilize diverse strategies to generate candidate boxes. For example, Chang [7] combines saliency and Objectness [8] with a graphical model to merge superpixels into figure segmentation; MCG [9] introduces a fast algorithm for computing multi-scale hierarchical segmentations.

Edge Boxes [2] is an effective and efficient window scoring region proposal method. It first computes the edges of a whole image, and make them into edges

groups. By comparing the similarity of the edge groups, every edge group is weighted. Using weighted edge groups, Zitnick gives scores to all the candidate regions finally. There are also amount of other window scoring proposal generating algorithms. For instance, the earliest well know proposal method Objectness [8] uses multiple cues including color, edges, location, size, and superpixel straddling to score salient locations in image; Bing [10] trains a simple linear classifier over edge features and applied in a sliding window manner.

3DOP [11] is also a window scoring region proposal method which uses stereo point cloud to generate 3D bounding boxes. Employing multiple point cloud features, it creates candidate proposals by minimizing an energy function.

More recently, the region proposal network (RPN) [12] addresses object proposals on convolution feature maps, leading to a significant speedup for proposal detection. By sliding a 3×3 window on the last convolution layer of VGG16 [13], RPN uses different sizes and different aspect ratios anchors to generate proposals. However, the receptive field of the 3×3 window and the several times pooling make it missed most of the small scale object.

3 Coarse to Fine Proposal Generating Framework

For the purpose of autonomous driving object detection, a coarse to fine object proposal generating framework has been put forward. We are aiming at generating high recall bounding boxes on pedestrian, vehicles and cyclists, which are playing main roles in daily traffic. As shown in Fig. 1 our coarse to fine object proposal generating framework consists of two components. In first step, we use coarse object proposal algorithms (Edge Boxes and Selective Search are chosen in our experiment) to produce a plenty of candidate boxes, Non-maximum suppression and cut-off are not conducted in this stage. In following step, the Bayesian probability model helps us re rating the candidates with original score in the previous stage and three more geometrical

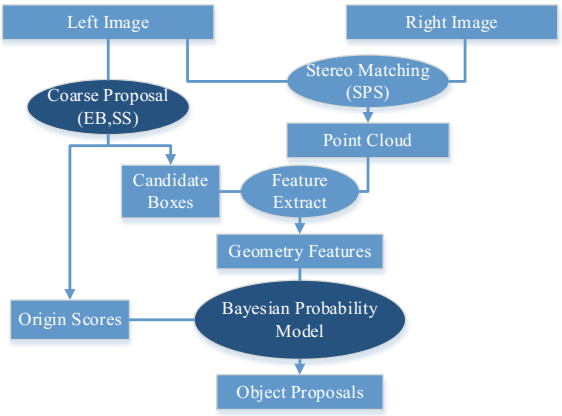


Fig. 1. Coarse to fine object proposal generating framework

features. In training stage, priori probability is calculated. In testing stage, the probability whether an interested object is contained can be easily estimated by Bayes formula:

$$P(obj|F) = \frac{P(F|obj)P(obj)}{P(F)} \tag{1}$$

where F is the value set of the features which will be introduce in detail in next Sec. IV. $P(obj)$ is the prior probability and $P(F|obj)$ is the conditional probability which are calculated in training stage.

4 Features

To reevaluate the score made by coarse object proposal methods, three low level geometrical features are used together with the original score. The three geometrical features are all computed from binocular images.

A. The Height Over the Ground Considering the situation of driverless vehicles, pedestrians, cars and cyclists are all standing against the ground. The average height over the ground of each bounding box is an important cue to distinguish those boxes containing a standing object from those not. Using binocular images, we can model the 3D point cloud *XYZ* through stereo matching approaches, such as SPS-St [14], PBCP [15], Displets [16], etc. SPS-St is chosen in our experiment for convenient and efficient. Then, we use RANSIC [17] to figure out the approximate ground plane equation. Every pixel in the left view is mapped to point cloud *XYZ*, so the height over ground plane is easily calculated by the distance formula of point to plane. We directly calculate the height map of all pixels in image and make it an integral channel denoted by H for reducing duplicate computation. As a result, the average height over the ground of every bounding box can be described as:

$$Ht = H(x+w, y+h) - H(x, y) \tag{2}$$

where (x, y, w, h) indicates the location of a bounding box in image. (x, y) is the left top corner coordinate value, and (w, h) represents width and height of the bounding box.

B. Actual Area of Bounding Box The objects normally have multifarious scales in 2D RGB images because of different distance to camera sensor, but when changed to stereo images, it is much easier to solve the multiscale problem. The area of an object is fixed in real world, and the area of similar objects varies in a limit scope. Directed by the pinhole imaging principle, the actual area of an object is proportional to the square of the distance. In point cloud *XYZ*, camera sensor is selected as origin of coordinates, where coordinate indicated (0,0,0). Taking advantage of the distance formula of point to point, the real distance between camera sensor and each pixel is calculated. We choose the distance between camera sensor and the center pixel of every bounding box as an estimate of the bounding box distance to camera sensor, the actual area of bounding box is evaluated by:

$$Ad = w \times h \times \sqrt{X^2(x_c, y_c) + Y^2(x_c, y_c) + Z^2(x_c, y_c)} \tag{3}$$

where $(x_c,\ y_c)$ is the center of bounding box, which is computed from $(x + w/2,\ y + h/2)$.

C. Aspect Ratio The objects needed to be detected always get a fixed aspect ratio range, it remains unchanged when captured in image. Statistically, as illustrated in Fig. 2-a, the aspect ratio of pedestrian varies in a narrow range from 0.2 to 0.8, and mostly of the aspect ratios focus on a tighter range around 0.4. Figure 2-b shows that cars have much broader variation scope, range from 0.5 to 3.5, it mainly caused by different facing directions. In Fig. 2-c, aspect ratio of cyclists varies from 0.5 to 1.5. Altogether, the objects we are interested in have a relatively fixed aspect ratio range. According to aspect ratio distribution, boxes with unsuitable aspect ratio will get lower probability in Bayesian probability model and finally will be cut off. Based on the bounding boxes generated by coarse object proposal method, the aspect ratio is represented as follow:

$$Ar = \frac{w}{h} \tag{4}$$

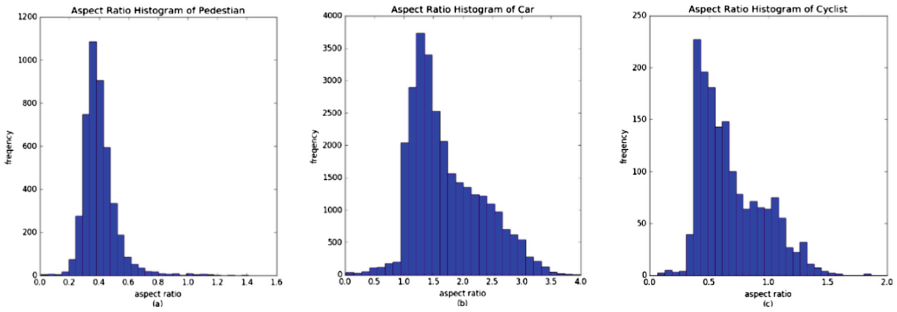


Fig. 2. Aspect ratio histograms of pedestrian, car and cyclist.

5 Experiments Evaluation

In this section, implementing details and experiment result evaluation will be carefully discussed. We evaluate our approach on the challenging KITTI detection benchmark [1], which has 7,481 delicately annotated training and 7,518 unannotated test images. It was established for autonomous driving experiment and evaluation. The benchmark contains three object classes: *Car*, *Pedestrian*, and *Cyclist*. Evaluation is done for each class in three regimes: Easy, Moderate and Hard, which contain objects of different occlusion and truncation levels. We split the 7,481 training images into a training set (3,712 images) and a validation set (3,769 images) as it mentioned in [11]. We ensure that the training and validation set do not contain images from the same video sequences, and evaluate the performance of our proposals on the validation set.

A. Metrics To evaluate proposals, we use the oracle recall as the metric, following [18, 19]. A ground truth object is said to be recalled if at least one proposal overlaps with it with IoU above a certain threshold. We set the IoU threshold to 70% for Car, and 50% for Pedestrian and Cyclist, following the standard KITTI's setup. The oracle recall is then computed as the percentage of recalled ground truth objects. we also report average recall (AR) [19], which has been shown to be highly correlated with the object detection performance.

B. Baseline We compare our proposal method with two most frequently applied approaches in object detection on the validation set: Edge Boxes (EB), which achieves highest recall among window scoring proposal methods, and Selective Search (SS), which achieves best performance amount grouping proposal methods. We also compare our proposal result with 3DOP, the best proposal approach on KITTI in present.

C. Implementing Details We choose Edge Boxes and Selective Search as our first stage coarse proposal generating method. Take advantage of the intermediate result, thousands of candidate boxes were generated with original scores. Compared with the annotated ground truth, the candidate which has a higher IoU than 0.6 is selected as positive sample in training stage, also 600 negative samples are randomly selected in the rest candidates per image. In testing stage, 50000 candidate bounding boxes are chosen per image in coarse proposal generating process, the recall of 50000 boxes is illustrated in Table 1. It can be observed that almost all interested objects are included in 50000 candidates. We select the count of 50000 for both computation and precision concern. More evaluations have been done on Edge Boxes and details will be presented in next Section.

Table 1. Recall of coarse proposal generating process

Method	Category								
	Pedestrian			Car			Cyclist		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
Edge boxes	0.9842	0.9381	0.9192	0.9916	0.9651	0.9455	0.9828	0.8967	0.8914
Selective search	0.9675	0.9541	0.9008	0.9913	0.9464	0.9132	0.9862	0.9601	0.9589

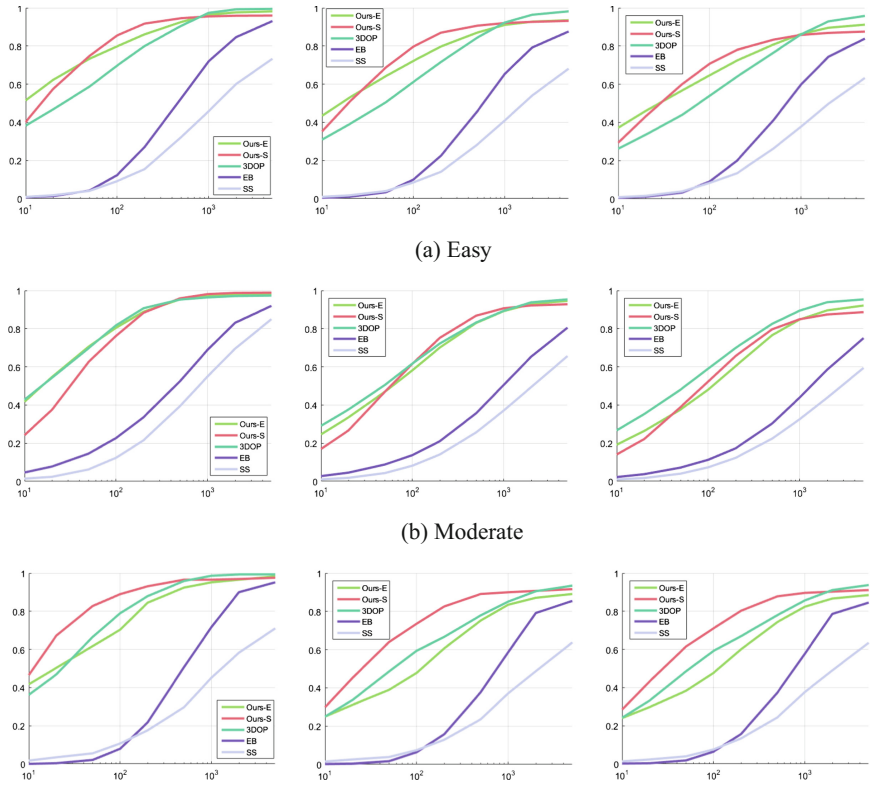
Easy, *Moderate* and *Hard* in table indicate the difficulty level of detection for each object category, which classified by different occlusion and truncation levels.

D. Proposal Recall We evaluate recall of the two variants of our approach: proposal based on Edge Boxes and proposal based on Selective Search. We denote the former variant as *Ours-E* and the latter as *Ours-S*.

Figure 3 shows recall as a function of the number of candidates. We can see that in general our coarse to fine object proposal generate framework greatly improved the result of our baseline Edge Boxes and Selective Search. By using 100, 500, and 1000 proposals, our approach gets much higher average recall than the state of art approach 3DOP in categories of *Cyclist* and *Pedestrian* of all difficulty levels. In the class of *Car*, competitive result is also generated. When 1000 candidates are selected, we achieve

more than 90% recall for all species of object in *Moderates* and *Hard* regimes, while for *Easy* we need only 200 proposals to reach the same recall.

In Fig. 4, we show recall as a function of the IoU overlap for 500 proposals. We obtain significantly higher recall over the baselines across all IoU overlap levels, particularly for Cyclist.



(a) Easy

(b) Moderate

(c) Hard

Fig. 3. Bounding box Recall vs number of Candidates: We use an overlap threshold of 0.7 for *Car*, and 0.5 for *Pedestrian* and *Cyclist*, following the KITTI evaluation. The number of horizontal coordinate represents the IoU, and the number of vertical coordinate indicates the average recall (AR). The first, second and third rows represent different object categories of *Pedestrian*, *Car*, *Cyclist*. From left to right are for *Easy*, Moderate, and *Hard* evaluation regimes, respectively. In every subplot, the number of horizontal coordinate and vertical coordinate indicate the number of candidates and the average recall (AR).

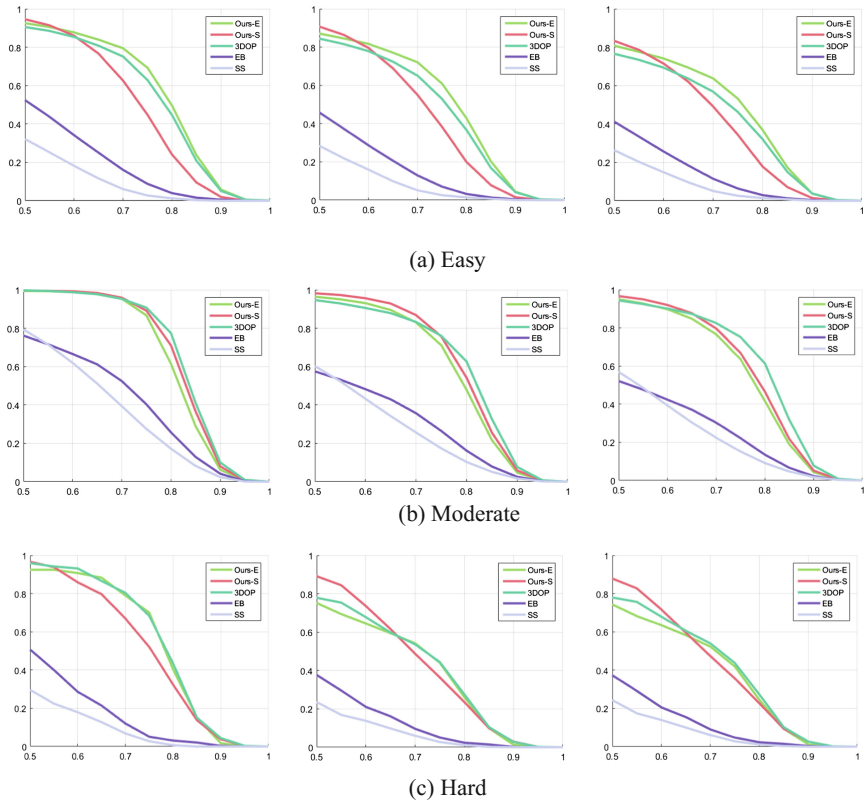


Fig. 4. Bounding box Recall vs IoU for 500 proposals: The number of horizontal coordinate represents the IoU, and the number of vertical coordinate indicates the average recall (AR). The rows (from left to right) and columns (from up to down) represent different object categories: *Pedestrian*, *Car*, *Cyclist*; and different difficulty level of detection: *Easy*, *Moderate*, *Hard*.

6 More Experiments

In order to balance computation with precision, many experiments have been done on the coarse object proposal approach Edge Boxes, experiments results are illustrated in Table 2. When the number of candidates comes up to 50000, recall hardly increases by

Table 2. Recall of Edge boxes with different number of candidates

Candidates	Category								
	Pedestrian			Car			Cyclist		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
30000	0.9693	0.9197	0.8916	0.9718	0.9045	0.8668	0.9552	0.8714	0.8602
50000	0.9842	0.9381	0.9192	0.9916	0.9651	0.9455	0.9828	0.8967	0.8914
100000	0.9877	0.9427	0.9284	0.9966	0.9829	0.9685	0.9897	0.9094	0.9030

Easy, *Moderate* and *Hard* in table indicate the difficulty level of detection for each object category, which classified by different occlusion and truncation levels.

simply adding candidate quantity. We finally set the sliding window step to 0.70 to generate about 50000 candidates per image.

7 Conclusion and Prospect

In this paper, we proposed a coarse to fine object proposal framework to generate high recall candidates for autonomous driving. Utilizing several simple low-level geometrical features and Bayesian probability models, we refined the immediate results produced by coarse proposal approaches, such as Selective Search and Edge Boxes, finally achieved state-of-art proposal result on all categories and all regimes on the challenging KITTI detection benchmark. In our coarse to fine proposal framework, the depth information generated from binocular images plays an important role in refining stage, it is an effective complement cue to RGB color information. More potential on depth information remaining to be excavated. The object detector will be catenated to our object proposal result later. And the code will be available on https://github.com/medivhliu.

References

1. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361 (2012)
2. Zitnick, C.L., Dollár, P.: Edge boxes: locating object proposals from edges. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 391–405. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10602-1_26
3. van de Sande, K.E.A., Uijlings, J.R.R., Gevers, T., Smeulders, A.W.M.: Segmentation as selective search for object recognition. In: IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, pp. 1879–1886, November 2011
4. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer Vision and Pattern Recognition, pp. 580–587. IEEE (2014)
5. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision IEEE, pp. 1440–1448 (2015)
6. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vis. **59**, 167–181 (2004)
7. Chang, K.Y., Liu, T.L., Chen, H.T., Lai, S.H.: Fusing generic objectness and visual saliency for salient object detection, pp. 914–921 (2011)
8. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 73–80. IEEE (2010)
9. Arbelaez, P., Ponttuset, J., Barron, J., Marques, F., Malik, J.: Multiscale combinatorial grouping. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 328–335 (2014)
10. Cheng, M.M., Zhang, Z., Lin, W.Y., Torr, P.: BING: binarized normed gradients for objectness estimation at 300fps, pp. 3286–3293 (2014)
11. Chen, X., Kundu, K., Zhu, Y.: 3D object proposals for accurate object class detection (2015)

12. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell., p. 1 (2016)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Comput. Sci. (2014)
14. Bigdeli, S.A., Budweiser, G., Zwicker, M.: Temporally coherent disparity maps using CRFs with fast 4D filtering. In: IAPR Asian Conference on Pattern Recognition IEEE (2015)
15. Seki, A., Pollefeys, M.: Patch based confidence prediction for dense disparity map. In: British Machine Vision Conference (BMVC) (2016)
16. Guney, F., Geiger, A.: Displets: resolving stereo ambiguities using object knowledge. In: Computer Vision and Pattern Recognition. IEEE (2015)
17. Guo, K., Li, N., Zhang, M.: The application of RANSIC in video mosaicing. In: Second International Conference on Electric Information and Control Engineering, pp. 652–655 (2012)
18. Wang, X., Yang, M., Zhu, S., Lin, Y.: Regionlets for generic object detection. IEEE Trans. Pattern Anal. Mach. Intell. **37**(10), 2071–2084 (2015)
19. Tuytelaars, T.: Dense interest points. In: IEEE Conference on Computer Vision & Pattern Recognition, pp. 2281–2288 (2010)

Study on Recognition and Management of Cartographic Topology Preprocessing Mode

Chengming Li(✉), Xiaoli Liu, Wei Wu, and Yong Yin

Chinese Academy of Surveying and Mapping,
Lianhuachi Xi Road, Haidian District, Beijing, China
83391860@qq.com

Abstract. Targeted at the problem of inconsistent spatial data found during automated synthesis of cartography, an algorithm that could be utilized to rapidly recognize and process the spatial relationship of such incorrect spatial data was proposed in this paper. Not only was this algorithm able to remove topology errors and reduce data redundancy, but was beneficial to improve automaticity of automated synthesis of cartography. Related results indicated that satisfying requirements of production practices, which was conducive to foundation database maintenance, was provided with rather high practicability and stability.

Keywords: Automated synthesis of cartography · Topology · Preprocessing · Model recognition

1 Introduction

Automated synthesis of cartography has always been one of the most challenging and creative problems in the field of cartology. With an aim to tackle such an issue, the key is to settle up experiences and judgment rules of experts in cartography into synthesizing operators that can be understood and executed by computers. During the concrete execution of synthesizing operators, much more attentions should be paid to the consistency of spatial data. As topological relation is an important content of spatial data consistency, the large-scale map database of cities at this present stage serves as the research object in this paper. It is found that some problems generally lie in the topological relation of library data. For example, although area objects among different layers can satisfy cartographic demands visually, in fact, there exist gaps and fragments that only can be searched by some present software. Moreover, they fail to carry out automated repair for these gaps and fragments. Consequently, it is difficult for such software to meet requirements for the automated synthesis of cartography concerning factor space constraints.

Therefore, topology preprocessing is studied to fix and accurately reflect the spatial relationships among space entities so that more perfect topological structures can be formed logically to satisfy demands of automated synthesis of cartography. In this paper, studies should be performed from two aspects of topology preprocessing mode recognition and management. One is study on recognition of information model; the other is the study on management of processing model.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 228–233, 2017.
DOI: 10.1007/978-981-10-3969-0_26

2 A Study on Information and Processing Models

2.1 Target Individual Information

In the large-scale urban database, the topological relation of space entities still has gaps and fragments. Through observational studies, such gaps and fragments can be roughly divided into three types of intersection, separation and interleave, as shown in Fig. 1. Based on summaries made during working practices, continuity and connectivity of topology are used to solve such a problem. When changes in measuring scale of a map take place, size and shape and it both vary correspondingly, so are some properties of map graphics, such as length, area, angle and relative distances between each other of the map graphics. However, some graphical properties may remain unchanged, including contiguity, inclusiveness and intersection of map graphics as well as geometric types (e.g., dots, lines and planes) of elements, etc. These properties that keep invariant during continuous changes of graphs are referred to as the topological attribute.

(a) Intersection (b) Separation (c) Interleave

Fig. 1. Types of gaps and fragments

Entities in geographic space are diverse and have shapes ever-changing. Nevertheless, they are represented by three elements of the dot, the line and the plane on the map. In a two-dimensional plane, they respectively correspond to three basic graphic elements known as topological elements, such as junction point, arc and region.

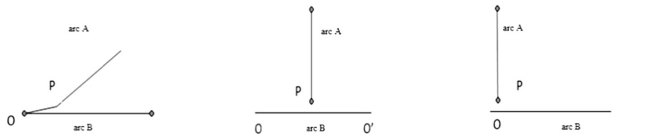
In detail, junction point consists of isolated points, endpoints of arc, joints of arc, inner points of polygon and boundary points of polygon, etc. Arc refers to the ordered segment between two junction points that can be identical or different. Region stands for a polygonal domain enclosed into by multiple arc closure chains and can be denoted by polygon.

The major research object of this paper is the spatial relationship of arcs in the topology, which is also known as the relation between dots in two arcs. Furthermore, as for such dots on the arc, they include both space coordinates and attribute information. Therefore, at the time of considering spatial relationship between dots, the weight relation of them should also be taken into account.

2.2 Recognition of Target Information

To solve gaps and fragments in space entities, such space entities should be recognized and target information identified is classified into three types.

Deleting point: Arc A and B intersect on point O; besides, the distance from point P on arc A to arc B is less than the threshold, and, the distance from P to O is also less than this threshold, point P is the deleting point, as presented in Fig. 2(a).



(a) Deleting point P (b) Projection Point P (c) Fitting Point P

Fig. 2. Types of gaps and fragments

Projection Point: In the case that arcs A and B fail to intersect, the distance from point P on arc A to arc B is less than threshold while distances from P to point O and O' on arc B are larger than threshold, point P is the projection point, as presented in Fig. 2(b).

Fitting Point: Under a circumstance that arcs A and B do not intersect, the distance from point P on arc A to point O on arc B is less than threshold, both P and O are fitting points, as presented in Fig. 2(c).

2.3 Processing Model

The basic processing model is divided into three types. First, the deleting point is a direct deleting point P, the same as point P in Fig. 2(a). Secondly, the projection point is P' projected onto arc by point P; the projection point P' is inserted onto the arc and P moves up to P', as point P in Fig. 2(b). Thirdly, there are two processing modes for fitting points and fitting processing is carried out for dots of the third type; as a result, a new point P' is fitted or a point P' with the maximum weight is selected; subsequently, all points move up to P', as points P and O in Fig. 2(c).

3 Algorithm Implementation

3.1 Recognition of Target Information

In order to find the target information rapidly, R-tree for outsourcing frame of arc is established to find paired arcs that may be target information. Then, dots on the paired arcs are estimated in detail to improve algorithm efficiency.

Step 1: The R-tree is set up for all outsourcing frames of arcs of spatial data and a range of an additional threshold is expanded outward for the outsourcing frame of every arc to search for all possible intersecting paired arcs in arc R tree. Then, the paired arcs are put into an array of possible intersecting arcs.

Step 2: As for arc A and arc B in pair in the array, every point P on arc A is estimated to be on arc B or not; if yes, they serve as the next paired arcs; if not, proceed to Step 3 up until the final point on arc A. After the completion of arc query, search for the next pair to the end of the final paired arcs.

Step 3: With regard to outsourcing frames of segments composed by every two adjacent points on arc B, R tree is established for them; then, a range of an additional threshold is expanded for the outsourcing frame of point P on arc A, to search for segments on arc B that conform to target information. If not, repeat Step 2 up until all paired arcs in the array participate into target information recognition; otherwise, carry out Step 4 for every segment found.

Step 4: After segments conforming to target information are found, the buffer circle of point P is used to perform intersection judgment together with segments. If no intersection, they are not recorded; otherwise, arc A with point P and arc B with segment are noted down accompanied with recognition and identification of target information in three types.

The relation between the distances PA and PB from point P to endpoints A and B of segment AB, and the threshold Distance Epsilon is adopted to identify and distinguish target information of three types:

If PA is less than Distance Epsilon, point A is on arc A, and the distance from fitting points P and A to the next dot of point P on arc A (along the direction of AP) is less than Distance Epsilon as well, then P is the deleting point.

If PB is less than Distance Epsilon, point B is on arc A, and the distance from fitting points P and B to the next dot of point P on arc A (along the direction of BP) is less than Distance Epsilon as well, then P is the deleting point.

If PA and PB are both longer than the Distance Epsilon, then, P is the projection point.

If PA is no more than the Distance Epsilon and PB is more than the Distance Epsilon, then, P is the fitting point.

If PA is more than the Distance Epsilon and PB is no more than the Distance Epsilon, then, P is the fitting point.

If PA and PB are less than the Distance Epsilon, then, P is the fitting point.

3.2 Clustering of Target Information

According to the spatial relationship of target information, all are clustered into various processing units.

Considering that deleting points require no clustering, the clustering of target information only takes projection and fitting points into account.

Step 1. Target information is put into a processing unit newly created; and, the outsourcing frame of point P on arc A in the target information is expanded outward with a range of one threshold and then put into R tree.

Step 2. In the R tree, search point P on the next arc A of the target information; if any target information can be found, it is put into the processing unit found; otherwise, return to Step 1 up until all target information is processed completely.

The start Index of target information is greater than true Index, this is the projection point; otherwise, it is the fitting point that can be divided into nodes and intermediate points. Judgment condition for nodes is that the index of points on arc recorded should be zero or equal to the number of arc points minus 1.

3.3 Updating of Target Information

When the processing units containing projection points are processed, a new projection point P' should be inserted on an arc, which may cause changes in the index of all target information that has been recorded on the arc. Hence, such target information recorded should be updated. Firstly, spatial index is created for target information to search for target information requiring updating and judge whether the index of such a point is larger than that of the insertion point P. If yes, updating is needed, that is the index of this point is equal to (its index +1); otherwise, it remains unchanged.

4 Experimental Testing

Below, road network data of 650 km^2 are taken as the example. To be specific, topology construction, and topology preprocessing and topology reconfiguration need 179 s in total in terms of efficiency, among which, the topology preprocessing requires 123 s. Moreover, from the perspective of accuracy, 13,049 articles of target information are identified with an accuracy of 100% (see Table 1).

Table 1. Table of topology preprocessing result comparisons for road surface

Name of layer	The number of plane elements	The number of arcs	The number of point elements
Road surface	16651	27638	513931
Intersection	15115	26050	263656
Road surface covered	627	1148	7026
Intersection covered	1168	1736	18457
In total	33561	92852	795684
Total time (in second)	Time of topology preprocessing (in second)	The number of target information identified	Accuracy of recognition
179	123	13049	100%

5 Conclusions

In the large-scale map database, the spatial relationship of spatial data derived from library data is found to have some problems generally. In this paper, a rapid processing algorithm is presented for such problems. With regard to large-scale urban data, including buildings and road surfaces, etc., large quantities of experiments are performed. The experimental results prove the validity of employing topological relationship to express and repair the spatial relationship of spatial data. Based on the recognition and processing model, topological relationship is adopted to process spatial data. On one hand, as far as accuracy is concerned, data consistency required by automated synthesis of cartography is satisfied; on the other hand, as for the corresponding validity, demands of actual production practices are also met.

The research results in this paper laid a good topology basis for the road network synthesis, especially the road stroke connecting. Use the algorithm of this paper, the road network could form a more intuitive and full road stroke, and it could be convenient to calculate stroke grade roads.

Because of the characteristic of the city is different, the large-scale map of city will also be different. In this paper, the urban road network is the square and circular, the star and irregular shape did not take into account. Aiming at this point, the following research will continue to modify the algorithm to improve the adaptability of algorithm.

Acknowledgments. This project is supported by the Special Scientific Research Fund of Surveying and Mapping Geographic Information Public Welfare Profession (201512020), National SciTech Support Plan (2015BAJ06B01), and Basic Scientific Research Business Expense Project of the Chinese Academy of Surveying & Mapping (7771606).

References

1. Chen, S., et al.: Introduction to Geographic Information System. Science Press, Beijing (2000)
2. Wu, H.: Application of principle of convex hull in point cluster synthesis. Eng. Surv. Mapp. **6**(1), 1–6 (1997). Wuhan
3. Wang, P.: A topology processing method for road network generalization. GeoInf. Sci. **25** (1), 65–68 (2009)
4. Wu, L.: Principle, Method and Application of Geographic Information System. Science Press, Beijing (2001)
5. Wang, J., Wu, F.: Principle and Method of Digital Automated Synthesis of Cartography. The Chinese People's Liberation Army Publishing House, Beijing (1997)
6. Li, J.: Topology check and data preprocessing for GIS of lands and housing. Surv. Mapp. Shanxi **10**(1), 9–12 (2003)
7. Sun, D., Xiao, F., Liao, X., et al.: Topological structure construction algorithm for urban road network based on preprocessing. Comput. Eng. Appl. **44**(23), 233–235 (2008)
8. Liu, Y.: Expression, recognition and synthesis of spatial graphics. The PLA Information Engineering University (2005)
9. Cheng, S.: Construction and update of topological relationship in GIS. Master's thesis of the PLA Information Engineering University (2002)
10. Cheng, B.: Study and practice of automated synthesis of cartography for road network. The PLA Information Engineering University (2006)
11. Wu, H.: A Study on Basic Theory and Techniques of Map Generalization. Surveying & Mapping Press, Beijing (2004)
12. Guo, Q.: A study on new theory and techniques of automated synthesis of cartography. Doctoral dissertation of Wuhan Technical University of Surveying and Mapping, Wuhan (1998)

Research on Hot Topic Discovery Technology of Micro-blog Based on Biterm Topic Model

Jun Feng(✉) and Yu Fang

Department of Electronic and Information Engineering,
Tongji University, Shanghai, China
fengjunasd@126.com, fangyu@tongji.edu.cn

Abstract. In order to overcome data sparsity and expression diversity problems of short text and to improve the quality of clustering, this paper proposes a text feature enhancement method based on biterm topic model (BTM). First, we obtain the high frequency word matrix of underlying topic based on the extraction on the corpus using BTM and then strengthen the traditional vector space model (VSM) selectively with this matrix to reduce vector dimension and highlight the main features. Also, we propose a heat calculation equation combining with propagation characteristic and time effect of micro-blogs so that we can better demonstrate the evolution of a topic and analyze it. Experiments show that our method has achieved good results in improving the clustering quality and the heat calculation equation is also beneficial to the discovery and evolution of hot topics.

Keywords: Biterm topic model · Feature enhancement · Topic discovery · Hot topic evolution

1 Introduction

The hot topic discovery technology using clustering analysis or topic extraction is to dig out meaningful content to which users pay their attentions from a large amount of information. It belongs to Topic Detection and Tracking (TDT) [1] and can be used in the entire area or in a specific domain. For some hot topics, it can completely find the attitudes of people and the subsequent of popular events. More important is that hot topic discovery can find some emerging hot topics without a lot of reports.

As one of the most important micro media forms, micro-blog has lots of features such as wide information coverage, real-time, highly interactive and simple metadata. However, this short text will suffer from severe data sparsity problem and its oral and diverse expression is not conducive to the selection of characteristics too.

To solve the problems above, this paper focuses on the research on data sparsity and expression diversity by applying biterm topic model (BTM) [2] in topic extraction of micro-blogs and strengthening the VSM by topic-word matrix. This can reduce vector dimension and preserve more original information at the same time. Also, we merge words that potentially express the same topic to solve the diversity problem. What's more, we improve the K-means algorithm with propagation characteristic and time

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 234–244, 2017.
DOI: 10.1007/978-981-10-3969-0_27

effect. The K value is adaptive and the clustering is incremental and a heat calculation equation is proposed to describe the degree of hot events and their evolution process.

2 Related Works

2.1 Research on Short Text Clustering

As a typical short text, micro-blog will suffer from severe data sparsity problem. At present, the improvement of short text clustering are mainly based on feature selection. In 2002, FTC algorithm [3] proposed by Beil et al. holds that some specific words will show in documents sharing the same category. This means those words in a certain degree can be used as a standard of clustering partition as well as a description of a cluster. Hu et al. proposed a feature extraction algorithm [4] based on repeated strings. Since repeated string is involved in concepts of word order and semantic and is more convenient than vocabulary statistics, thus it is an improvement. Also, another way to improve is short text feature extensions. Gabrilovish [5] proposed to improve the accuracy of the calculation of the similarity by enrich the short text with Wikipedia. Hotho et al. [6] use the Hownet ontology to translate the word vector into the concept vector to measure the similarity through the concept similarity. Frey and Dueck [7] put keywords into search engine and select some key information to fuse with the original vector. With the development of topic model, Song and Zhang [8] apply LDA to micro-blog topic extraction and treat a set of micro-blogs of each user as a document since LDA perform not well in short text. Tang [9] use BTM to express the topic combining with traditional VSM to make up for the lack of VSM. Zhang [10] use topic-word matrix of BTM to expand the VSM and solve the data sparsity problem to a certain extent. Nowadays, more researches are took on topic model include variant model or combining with algorithm. Wang [11] use content features and user features of Sina Weibo to construct mixture LDA model and Wu et al. [12] proposed a EM-LDA model to dig out topics in micro-blog. Wang and Peng [13] combine TE-LDA model with ARIMA algorithm which can capture topics changed over time.

2.2 Research on Hot Topic Discovery Based on Micro-blog

Micro-blog is a kind of interaction between users and these characteristics like comment and forward can well describe the propagation and diffusion of events. Also, the time effect of fast outbreak slow down can well describe the evolution of events. The research and application of micro-blog involve many aspects mainly concentrated in the public opinion monitoring, emotion analysis, information recommendation and hot topic discovery. Jiang [14] elaborate the characteristics of public opinion impact and make a tentative discussion on how to guide well the public opinion. O'Connor et al. [15] dig out users' attitudes to major events from a large number of micro-blogs using emotional analysis technology and the results were compared with the traditional public opinion survey. Cheng et al. [16] proposed a micro-blog recommendation method based on information propagation theory. Some hot topics are recommended to users after considering aspects of information amount, reading cost and time delay.

3 Feature Enhancement Based on Biterm Topic Model

Micro-blog is a kind of short text less than 140 words. According to simple statistics, the lengths of 56.8% micro-blogs are less than 15 characters and 83.3% are less than 25 words. The average length of micro-blogs is only 17.8 words. Short text will suffer from severe data sparsity problem since only three or four significant keywords is meaningful after feature selection into VSM. First, this chapter introduces the principle of BTM and topic extraction. Then, we obtain the matrix of high frequency words according to the topic-word distribution matrix. Finally, we reduce the dimension and strengthen the feature of the basic vector using high frequency word matrix.

3.1 Obtain the Underlying High Frequency Words Matrix by BTM

The BTM is proposed based on Latent Dirichlet Allocation (LDA) and mixture of unigrams to strengthen the topic model learning by word co-occurrence patterns (biterm: the disorder word co-occurrence in a short text window) and to infer topic distribution using a variety of sampling topics of the whole corpus. Not only to keep the relationship between words but also to infer probability distribution over different topics since different biterms are independent of each other. BTM is a generative model with the whole corpus sharing the same global topic distribution of θ and each topic corresponds to a Multinomial distribution of φ, where θ and φ is a Dirichlet prior distribution with a super parameter α and β respectively. For word pairs of the whole corpus, the BTM defines the following generation process:

(1) Word distribution: For each topic z, describe the word distribution $\varphi_z \sim \text{Dir}(\beta)$.
(2) Topic distribution: Describes a global topic distribution for the entire short text collection $\theta \sim \text{Dir}(\alpha)$.
(3) Extraction: The following operation is performed for each word pair b in the set B, assuming that $b = (w_i, w_j)$.
   (a) Topic: Extract topic z from the global topic distribution θ, namely $z \sim \text{Multi}(\theta)$.
   (b) Words: Extract two words w_i, w_j from topic z, namely $(w_i, w_j) \sim \text{Multi}(\varphi_z)$.

According to above, the joint probability formula of *b* is:

$$P(b) = \sum\nolimits_z P(z)P(w_i|z)P\left(w_j|z\right) \tag{1}$$

So the probability over the whole corpus is:

$$P(B) = \prod\nolimits_{(i,j)} \sum\nolimits_z \theta_z \varphi_{i|z} \varphi_{j|z} \tag{2}$$

The topology of the BTM is shown in Fig. 1:

Three results will be obtained using BTM to carry on topic extraction to micro-blogs collection include topic distribution result (.pz), topic-word distribution result (.pz_w) and document-topic distribution result (.pz_d). Represent VSM with

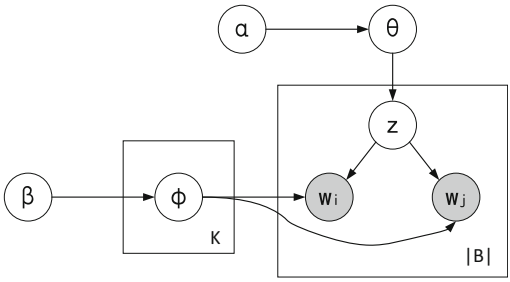


Fig. 1. The topology of the BTM model.

document-topic matrix and obtain the optimal *k* value judged by the clustering method with accuracy rate. Then, sort topic-word matrix in descending order of probability and take a certain proportion of words as high frequency word matrix to strengthen the VSM.

3.2 Strengthen Vector Space Model with High Frequency Words Matrix

First of all, simply treatment will be carried on micro-blogs to form the vector including Chinese segmentation, stop words pretreatment and then retain only the verb and noun through tagging function of the segmentation tool. Last, remove words whose frequency is less than a certain value.

The enhancement of VSM is processed by choice. As everyone knows, a word can belong to different topics. For example, the word "nutrition" belongs to both fitness topic and medical topic, so we assume fitness and medical are underlying topics of "nutrition" which means we need to strengthen the characteristics respectively.

In general, for each feature *f* in the document, we want to discuss in three cases:

(1) Single topic: Feature f only belongs to the topic of T_i, then we identify f by T_i that is to use weight of T_i as weight of f.
(2) Multi topics: Underlying topics of Feature f include T_{i1}, T_{i2}… T_{im}, we need to use these m topics to identify the current document.
(3) No topic: Features f does not belong to any topic, then we do not strengthen this feature just leaving f itself to identify the document.

For example, the vector $V = (f_1: w_1, f_2: w_2, f_3: w_3, f_4: w_4, f_5: w_5, f_6: w_6, f_7: w_7)$, assuming that f_1, f_2, f_3, f_4 belong to T_1, and f_4, f_5 belong to T_2 and f_6, f_7 do not belong to any topic. Then the vector after enhancement is $V' = (T_1: w_1 + w_2 + w_3 + w_4, T_2: w_4 + w_5, f_6: w_6, f_7: w_7)$. From which we can see:

(1) Single topic: After merging features, the main topic T_1 of the text is highlighted.
(2) Multi topics: This kind of selective enhancement is also taking into account that f_4 to be a minor underlying topic T_2.
(3) No topic: As for f_6, f_7 which do not belong to any topic, we just leave them of sparsity as a suppression.

The VSM after enhancement reduces the dimension in a certain degree as well as highlighting the main underlying topic of document. So if we keep a certain vector dimension M, the vector after enhancement can retain more original information and highlight the main features to describe micro-blogs in a better way.

4 Hot Topic Discovery and Heat Evolution of Micro-blogs

4.1 Formation of Hot Topic and Heat Affected Factors

As an important form of network topics, micro-blog can be seen as a kind of text stream on the timeline and the hot topic is contained in it. After release of a micro-blog, due to a series of reasons, may be the influence of users or the popularity of contents, micro-blog has been widely concerned with a large number of interaction like forward and comment. Then a large number of relative micro-blogs outbreak, a hot topic is emerging.

Sum up the impact factors into two aspects: the propagation and time effect. Propagation effects include:

(1) Numbers of likes: reflecting favorite and recognition of the micro-blog.
(2) Numbers of comments: reflecting the degree of interaction for this topic.
(3) Numbers of forwards: reflecting the propagation of the topic.

And time effect means the heat of a micro-blog will tend to be reduced as hot topic is no longer concerned. That is, as the hot events outbreak, evolve and then disappear, the heat of micro-blogs will change.

4.2 Heat Calculation Equation of Micro-blog

According to the analysis of the factors above, we proposed heat calculation equation of a single micro-blog:

$$H = u \cdot H_0 \tag{3}$$

Where u is time effect factor and H_0 is a static base heat.

(1) Time effect factor u: The longer the time is, the smaller the time effect is. Below is the specific equation:

$$u = e^{\frac{t-t_0}{\alpha}} \tag{4}$$

Where t is the publish time of a micro-blog, t_0 is current time and α is time factor.

(2) Static base heat H_0: H_0 is only related with interactions of micro-blog. Below is the specific equation:

$$H_0 = \log(f+1) + \frac{r - r_0}{\sqrt{r - r_0}} + c \tag{5}$$

Where f is number of fans, r is number of forwards and c is number of comments. r_0 is the balance factor with 1/1000 value of f which means when the number of fans of a user is very high, a micro-blog is meaningful only when the number of forwards achieves a basic forward level so that the celebrity effect is weakened to a certain degree.

4.3 Incremental K-means Clustering Algorithm with Adaptive K Value

Hot topic is an evolution process with a lot of micro-blogs produced every day. When using a clustering algorithm to discover hot topic, it must be conducted in an incremental form as new micro-blogs continue to join in the topic and the K value must be changed following changes of topics. According to requirements above, this paper improves the classical K-means clustering algorithm to overcome the initial and adaptive problems of K value and change algorithm into incremental form.

(1) Initial K value and clustering centers: In K-means algorithm, the choice of K value and initial cluster centers will greatly influence the effect of algorithm. Before clustering, the number of topics is unknown so we need to obtain the K value through other ways. Basic hierarchical clustering with a certain threshold is used in this paper and the local optimal result of several clusters will be used as input of K-means algorithm and after several iterations to obtain the global optimal solution.
(2) Incremental clustering: In order to find the evolution process of a topic, we need to observe changes of hot topics every period time. We will divide a span of time into a number of increments and use hierarchical clustering in the first increment to form the initial cluster centers as an input. In the following K-means clustering, each change of incremental clustering result will be recorded.
(3) Adaptive K value: With the new generation of micro-blogs, the topic and number of topics will change. By setting a threshold in each iteration process, a micro-blog with a similarity less than the threshold will not be classified but added to the temporary list. After each iteration, whether to create a new topic depends on calculating the dissimilarity with the existing clusters. That is, if dissimilarity between a micro-blog in a temporary list and the existing cluster is greater than the threshold, then we create a new cluster and K value plus one until all the incremental clustering done.

5 Experiments

5.1 Data Preparation

The test dataset of this paper is crawled from Sina using crawler program with sorts of selected keywords ranging from January 1, 2016 to June 30, 2016 with a total of 110389 and 15 topics involved. After preliminary screening of retaining the original ones and removing duplicate ones or those of less than ten words, we picked up ten

topics from the screening results. They are Kobe 20 Years, Baidu Post Bar Event, The Spring Festival Gala, College Entrance Examination, National Football Team Top Twelve, Heyi Hotel Attack Event, Qvodplayer Jurisprudence Case, Champions League Final, Man-machine Chess Game, Wei Zexi Baidu Promotion Event, Gravitational Waves and Zhihu Yao Tong Fraud. There are 1000 each topic and 10000 in total.

Segmentation tool is a java version of ICTCLAS of the Chinese Academy of Sciences and user import dictionary is vocabulary table of different fields from Sogou input, 149568 in total. We translate the 10000 chosen micro-blogs into VSM after word segmentation, stop words filtering, low frequency words filtering and calculating the TF-IDF value. Finally, according to publish time of each micro-blog, the dataset is divided into several sub sets with a certain time interval.

5.2 Analysis of Feature Enhancement Effect

BTM is an unsupervised model and topic number k must be set before modeling. The accuracy is verified by standard K-means clustering algorithm. Since different k value will affect the extraction effect, the experimental k value will be set with 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50 respectively and number of iterations is 500, $\alpha = 50/k$, $\varphi = 0.01$, The result of K-means will be influenced by initial centers, so results will be the average accuracy rate of 10 times of experiments shown in Table 1:

Table 1. Accuracy of different k value

K value	7	8	9	10	11	12	13	14
Accuracy	0.486	0.544	0.570	0.607	0.700	0.764	0.668	0.707
K value	15	20	25	30	35	40	45	50
Accuracy	0.650	0.600	0.437	0.527	0.530	0.440	0.447	0.449

As we can see from the table, when the k is 12, the accuracy of clustering is optimal, so the following experiments will be based on $k = 12$. By using BTM output file k12.pz_w, we sort word frequency of each subject in a descending order and keep top 2% high frequency words as topic features to form the underlying topic words matrix.

In order to verify the enhancement effect of underlying topic words matrix on initial VSM, using VSM without enhancement as compared, we conduct experiments on the dimension of 100, 300, 500, 700, 900, 1100, 1300 and 1500 respectively. The algorithm is standard K-means clustering with optimal k value of 12 and random initial centers. We use the cosine similarity to calculate text similarity and evaluate clustering results with precision rate, recall rate and F-measure value on an average result of ten experiments. And we also compared our results with LDA and another improved method proposed in Ref. [9].

As we can see in Fig. 2, there is almost no improvement on LDA compared to VSM and the results are unstable because LDA is not suitable for short text like micro-blog. The effect of VSM combining with BTM and the effect of our enhanced

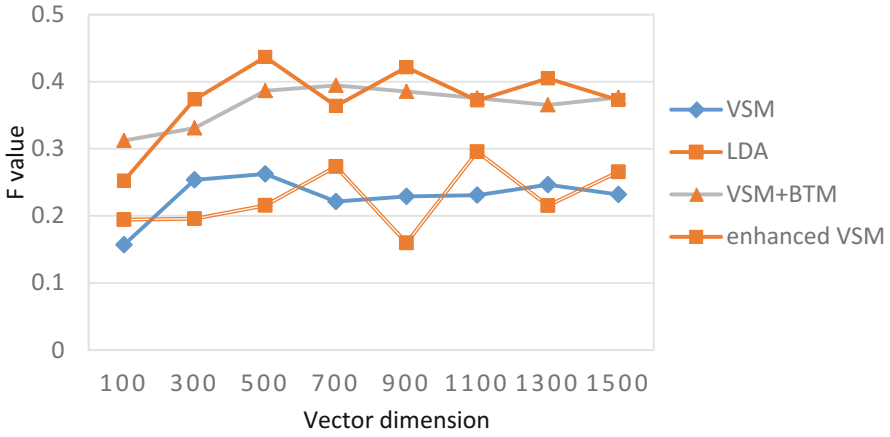


Fig. 2. The comparison between VSM and BTM model.

VSM are both improved compared to traditional VSM. And ours is slightly better than another, which is mainly because keywords hold a heave weight in short text, thus after enhancement, keywords can be even more representative for the topic of short text. This result shows that the vector after enhancement can better describe the information of micro-blogs.

5.3 Hot Topic Discovery and Its Evolution

In order to find hot topics, we need to understand the characteristics of hot topics, so we select two hot events and two general events to make a broken line of micro-blogs' heat shown in Fig. 3:

As we can see from above, hot events generally have a cycle of happen, burst, down and end, presenting in the figure is a line rise suddenly at first and then slowly decline and finally disappear. However, for normal events, the line will up and down around a certain value. According to this feature, we can find hot topics from all topics.

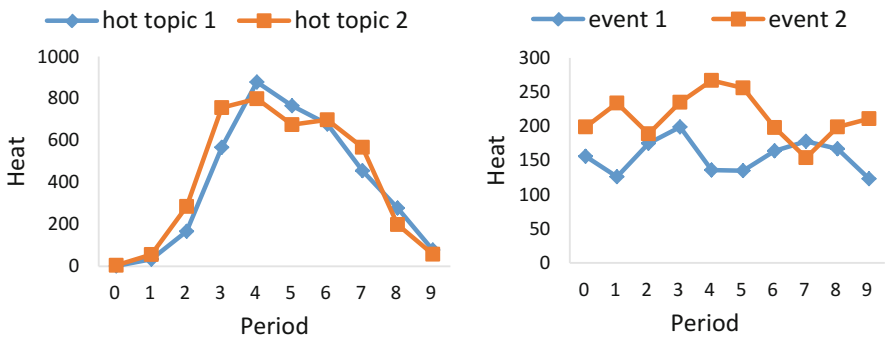


Fig. 3. The characteristics of hot and normal events.

According to the experiments of last section, we choose 500 as our experimental dimension since it contributes to a relatively good effect. Different from above standard experiments, we use the incremental clustering algorithm with dataset from January 8, 2016 to February 9, 2016, some events of this month is shown as follows (Table 2):

Table 2. The duration of some hot topics

Events	Number	Time interval
Qvodplayer	A	2016.01.08–2016.01.11
Baidu Post Bar	B	2016.01.14–2016.01.21
Zhihu Tongyao	C	2016.01.16–2016.01.21
Spring Festival	D	2016.01.21–2016.02.09
Highway daily	E	2016.01.08–2016.02.09
Others	-	2016.01.08–2016.02.09

Dataset is divided into 11 increments with a unit of 3 days. The first increments carry on hierarchical clustering algorithm to get $K = 2$ and we set $\alpha = 0.3$, $\varphi = 0.8$, then calculate each incremental iterative process.

From Fig. 3, we can see that A, B and C are hot events and their occurrence of peak time is consistent with the actual time. The D appeared two peaks, one appeared in the Spring Festival holiday and the other is in New Year's Day. The second peak higher is also consistent with our cognitive. While E is a daily topic with a smooth heat every day, so it is not a hot event (Fig. 4).

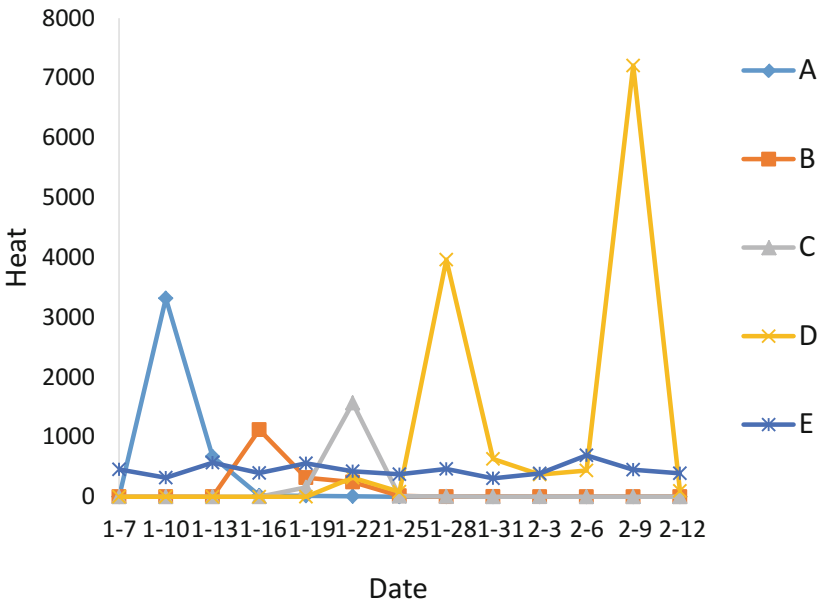


Fig. 4. The heat evolution of topics of micro-blogs.

6 Conclusion

This paper proposes an enhancement idea for traditional VSM based on BTM. According to the data sparsity problem of micro-blog and its oral and diverse expression, we strengthen words selectively which share the same underlying topic. Reduce dimension of vectors, at the same time, relatively preserve more original information and highlight the main features. Experiments show that our method has a certain promotion in accuracy compared to the traditional one and get a better clustering quality. This result provides the basis for a hot topic discovery. Then, this paper also put forward a heat calculation equation by analysis of propagation characteristic and time effect of micro-blogs. Experiments show that results of this equation can well describe the evolution process of hot topics so as to find a hot topic.

In addition, this paper only improves the representation of the vector and still has a lot to improve in the calculation of similarity or similarity from the semantic angle, which will need further research.

References

1. Allan, J.: Introduction to topic detection and tracking. In: Allan, J. (ed.) Topic Detection and Tracking, pp. 1–16. Springer US, New York (2002)
2. Yan, X., Guo, J., Lan, Y.: A biterm topic model for short texts. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1445–1456. ACM (2013)
3. Beil, F., Ester, M., Xu, X.: Frequent term-based text clustering. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 436–442. ACM (2002)
4. Hu, J., Xu, H., Liu, Y.: Algorithm of repeats-based term extraction and its application in text clustering. Comput. Eng. **33**, 65–67 (2007)
5. Gabrilovich, E.: Feature generation for textual information retrieval using world knowledge. ACM SIGIR Forum **41**, 123 (2007)
6. Hotho, A., Staab, S., Stumme, G.: Ontologies improve text document clustering. In: Third IEEE International Conference on Data Mining, pp. 541–544 (2003)
7. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**, 972–976 (2007)
8. Song, L., Zhang, P.: System design of micro-blog public opinion based on LDA topic modeling method. Netw. Secur. Technol. Appl. **4**, 5–6 (2014). (in Chinese)
9. Tang, Q.: Short text clustering method based on BTM. Anhui University, Hefei (2014). (in Chinese)
10. Zhang, Y.: A short text similarity calculation method based on feature extension using BTM topic mode. Anhui University, Hefei (2014). (in Chinese)
11. Wang, Y.: Topic model based on mixture LDA model in microblogging services. Nanjing University of Posts and Telecommunications, Nanjing (2015). (in Chinese)
12. Wu, W., Wu, Q., Gu, J.: Hot topic extraction from E-commerce microblog based on EM-LDA integrated model. Mod. Libr. Inf. Technol. **11**, 33–40 (2015). (in Chinese)
13. Wang, H., Peng, Y.: Public opinion hotspots discovery based on topic model and ARIMA algorithm. Technology Square (2016). (in Chinese)

14. Jiang, H.: Characteristics of micro blog and its influence on public opinion. News Lovers First Half **5**, 85–86 (2011). (in Chinese)
15. O'Connor, B., Balasubramanyan, R., Routledge, B.R.: From tweets to polls: linking text sentiment to public opinion time series. In: ICWSM, vol. 11, pp. 122–129 (2010)
16. Cheng, J., Sun, A.R., Hu, D.: An information diffusion based recommendation framework for micro-blogging. J. Assoc. Inf. **12**, 463 (2010)

A Deduplication Algorithm Based on Data Similarity and Delta Encoding

Bin Song[1], Limin Xiao[1,2(✉)], Guangjun Qin[1,2], Li Ruan[1,3], and Shida Qiu[1,3]

[1] State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
xiaolm@buaa.edu.cn
[2] National Engineering Research Center for Science and Technology Resources Sharing Service, Beijing 100191, China
[3] Space Star Technology Co., Ltd., Beijing 100086, China

Abstract. Satellite applications such as remote sensing application are overwhelmed with vast quantities of data. Nevertheless, the storage resources in the satellite are so limited that it should be used more efficient. The similarity between the remote sensing data is high, but the dissimilar parts of the data distribute irregularly. When using the traditional deduplication algorithm to split the file into chunks, a large amount of chunks are exactly similar but not the same, which results in the bad effect of data deduplication. We propose a deduplication algorithm based on data similarity and delta encoding to reduce the usage of storage resources. The data similarity analysis can find out the similar data. The delta encoding technology can reduce the usage of storage resources. Through experiments on remote sensing application data, we have achieved deduplication ratios up to 30:1, and analyzed how the chunksize affect the experiment results.

Keywords: Deduplication · Similarity · Delta encoding · Satellite

1 Introduction

Satellite applications such as remote sensing application play a significant role in addressing challenges in global resources and environmental change, which are overwhelmed with vast quantities of data [1]. Nevertheless, the storage resources in the satellite are so limited that it should be used more efficient. Using deduplication technology to process the satellite application data is a viable solution. Deduplication technology [2] typically use block file compression methods to eliminate redundancy as is shown in Fig. 1.

First, splitting the file into small chunks. Second, calculating the fingerprint of the chunks using strong anti-collision hash functions such as MD5 [3] and SHA-1 [4]. Third, detecting duplicate data by fingerprint matching and deduplicate it. This method is based on identical detection.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 245–253, 2017.
DOI: 10.1007/978-981-10-3969-0_28

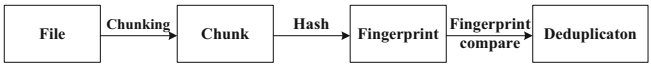


Fig. 1. The basic flow of typical deduplication technology

Nevertheless, according to the analysis of remote sensing data and the experiment results, traditional deduplication technology performs poor under satellite application scenarios. This is because the similarity between the remote sensing data is high, but the dissimilar parts of the data distribute irregularly. Taking into account the presence of noise, this situation get worse. When using the traditional deduplication algorithm to split the file into chunks, a large amount of chunks are exactly similar but not the same. As a result, a lot of similar chunks are ignored which result in low deduplication ratio.

In order to solve these problems, we propose a deduplication algorithm based on data similarity and delta encoding. Data similarity analysis can find out the similar data within massive satellite application data. The delta encoding technology can deduplicate chunks in byte granularity. In this way, we achieved the highest deduplication ratios up to 30:1 which are much higher than traditional deduplication methods.

The rest of this paper is organized as follows. The next section overviews some related work and Sect. 3 discusses the analysis of remote sensing data and key issues to be resolved in the processing of remote sensing data. In Sect. 4, we address the design and implementations of the deduplication algorithm based on similarity analysis and delta encoding. Section 5 discusses the experiment analysis of the algorithm's performance for remote sensing applications with different chunksize, and Sect. 6 concludes this paper.

2 Related Work

Deduplication technology can be categorized on the basis of various design considerations such as chunking granularity and deduplication timing.

Chunking Granularity. Different chunk based deduplication strategies are available to remove the redundant data as discussed below:

(1) Whole file chunking (WFC) [5]: The whole file chunking does not break files into a smaller chunk, rather than it treats the whole file as a chunk. It finds the hash value for the entire chunk which is the file index. If a new incoming file matches with the file index, then it is considered as duplicate and it points to the existing file index.
(2) Fixed-size partition (FSP) [6]: In this data deduplication algorithm, it breaks the files into equals sized chunks in which the chunk boundaries are fixed such as 4 KB, 8 KB, etc. As a result, content based checksum identified the chunks and stores only the index which does not exist.
(3) Content-defined chunking (CDC) [7]: In CDC deduplication algorithm, chunking boundaries are determined based on the contents of the file, so it is more resistant to the insertion and deletion. Similar to fixed size chunking method, CDC has

three important steps as follows: Dividing the file into variable blocks based on the chunk boundaries using such as Rabin finger printing algorithm. Generating the hash values for each block using common hashing technique such as MD5 or SHA-1. Identify the redundant data from the hash values.

Deduplication Timing. Another important criteria is when to duplicate the data. Data can be processed at three places, before being written into a disk (Inline) or after writing to the disk (Post), or both before and after written to the disk (Hybrid) [8].

(1) An Inline deduplication can be done the client side or when the data is transferring from the client to the server. Inline deduplication is a process where the data is deduplicated before it is written to disk. If a block of data arrives into the process, it analysis whether the data block has been processed already or not. If the data processed before, it pulled away from the redundant block then writes a reference to that block. If it identify the block of data is unique, the process writes the block into the storage. This method of deduplication used in work in RAM. The advantage of inline deduplication is that it does not require extra disk space.
(2) Post Process deduplication: Post-Processing operations are performed on the server side. The source data is written to the backup server storage and duplicates are cleared later. Post-processing deduplication process gets started once it is written to disk. The advantage of the post-process deduplication is that the performance is higher than In-Line deduplication. Another benefit of this method is the ability to share the index and metadata, hence clustering for higher availability is easier, and data replication can be much more efficient. The disadvantage is that the need for the fast disk cache, which typically makes the initial purchase price higher than in line-based solutions.

3 Problem Definition

3.1 The Analysis of Remote Sensing Data

The similarity between the remote sensing data is high, but the dissimilar parts of the data distribute irregularly. Taking into account the presence of noise, this situation get worse. When using the traditional deduplication algorithm to split the file into chunks, a large amount of chunks are exactly similar but not the same. However, most of the traditional deduplication technology is based on fingerprint detecting. As a result, a lot of similar chunks are ignored which result in low deduplication ratio. For example, Fig. 2 shows two similar remote sensing (RS) images of great lakes that were recorded by satellite a week apart. From the visual point of view, the two image are similar. However, the different parts of the two images distribute irregularly. This situation is made worse by the noise.

When using FSP algorithm to compare the two images, we split each of the image into 100 chunks. Calculating and comparing the 200 chunks' fingerprints with MD5 algorithm, only 4% fingerprints are the same. When using CDC algorithm, Rabin finger printing algorithm is used to determine boundary, the experiment results show that only

Fig. 2. Two similar RS images

23% of the chunks' fingerprints are the same. However, the similarity is much higher than the experiment using FSP or CDC algorithm. Using the deduplication algorithm based on data similarity and delta encoding, 87.5% redundant data can be removed.

3.2 Key Issues to Be Resolved

According to the above discussion, the effect of using traditional deduplication algorithm to process the RS data is not so good. There are several key issues to be resolved, and we propose corresponding solutions as follows.

What Data to Process. For those data with high redundancy, deduplication processing can eliminate redundancy, which will save a lot of storage space. Nevertheless, if the data is not redundant, the result of deduplication processing is not so good. Besides, it will waste a lot of computing resources. The redundancy of data produced by satellite applications is unknown. Hence, we propose a similarity measure algorithm to analysis the data, and deduplicate those with high similarity. The similarity measure algorithm is based on the principle of locality and bloom filter algorithm, which will be discussed in Sect. 4.

How to Deduplicate the Data. We propose a deduplication algorithm which can detect the same part in byte level and encode the different part to patch files using delta encoding algorithm. Through experiments on remote sensing application data, we have achieved deduplication ratios up to 30:1, and analyzed how the chunksize affect the experiment results.

4 Design and Implementation

The overview of deduplication algorithm is shown in Fig. 3. The file is split into several chunks base on chunking algorithm. Data similarity analysis algorithm find out the similar data to deduplicate. Delta encoding algorithm encode the chunks into libs and patches. Then, chunk libs and chunk patches are stored to disk, the index is updated.

As is depicted in Fig. 4, the flow of deduplication algorithm is consist of three steps, chunking, data similarity analysis and delta encoding. The first step, chunking, is to split the file into several chunks. The chunksize is related to the experiment results,

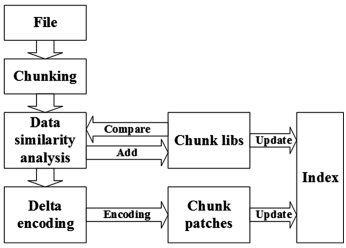


Fig. 3. Overview of deduplication algorithm

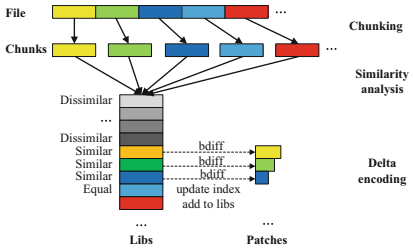


Fig. 4. Flow of deduplication algorithm

which will be discussed in Sect. 4. The second step, similarity measure, is to find out the similar data in all of the data set. If the chunk is similar to a chunk in the chunk libs, it will reach the third step. If the chunk is equal to a chunk in the chunk libs, the chunk libs index will be updated. If the chunk is not similar to all the chunks in the libs, it will be added to the chunk libs. The third step, delta encoding, is to encode the similar data base on bdiff algorithm.

4.1 Data Similarity Analysis Algorithm

Chunking. Chunking phase can be classified into several categories: Whole file chunking (WFC), Fixed-size partition (FSP), and Content-defined Chunking (CDC). FSP simply splits data into equal chunks that are independent of the content of the data being stored. Because the larger the chunksize is, the higher bdiff algorithm time complexity is. We use FSP algorithm to split the file into several chunks. Sliding Block chunking(SBC) is a improvement of CDC algorithm. As is depicted in Fig. 5, SBC divides files into fixed size and non-overlapping chunks and calculates its signatures. This approach normally performs a finer granularity matching. So we use SBC algorithm to compute bloom filter vector.

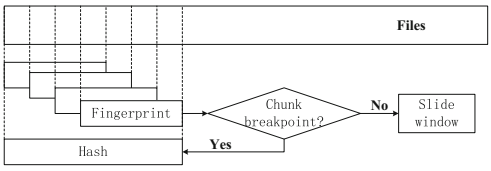


Fig. 5. SBC algorithm

Data Similarity Analysis. Bloom filter [9] is a simple space-efficient randomized data structure for representing a set in order to support membership queries [10]. We implementation similarity analysis based on bloom filter. For example, we compute

two m-bit long vector for chunk1 and chunk2. Then we get two bloom filters representing sets c1 and c2 with the same number of bits and using the same hash functions. A bloom filter that represents the union of two sets can be obtained by taking the OR of the two bit vectors of the original bloom filters. Accordingly, the inner product of the two bit vectors is a measure of their similarity. The similarity of chunk1 and chunk2 is therefore:

$$\mathrm{Sim}(c1, c2) = \frac{|\mathrm{fingerprints}(c1) \cap \mathrm{fingerprints}(c2)|}{|\mathrm{fingerprints}(c1) \cup \mathrm{fingerprints}(c2)|}$$

4.2 Delta Encoding

The Fundamental of Delta Encoding. As is depicted in Fig. 6, we use delta algorithms to encode one similar file in terms of another. The patch file is much smaller that the original file, while can save a lot of storage resources.

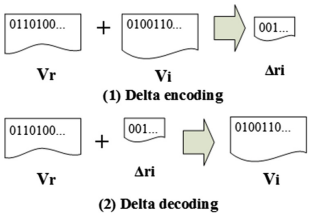


Fig. 6. Delta operation

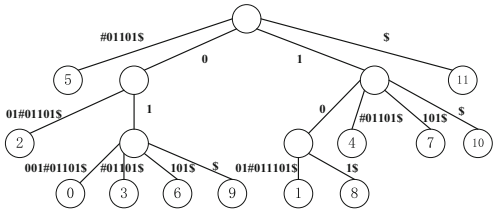


Fig. 7. The suffix tree for 010011#

Bdiff Algorithm. Bdiff algorithm [11] is an implementation of delta encoding. It builds a suffix tree for the first file. This tree search substrings of the second file to find matches in the first file. As an example, let S = 010011, and T = 011011. The suffix tree for S#T$ is shown in Fig. 7. Each arc in a suffix tree is associated with a substring of S#T$. To find Δs(T), start by tracing the string S#T$ in the suffix tree, beginning with the root. Finding the deepest non-leaf node, and the leaf node of the node has both # and $. Then, the traced path is part of the longest common subsequence.

4.3 The Dynamic Update of Chunk Libs

Data reference locality principle is a term for the phenomenon in which the same values, or related storage locations, are frequently accessed, depending on the memory access pattern. Figure 8 shows the dynamic update of chunk libs. In the implementation of similarity analysis, we update the chunk libs dynamically based on the principle of locality, which can significantly improve the efficiency of similarity analysis. In the experiment, we set the capacity of chunk libs to 30. The new chunk is only compared

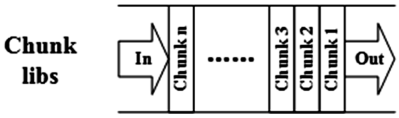


Fig. 8. Dynamic update of chunk libs

with the latest 30 library chunks. We will research the relationship between the capacity of chunk libs and deduplication ratios in future work.

5 Experiments and Discussion

We select "Data Collection of Global Lake Variations in 2014" from the National Integrated Earth Observation Data Sharing Platform to experiment and analysis. The remote sensing data file is tiff format. The amount of data is around equal to 1 GB.

As is shown in Fig. 9, with the increase of chunksize, deduplication ratios increases rapidly first and then decreases. When the chunksize is 1 MB, the deduplication ratios are maximized up to nearly 30:1. Because the smaller the particle size of chunk is, the greater the proportion of chunks whose similarity is above the threshold to the total chunks is. Therefore, more chunks will be added into the libs, which will result in the growth of libs' size and decrease of deduplication ratios. With the continuous increase of chunksize, fewer chunks will be added to the libs, but the corresponding patches' size will be bigger. The deduplication ratios will also decrease.

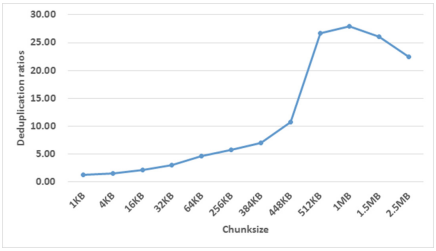


Fig. 9. The relationship between chunksize and deduplication ratios

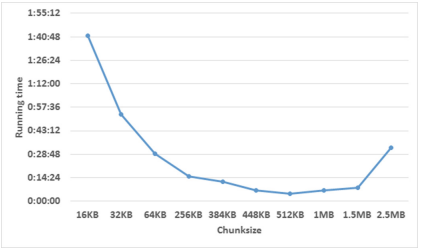


Fig. 10. The relationship between chunksize and program running time

Figure 10 shows the relationship between chunksize and program running time. With the increase of chunksize, running time decreases rapidly first and then increases. This is because the smaller particle size of chunk is, the more time will be consumed in chunking and similarity analysis. With the continuous increase of chunksize, delta encoding will consume a lot of time. The program running time will also increase. Generally speaking, the program running time of the algorithm is longer than that of the traditional algorithm, and it is better to use post process deduplication [2] method.

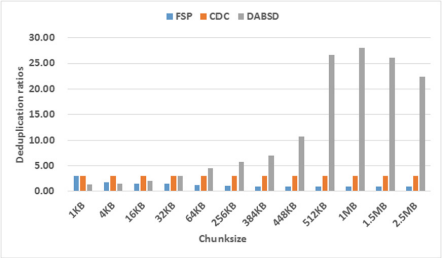


Fig. 11. The deduplication ratio compared with traditional deduplication algorithm

Figure 11 shows the experiment results of our algorithm comparing with the traditional deduplication algorithm. DABSD is short for deduplication algorithm based on similarity analysis and delta encoding. We choose FSP and CDC algorithm to conduct the comparative experiment. With the increase of the chunksize, deduplication ratios decrease using FSP algorithm, the biggest deduplication ratios are up to 3.7:1. The experiment results of CDC algorithm is nothing to do with the chunksize, the deduplication ratios are nearly 2.95:1. When the chunksize is small, traditional deduplication algorithm is better, but the deduplication ratio is still not high. With the increase of the chunksize, our algorithm is much better than traditional methods.

6 Conclusion and Future Work

In this paper, we present a deduplication algorithm for massive satellite application data storage. In the approach, we design a data similarity analysis algorithm and dynamic update strategy to measure similar file chunks. Afterwards, using bdiff algorithm to process redundant data. We also analyze the relationship between chunksize and the deduplication ratios. Overall, this approach is a viable way for massive satellite application data storage. In future work, more research works could be focused on the relationship between the capacity of chunk libs and deduplication ratios, and how to integrate the algorithm into file system.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant No. 61370059, the National Natural Science Foundation of China under Grant No. 61232009, Beijing Natural Science Foundation under Grant No. 4152030, the fund of the State Key Laboratory of Software Development Environment under Grant No. SKLSDE-2016ZX-13, the Open Research Fund of The Academy of Satellite Application under Grant No. Y20A-E03 and the Open Project Program of National Engineering Research Center for Science & Technology Resources Sharing Service (Beihang University).

References

1. Wang, L., Ma, Y., Zomaya, A.Y., et al.: A parallel file system with application-aware data layout policies for massive remote sensing image processing in digital earth. IEEE Trans. Parallel Distrib. Syst. **26**(6), 1497–1508 (2015)

2. Meyer, D.T., Bolosky, W.J.: A study of practical deduplication. ACM Trans. Storage (TOS) **7**(4), 14 (2012)
3. Rivest, R.: The MD5 message-digest algorithm. RFC Editor (1992)
4. Eastlake 3rd, D., Jones, P.: US secure hash algorithm 1 (SHA1) (2001)
5. Manogar, E., Abirami, S.: A study on data deduplication techniques for optimized storage. In: 2014 Sixth International Conference on Advanced Computing (ICoAC), pp. 161–166. IEEE (2014)
6. Bobbarjung, D.R., Jagannathan, S., Dubnicki, C.: Improving duplicate elimination in storage systems. ACM Trans. Storage **2**(4), 424–448 (2006)
7. Kruus, E., Ungureanu, C., Dubnicki, C.: Bimodal content defined chunking for backup streams. In: FAST, pp. 239–252 (2010)
8. Manogar, E., Abirami, S.: A study on data deduplication techniques for optimized storage. In: 2014 Sixth International Conference on Advanced Computing (ICoAC), pp. 161–166. IEEE (2014)
9. Bloom, B.H.: Space/time trade-offs in hash coding with allowable errors. Commun. ACM **13** (7), 422–426 (1970)
10. Broder, A., Mitzenmacher, M.: Network applications of bloom filters: a survey. Internet Math. **1**(4), 485–509 (2003)
11. Hunt, J.J., Vo, K.P., Tichy, W.F.: An empirical study of delta algorithms. In: Sommerville, I. (ed.) SCM 1996. LNCS, vol. 1167, pp. 49–66. Springer, Heidelberg (1996). doi:10.1007/BFb0023080

Area Constrained Space Information Flow

Alfred Uwitonze[1], Jiaqing Huang[1(✉)], Yuanqing Ye[2], and Wenqing Cheng[1]

[1] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, People's Republic of China
jqhuang@mail.hust.edu.cn
[2] Department of Electrical and Computer Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

Abstract. Departing from *Network Information Flow* (NIF) that studies network coding in graphs, *Space Information Flow* (SIF) is a new research direction that studies network coding in geometric space, such as Euclidean space. This work focuses on the problem of *Area Constrained Space Information Flow* (ACSIF), which is a more practical branch of SIF that considers the constraint on the area where the candidate relay nodes should be placed. One of the key open problems in ACSIF is to design the algorithms that compute the min-cost in multicast networks. This work proposes the first heuristic algorithm based on Delaunay Triangulation (DT) and Linear Programming (LP) techniques that can build a min-cost multicast communication network for N ($N \geq 3$) given terminal nodes in 2-D Euclidean space by taking into account the constrained area enclosed in circle with radius R around the terminal nodes. The proposed algorithm has a polynomial computational complexity and the simulation results show that it is effective.

Keywords: Space information flow · Network coding in space · Network information flow · Network coding · Delaunay triangulation

1 Introduction

Network Information Flow (NIF) [1] studies *network coding in graphs* and was proposed in 2000. Comparatively to NIF, *Space Information Flow* (SIF) [2] studies *network coding in geometric space* and was proposed in 2011. Although both SIF and NIF use network coding as their fundamental technique to transmit information, SIF allows additional set of relay nodes to be introduced to connect a set of given terminal nodes at any location within the network, something which is not accepted by NIF. SIF is also different from the Euclidean Steiner Minimal Tree (ESMT), with ESMT being the optimal routing in space. In network coding in geometric space model, we use *relay* nodes to distinguish from *Steiner* nodes that is adopted by routing in space. An ESMT can connect a set of given terminal nodes in a Euclidean space with free introduction of extra set of *Steiner* nodes. SIF is also different with Minimum Spanning Tree (MST) in that MST interconnects all the terminal nodes of a given set with a shortest possible network of direct links, without any additional relay node [3].

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 254–259, 2017.
DOI: 10.1007/978-981-10-3969-0_29

The goal of SIF is to minimize the total bandwidth-distance sum product ('network volume'), while sustaining given end-to-end communication rates. The pentagram example [4] is the first single multicast SIF to demonstrate that the performance of SIF can be strictly better than that of ESMT, with the *Cost Advantage* (CA) [2] being strictly bigger than 1, where CA is defined as the ratio of minimum cost necessary for achieving a target throughput by routing over that of network coding, under the natural assumption that the cost of a link in a network is proportional to its length as well as its information flow rate.

Regarding the applications of SIF, a potential scenario is the planning of wireless sensors in space [5].

With regard to routing in space, the properties of optimal ESMT were studied by Gilbert and Pollak [6]. According to Van Laarhoven [7], the computational complexity of ESMT is NP-Hard. In line with SIF, Li and Wu [2] studied the problem of multiple-unicast network coding in space. Yin *et al*. [8] proved the upper-bounds on the number of relay nodes required for SIF. Xiahou *et al*. [9] proposed a geometric space framework to analyze the multiple-unicast network coding conjecture in undirected graph. Hu *et al*. [10] proposed a heuristic algorithm based on iterative technique to solve the problem of min-cost video multicast via constrained SIF. Huang *et al*. [4] proposed the first heuristic algorithm to compute the optimal SIF solutions, for the case of a single multicast. In a subsequent work, Huang and Li [5] proposed a heuristic SIF algorithm based on non-uniform recursive space partitioning to compute SIF. As for MST, its computational complexity is polynomial [3].

However, the problem of Area Constrained Space Information Flow (ACSIF) has not been investigated in previous SIF research works. To the best of our knowledge, this is the first work to explore the problem of ACSIF. The main contribution of our paper can be summarized as follows: We propose the first heuristic algorithm with a polynomial computational complexity that uses SIF to compute the min-cost in multicast networks, as well as the corresponding topology, under the restricted area to the candidate relay nodes placement.

The rest of this paper is organized as follows: The problem formulation and definitions are discussed in Sect. 2. Section 3 describes the detailed steps of the new heuristic algorithm for ACSIF. Section 4 presents the simulation results, while Sect. 5 concludes the paper.

2 Problem Formulation and Definitions

Although our idea can be extended to a general space case, in this work we focus on the problem of min-cost multicast network coding in 2-D Euclidean space. Given N ($N \geq 3$) terminal nodes T_1, T_2,…, T_N with coordinates in a 2-D space, the restricted area to candidate relay nodes placement enclosed in circles, each with radius R, that are located around the terminal nodes and a multicast session from one source node O to the N terminals as sinks. The goal is to construct a min-cost multicast network using SIF, which allows to introduce extra relay nodes under the given restricted area to relay nodes placement. We define the total cost of the network as $\sum_{uv \in E} w(uv)f(uv)$, where $f(uv)$ is the information flow rate of a link uv in space, and $w(uv)$ is the weight of a link

uv, which equals to the Euclidean distance $\|uv\|$ of uv [2]. The positions of the relay nodes and the flow rate assignments on the connection links are the two factors that determine the network cost in general. We call these two variables *positions* and *flow rate assignments*. We assume that the connection links of two optimal relay nodes overlap with the restricted area. The *positions* of terminal nodes are fixed, while the *positions* of relay nodes are not. The flow rate assignments will also determine the connection topology of all nodes, since a link with a zero flow rate indicates that the link does not exist. Our objective is to achieve the minimum cost by tuning these two sets of variables while taking into account the restricted area to the candidate relay nodes placement.

The area A_C enclosed by a circle is defined as $A_C = \Pi R^2$, where the radius R is the distance from the center of a circle to a point on the circle.

3 The Proposed Heuristic Algorithm for ACSIF

The aim of our algorithm is to use additional relay nodes to establish a min-cost multicast network connection from *N (N ≥ 3)* given terminal nodes in 2-D Euclidean space, under the restricted area to the candidate relay nodes placement. The proposed algorithm suggests alternative positions of optimal relay nodes whose connection links overlap with the restricted area.

3.1 Detailed Description of ACSIF Algorithm

Our algorithm uses two key techniques: DT and LP. DT helps to generate at most 2*N*-5 DT triangles from *N (N ≥ 3)* given terminal nodes [11]. It then helps to obtain the possible candidate relay nodes from all generated DT-triangles. LP is used to compute the minimum cost as well as the flow rates on the connection links.

Our algorithm adopts the following LP model:

Minimize $cost = \sum_{\overrightarrow{uv} \in A} w(\overrightarrow{uv}) f(\overrightarrow{uv})$

Subject to:

$$\begin{cases} \sum_{v \in V_{\uparrow}(u)} f_i(\overrightarrow{vu}) = \sum_{v \in V_{\downarrow}(u)} f_i(\overrightarrow{uv}) & \forall i, \forall u \\ f_i\left(\overrightarrow{T_i S}\right) = r & \forall i \\ f_i(\overrightarrow{uv}) \leq f(\overrightarrow{uv}) & \forall i, \forall \overrightarrow{uv} \\ f(\overrightarrow{uv}) \geq 0, f_i(\overrightarrow{uv}) \geq 0 & \forall i, \forall \overrightarrow{uv} \end{cases} \tag{1}$$

The LP model (Eq. (1)) is based on an undirected complete graph, denoted as *G = (V, E)*, where V is the set of N given *terminal* nodes and M additional *relay* nodes, while E is the set of undirected links. Due to the bi-directed possibilities of transmission in space, we make links bi-directed and denote a set of directed links as $A = \left\{\overrightarrow{uv}, \overrightarrow{vu} | uv \in E\right\}$. In the LP objective function, the decision variable $f(\overrightarrow{uv})$ represents the combined effective flow rate on a link $\overrightarrow{uv}$. The coefficient (i.e. weight) $w(\overrightarrow{uv})$ represents the Euclidean distance $|\overrightarrow{uv}| (= |\overrightarrow{vu}| = |uv|)$. In the LP constraints,

$f_i(\overrightarrow{uv})$ represents the rate of network information flow $S \rightarrow T_i$ on a link $\overrightarrow{uv}$, i.e. for every network information flow $S \rightarrow T_i$, there is a *conceptual* [12] flow $f_i(\overrightarrow{uv})$. We call it *conceptual* because different conceptual flows share instead of competing for available bandwidth on the same link [12]. The final flow rate $f(\overrightarrow{uv})$ of a link uv equals to the maximum among all $f_i(\overrightarrow{uv})$ and should be not less than the maximum conceptual rate, which will directly affect the total *cost*. We have both $f_i(\overrightarrow{uv})$ and $f_i(\overrightarrow{vu})$ to indicate the flows in two directions. $V_{\uparrow}(u)$ and $V_{\downarrow}(u)$ denote upstream and downstream adjacent set of u in V, respectively. We assume that there is a conceptual link from each sink T_i back to the source S with the rate r, for concise representation of flow conservation constraints [12]. The details about steps involved are shown in Algorithm 1.

Algorithm 1 A Heuristic Algorithm for ACSIF

Require: Input: *N (N ≥ 3) terminal nodes*, the constrained area enclosed in circle with radius *R*, a multicast session
Ensure: Output: An Area Constrained SIF solution
1: Initialize the total set of all possible candidate relay nodes $R_{total} = \emptyset$;
2: Construct all the DT-triangles by Delaunay Triangulation;
3: Initialize the subset of candidate relay nodes $R(m) = R'(m) = \emptyset$, MINICOST =+∞;
4: **for** $m = 1$ to 2, **do**
5: Construct the polygons of 3 and 4 edges by concatenating m adjacent DT-triangles;
6: Obtain all the possible candidate relay nodes R_{total};
7: Retain only the candidate relay nodes $R(m)$ that are located outside the constrained area;
8: Construct a complete graph with $(N + \sum_{m=1}^{m} |R(m)|)$ nodes;
9: Solve the LP model based on the complete graph and output MINCOST;
10: Keep only the candidate relay nodes $R'(m)$ that are located outside the constrained area and whose links do not overlap with the constrained area;
11: Construct the second complete graph with $(N + \sum_{m=1}^{m} |R'(m)|)$ nodes;
12: Solve the LP model based on the second complete graph and output an area constrained SIF *cost* and the network topology;
13: **end for**
14: **if** *cost* < MINCOST **then**
15: MINCOST=*cost*
16: **end if**
17: **if** The flow rates of all relay nodes that are located outside the constrained area and whose links do not overlap with the constrained area == 0 **then**
18: Output MINCOST and stop.
19: **end if**

3.2 Complexity Analysis of ACSIF Algorithm

The computational complexity of Delaunay Triangulation is $O(NlogN)$ [11]. The computational complexity of generating the candidate relay nodes from DT-triangles and quadrilaterals constructed by concatenating two adjacent DT-triangles is polynomial [13]. Given N ($N \geq 3$) terminal nodes, we can obtain at most $2N\text{-}5$ DT-triangles

by Delaunay Triangulation [11] and can then concatenate at most *N-2* adjacent DT-triangles and *2N-6* quadrilaterals. Thus, $|R_{total}| \leq 6N\text{-}16$, and the complexity of LP is $O((N + |R_{total}|)^2) = O((7N\text{-}16)^2) = O(N^2)$. Therefore, the total computational complexity of Algorithm 1 is $O(N^3 logN)$ and it is polynomial.

4 Simulation Results

We have simulated our heuristic algorithm in a 2-D Euclidean space. Our simulations used MATLAB to solve LPs. Our inputs consisted of 15 cases of 10 nodes data sets from OR-Library. We set the constrained area enclosed in 5 circles, each with radius $R = 0.040$. The optimal ESMT is computed by GeoSteiner 3.1 that implements an ESMT exact algorithm [13]. The MST is computed by implementing Prim's algorithm [3] in MATLAB. We inserted the 5 circles around the terminal nodes such that the connection links of two optimal relay nodes overlapped with the constrained area.

4.1 Cases of 10 Nodes Data Sets from OR-Library

We applied our algorithm to 10-points ($N = 10$) data sets from OR-Library [14], which contained 15 cases with different positions. Figure 1 shows the min-cost performance comparison between MST, ACSIF, SIF and optimal ESMT for all the 15 cases. As it can be seen from Fig. 1, ACSIF outperforms MST for all the cases. Both SIF and optimal ESMT achieve the same results. This can be attributed to the fact that SIF problem degrades into optimal ESMT problem when the flow rate f (uv) required is found to be always equal to 1 [8]. Furthermore, ACSIF performs very close to both SIF and ESMT. This shows the effectiveness of the proposed algorithm.

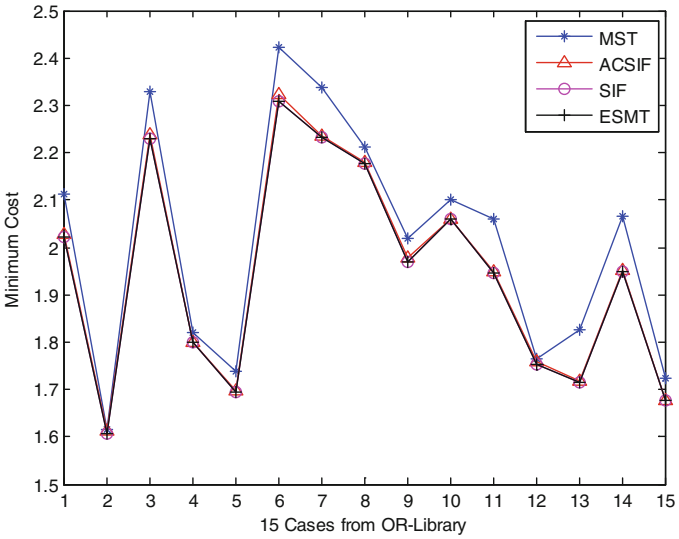


Fig. 1. Min-cost comparison between MST, ACSIF, SIF and optimal ESMT.

5 Conclusion

This paper presents a solution to the problem of Area Constrained Space Information Flow in single multicast networks by using a new $O(N^3 \log N)$ algorithm that takes into account the constrained area to the candidate relay nodes placement while computing the min-cost multicast for N ($N \geq 3$) given terminal nodes in 2-D Euclidean space. The algorithm suggests the alternative positions to the optimal relay nodes whose links overlap with the constrained area. The simulation results show that the proposed algorithm is effective.

Acknowledgments. This research was supported by National Natural Science Foundation of China (No. 61271227).

References

1. Ahlswede, R., Cai, N., Li, S.Y.R., Yeung, R.W.: Network information flow. IEEE Trans. Inf. Theory **46**(4), 1204–1216 (2000)
2. Li, Z., Wu, C.: Space information flow: multiple unicast. In: Proceedings of IEEE ISIT, pp. 1897–1901 (2012)
3. Prim, R.C.: Shortest connection networks and some generalizations. Bell Syst. Tech. J. **36** (6), 1389–1401 (1957)
4. Huang, J., Yin, X., Zhang, X., Du, X., Li, Z.: On space information flow: single multicast. In: Proceedings of IEEE NetCod, pp. 1–6 (2013)
5. Huang, J., Li, Z.: A recursive partitioning algorithm for space information flow. In: Proceedings of IEEE GLOBECOM, pp. 1460–1465 (2014)
6. Gilbert, E.N., Pollak, H.O.: Steiner minimal trees. SIAM J. Appl. Math. **16**(1), 1–29 (1968)
7. Van Laarhoven, J.M.: Exact and heuristic algorithms for the Euclidean Steiner tree problem. Ph.D. thesis, University of Iowa (2010)
8. Yin, X., Wang, Y., Wang, X., Xue, X., Li, Z.: Min-cost multicast networks in Euclidean space. In: Proceedings of IEEE ISIT, pp. 1316–1320 (2012)
9. Xiahou, T., Li, Z., Wu, C., Huang, J.: A geometric perspective to multiple-unicast network coding. IEEE Trans. Inf. Theory **60**(5), 2884–2895 (2014)
10. Hu, Y., Niu, D., Li, Z.: Internet video multicast via constrained space information flow. IEEE MMTC E-Lett. **9**(2), 17–19 (2014)
11. Smith, J.M., Lee, D.T., Liebman, J.S.: An O(n log n) heuristic for Steiner minimal tree problems on the Euclidean metric. Networks **11**(1), 23–39 (1981)
12. Li, Z.: Min-cost multicast of selfish information flows. In: Proceedings of IEEE INFOCOM, pp. 231–239 (2007)
13. Winter, P., Zachariasen, M.: Euclidean Steiner minimum trees: an improved exact algorithm. Networks **30**(3), 149–166 (1997)
14. Beasley, J.E.: OR-Library: distributing test problems by electronic mail. J. Oper. Res. Soc. **41**(11), 1069–1072 (1990)

Research on the Algorithm of Converting Files Generated by CALPOST to AVS/Express Platform

Xiaofei Shi[1(✉)], Yunfeng Ma[1], Qi Wang[1], Tingshuai Wang[1], Ping Wang[1], Shuai Wang[2], Xuezong Xu[3], Weike Xu[3], Zhongyi Wei[3], Nan Xiao[3], Caina Zhang[3], Xiaorui Ma[3], Yanwei Qian[3], and Kunyu Gao[4]

[1] College of Energy and Environment, Shenyang Aerospace University, Shenyang City, China
shixiaofei0707@163.com
[2] Shenyang Environmental Monitoring Centre, Shenyang City, China
rcdxph@126.com
[3] Smart Environmental Protection Department, CASIC Intelligence Industry Development Co., Ltd., Beijing City, China
weizhongyi@aiidc.com.cn
[4] Liaoning Province Environmental Monitor Centre, Shenyang City, China
tingshuai919@163.com

Abstract. The technology on visualizing the files generated by CALPOST efficiently and quickly is not so mature in environmental research. Based on the characteristic of the data stored in the CALPOST files, a CAL2AVS (namely CALPOST TO AVS/Express) algorithm was developed in the JAVA platform through Window 7 operation system. Finally, a test was taken in Shenyang City with two emission point sources which showed a better exhibition in rendering and quickness compared with other tools.

Keywords: Visualizing · CALPOST · CAL2AVS · Shenyang City · Rending

1 Introduction

The models focused on atmospheric environmental information abound, such as screen, AERMOD, CALPUFF, ADMS and so on. In China the government has recommended the CALPUFF as a guide in air quality environmental assessment according Guidelines for environmental impact assessment: Atmospheric Environment proclaimed on 31st, Dec., 2008 [1].

Many researches has adopted the CALPUFF air pollution dispersion model to carry out the air quality numerical simulation, such as: Gabriele Curci et al. adopted CALPUFF.

$$\left[X + \frac{(2n-1)\Delta x}{2}\right] * \left[Y + \frac{(2m-1)\Delta y}{2}\right] = Value_{mn}, (m, n \in N^*) \quad (1)$$

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 260–266, 2017.
DOI: 10.1007/978-981-10-3969-0_30

Model to study the minimum distance between the residents community and the source of pollution, concluded the safety minimum distance is 5 km [2]; CO2, NOX, CO and SO2 were simulated, in Abdul-Wahab's research [3]; Zou Xudong et al. used CALPUFF model to simulate the PM10 concentration distribution of the typical source in Shenyang City [4] and so on. But there are few researches on the visualization of the CALPOST files in rending and quickness. Wu Qunyong et al., through the GIS software, achieved the simulation results of the CALPUFF model to express the distribution of spatial and temporal [5]. Obviously, operating GIS software tool in the emergency management is not acceptable.

In this paper, an AVS/Express platform was introduced to show the distribution of air pollutant which was time-varying. Through this platform, the air quality condition can be quickly forecasting, therefore some measurements can be taken immediately to avoid the heavy pollution event.

2 Method

2.1 Files Generated by Calpost

The CALPOST program is a postprocessor designed to average and report concentration or wet/dry deposition flux results based on the hourly data containing in the files of CALPUFF output [6].

In order to get time-varying concentration of the air pollutant in CALPUFF output file, a control file CALPOST.INP was generated by the code we selected. After executed the CALPOST.INP in the command line, the files were generated. For example, on a DOS system of working inventory of CALPOST d:\…\CALPUFF\CALPOST.INP ('…' means working directory). Then, a number of files were obtained, each one contained the concentration value of echo cell we set and represented the distribution condition at the specified time.

Through analysis the characteristic of the concentration value, an equation was acquired as follow:

(1) where, X is the coordinate of left lower corner in longitude, to East is positive;

Y is the coordinate of left lower corner in latitude, to North is positive;
Δx represents the cell distance in longitude (km);
Δy represents the cell distance in latitude (km);
$Value_{mn}$ is the value of the cell of (m, n) ($\mu g/m^3$).

2.2 Brief Introduction of AVS/Express

The AVS/Express platform is a software that can allow researchers to apply the hardware power to their problems without requiring programming expertise or great investment of time. In a visualization application, modules play an important in function in the visualization cycle: (1) Filtering the basic data into a more usable form;

(2) Mapping the filtered data into either geometric primitives or mapping the data into pixels; (3) Rendering the 3D geometries or 2D images into pictures on the display screen.

In this platform, there are four grid data types can be handled. See Fig. 1:

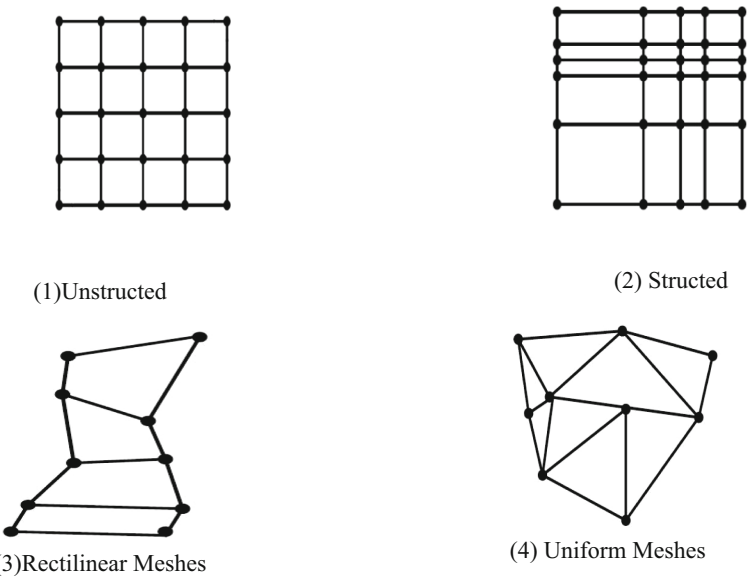


Fig. 1. Data types

In this paper, the concentration file data can be seen as Fig. 1(4).

The AVS/Express environment has three main interfaces: Network Editor; V language; Application Programmer Interfaces (APIs). In this paper, the Network Editor is chosen.

2.3 Achieve the CAL2AVS Algorithm by JAVA Platform

JAVA is a multi-platform developing language. In this research, Window 7 operation system was used to develop this algorithm. To compile the NetCDF format file, which can store time-varying variable data; in the Windows platform we need some additional compiled files, such as netcdf.dll, netcdf.lib, ncgen.exe, ncdump.exe and netcdf.exp. In this process, there are two necessary parts: one is the application of source packages, mainly MainFrmame.java and TransformFile.java included; the other is calling of library functions, mainly the libraries such as resource.jar, rt.jar, jsse.jar and jce.jar included. Part of the program code is as follows:

```
……
public TransformFile(double[] x, double[] y, List<Double> z,
File[] files) {
        this();
        this.x = x;
        this.y = y;
        this.z.addAll(z);
        this.files = files;
     public void getTime()
{for (int i = 0; i < files.length; i++)
      {String str = getTime(files[i].getName());
            if (times.contains(str) == false)
           {times.add(str);
           }
       }
       System.out.println(times.toString());
 }
……
```

3 Application Instance of Using the CAL2AVS Algorithm on Shenyang Air Pollutant Platform

3.1 Calpuff Running Information

In this research, the Shenyang City was taken for an example with this algorithm. The point (529.963 km E, 4625.475 km N) in fifty-first district was selected as the left lower corner coordinates of our simulation domain. In the horizontal direction, the resolution was 3 km, and the number of cell in the direction of the East and North were both 30; in the vertical direction, a total number of 9 layers was set, and the height of each level was 0 m, 20 m, 50 m, 100 m, 200 m, 400 m, 700 m, 1200 m and 1600 m. The two point emission sources were showed in Table 1:

Table 1. Two emission sources information

Name	X coordinate (km)	Y coordinate (km)	Height (m)	Exit temp. (K)	Emission rate (g/s)
P1	543.463	4638.075	45	383.15	150
P2	573.463	4678.075	58	383.15	200

3.2 Execute the Algorithm

- Input the grid information of the simulation domain. In this paper, the start point is (529.963 km, 4625.475 km), then add it to the number which is acquired by cell number 30 multiple distance 3 km, the end point (619.963 km, 4715.475 km) is obtained.

- Choose all the required files. From File menu, select Read item; then a dialog box is appeared, choose all the files generated by CALPOST; and then pressure the OK button. When a message 'Successfully Obtain Input File!' is exhibited that indicate this step is finished.
- Generate the CDL file. From File menu, choose Generate CDL File button, the process is running, when a message 'Successfully Generated CDL File!' is showed that indicated the CDL(Common Data form Language network) file which can be used to describe the NetCDF file data structure and specific the definition of the NetCDF object (dimension, variable and attribute) [7] is generated.
- Generated the NetCDF file. From File menu, choose Generate NetCDF File item, the process is running with about several seconds that depends on the number of files dealt with. When a message 'Successfully Generated NetCDF File!' is appeared, the file we need is acquired.

3.3 Display the NetCDF File in AVS/Express Platform

In the AVS/Express platform, a large number of application components are built in. In order to visualize the NetCDF file, there are three steps should be taken: (1) Find the corresponding templates; (2) Drive the templates in the network editor via the left key to generate sub objects; (3) Dataflow architecture. As shown in Fig. 2.

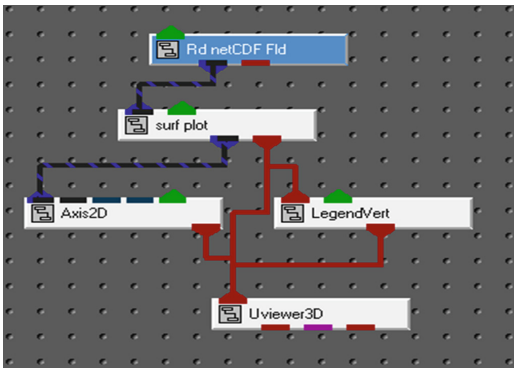


Fig. 2. Main instantiation chart

After the instantiation, the file generated in step 3.2 is introduced in SingleWindowApp interface by Rd netCDF Fld item. By slide the Step Time button, we can visual the distribution of the concentration at any time during the period quickly. Below are three pictures which show the distribution of air pollutant concentration at four different times, see Fig. 3:

From Fig. 3(1) we can see the maximum concentration value is about 13.33 $\mu g/m^3$, while in Fig. 3(2), Fig. 3(3) and Fig. 3(4) the maximum values are 5.73 $\mu g/m^3$, 3.87 $\mu g/m^3$ and 6.24 $\mu g/m^3$. Obviously, that is indicated the air pollutant concentration

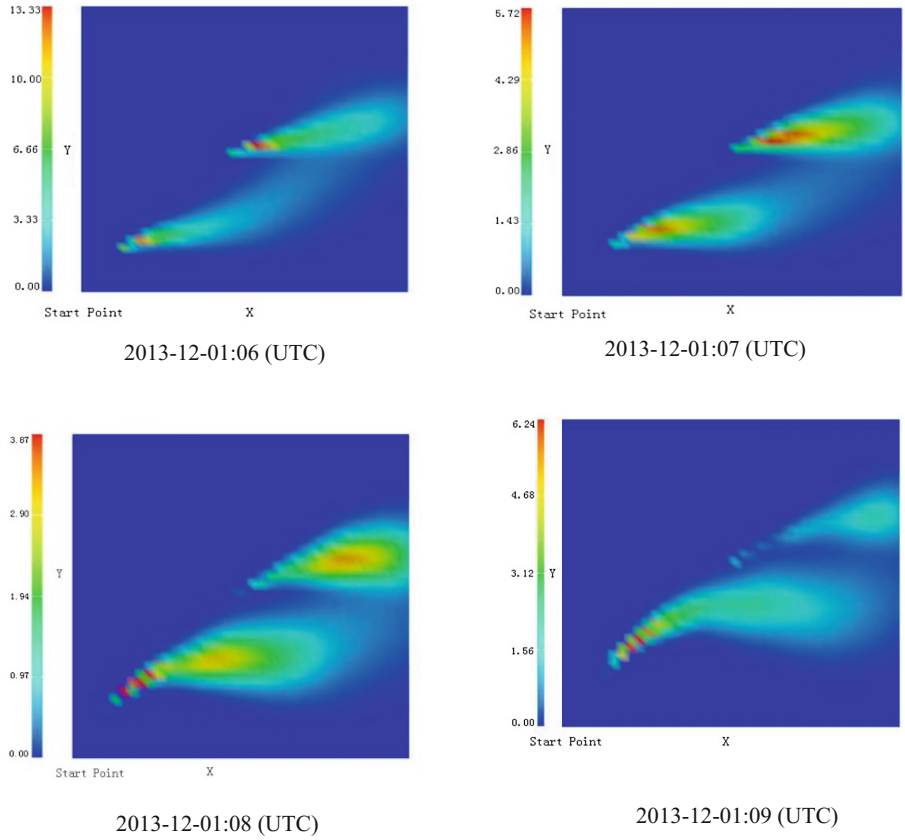


Fig. 3. Visualization application in concentration (μg/m^3) distribution

is diluted with time, but the area polluted is enlarging with time. However in Fig. 3(4) the maximum air pollutant concentration value reaches 6.24 μg/m^3 which is bigger than 3.87 μg/m^3 that indicates the wind direction is changing and perhaps the earth elevation obstacle the distribution. So, we can deduce the wind direction and the changing of wind direction easily.

4 Conclusions

Since the AVS/Express platform is a developing environment, which is mainly done visually, in combining and connecting modules in a LEGO-like fashion. While the CALPUFF model is an advanced non-steady-state meteorological and air quality modeling system which can simulate the air pollution condition. This algorithm is equivalent in transferring the number of files generated by CALPOST to one file without lose the original information. Through this platform, the air pollutant distribution situation can be quickly showed with wonderful rending effect. Also, in Fig. 2 a

changing concentration bar is showed which can get maximum value quickly, thus we can easily infer that whether further measurements should be taken or not. Due to the emergency air pollution event, this CAL2AVS algorithm will be used in daily business research.

References

1. HJ2.2-2008. Guidelines for environmental impact assessment: atmospheric environment. Ministry of Environmental Protection the People's Republic of China (2008)
2. Curci, G., Cinque, G., Tuccella, P., et al.: Modelling air quality impact of a biomass energy power plant in a mountain valley in Central Italy. Atmos. Environ. **62**, 248–255 (2012)
3. Abdul-Wahab, S.A., Obaid, J., Elkamel, A.: Modelling of greenhouse gas emissions from the steady state and non-steady state operations of a combined cycle power plant located in Ontarin, Canada. Fuel **136**, 103–112 (2014)
4. Zou, X., Yang, H., Zhang, Y., et al.: Distribution simulation analysis of PM10 concentration from typical sources of Shenyang in winter. Chin. J. Environ. Eng. **4**, 881–886 (2010)
5. Wu, Q., Huang, J., Sheng, L., et al.: Spatio-temporal and multi-dimensional visualization for the simulation result of CALPUFF model. J. Geo-Inf. Sci. **17**, 1–9 (2015)
6. Scire, J., Strimaitis, D., Yamartino, R.: A User's Guide for the CALPUFF Dispersion Model. Earth Tech. Inc., Concord (2000)
7. Rew, R., Davis, G., Emmerson, S., et al.: The NetCDF Users Guide, 2005–2009. University Corporation for Atmospheric Research (2011)

A Construction Method of Road and Residence Correlation Based on Urban Skeleton Network

Chuang Liu, Haizhong Qian(✉), Haiwei He, Xiao Wang, and Limin Xie

Zhengzhou Institute of Surveying and Mapping, Zhengzhou, China
Qianhaizhong2005@163.com

Abstract. For road and residence in large-scale urban map data are separated from each other, it is difficult to study their correlation. However, as road is tightly related to residence, it is necessary to deeply explore the connection and establish the clear correlation between them. In the paper, excellent characteristics of urban skeleton network were utilized to establish distinct correlation between road and residence and new comprehensive idea were provided for overall collaboration and integration of road and residence.

Keywords: Urban skeleton line · Coordination cartographic generalization · Correlation · Restraint Delaunay triangulation network

1 Introduction

Road and residence are two major elements in human productive activity and also main objects accounting for above 95% information amount in urban map. It shows that road integration and residence integration are also key points of map generalization. Currently, a majority of existing researches on automatic integration of road and that of residence only pay attention to either of road or residence for integration. However, it will inevitably destroy spatial collaboration relation among map elements through separable integration of single element, which results in variously spatial and semantical conflicts and also increases additional burden to artificial error correction. In the paper, a road and residence correlation construction method based on urban skeleton network was proposed and a model of correlation between road and residence was established to provide a new idea for collaboration and integration of road and residence.

Cooperative research on road and residence started later and in the early stage, it mainly focused on man-machine coordination that means integrated task was mutually accomplished by interaction between operator and computer. Currently, relevant research mainly put emphasis on local comprehensive processing for specific situation. For example, Pingtao Wang and Keshi Doihara utilized road network to divide construction into different block units to perform integration [1]; Patrick Revell and Nicolas Regnauld took advantage of triangulation network and Agent system to integrate building and road axis [2]; Qian Haizhong proposed block integration method based on dimension reduction technique and used characteristic of geometry spatial

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 267–275, 2017.
DOI: 10.1007/978-981-10-3969-0_31

complementary between large-scale urban buildings and streets to perform integration like combination, simplification and displacement on blocks according to skeleton line of architecture and street [3]; Deng Hongyan proposed a cartographic generalization inspection method based on design for quality and provided a new idea for connection between road and residence [4]; He Haiwei came up with a simplification strategy to avoid a conflict with residence in road simplification [5].

Above researches greatly promotes the development of collaboration and integration of road and residence. However, further research discovered that study of cartographic generalization mainly focuses on single-element integration and that integration in consideration of two elements was only to solve local conflict. Thus, it should make further comprehensive exploration on collaboration and automatic integration of road and residence. This paper links road and residence together that was uncorrelated with each other in map by utilizing urban skeleton network technology, which will lay a foundation for correlation between road and residence and coordination and integration of road and residence.

2 Overview of Urban Skeleton Network

Urban skeleton network [6] refers to use one-dimensional linear element to summarize 2D urban planar space. While simplifying urban space, urban skeleton network better maintains the morphological characteristics and distribution condition of urban space. As shown in Fig. 1, urban skeleton network consists of skeleton lines of road and blank area. Specifically, place with roads in map is directly deemed as urban skeleton network; Skeleton lines are extracted from the blank area except for road and residence in map. Skeleton lines from these two sources in urban skeleton network are respectively called road skeleton line and blank-area skeleton line.

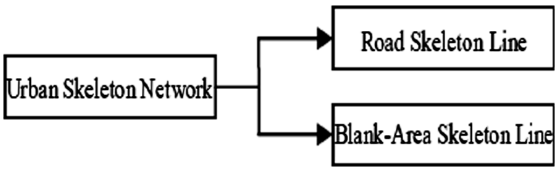


Fig. 1. Structure of urban skeleton line network

Urban blank area refers to the total area except for road and residence in map (Fig. 2). In particularly, hole in residence is regarded as interior zone of residence and belongs to residence rather than blank area. Blank area can be classified into area feature according to element type but it is special area feature which is distinctively different from common residence, plant, river system and other area features, equipped with characteristics as follows:

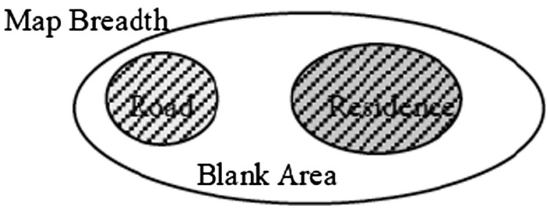


Fig. 2. Diagram of map composition

(1) The domain is usually large. Blank area is collection of all areas except for road and residence, and usually represents scope of one or more blocks, so its overage area is extensive and regional area is usually large;
(2) Interior has many holes. Blank area is area feature with hole and contains more holes than other area features with holes. Number of holes and involved number of residences are similar and outline of them is similar. For urban residence distributes unevenly in map with various types and patterns, holes in blank area distribute unevenly, with complicated outline and form.
(3) Complementary with road and residence. Blank area, road and residence jointly form the overall range of map and the former is complementary set for the latter two.

3 Construction Method of Urban Skeleton Network

It can be concluded from the definition and composition of urban skeleton network that the key to construction of urban skeleton network is extraction of skeleton line in blank area. Extracting quality of skeleton line in blank area determines whether it can obtain urban skeleton network that meets follow-up matching requirement.

In the paper, restraint Delaunay triangulation network was utilized to extract skeleton line from blank area. Delaunay triangulation network can sensitively capture detailed information from image. So it can satisfy characteristics including large coverage of bland area, many holes and irregularity.

Concrete steps for extraction of skeleton line from blank area as follows:

(1) Data Node Encryption

In order to establish restraint Delaunay triangulation network that can meet the requirement for extraction of follow-up skeleton line, it should conduct node encryption for data of road and residence before establishing network. Figure 3 shows the comparison before and after data node encryption of road and residence, and original nodes should be kept in the process of encryption and encryption is only conducted between two original nodes.

Similar triangle principle was utilized to solve encryption node coordinate. Set the first node coordinate N_1 in Encryption segment as (x_{start}, y_{start}), the end node coordinate N_2 as (x_{end}, y_{end}), and the pass point coordinate P as (x, y) in Fig. 4. According to similar triangles principle, it could be obtained that

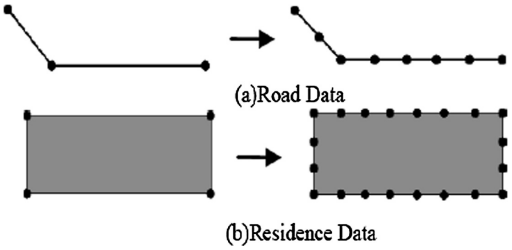


Fig. 3. Comparison of data node before and after encryption

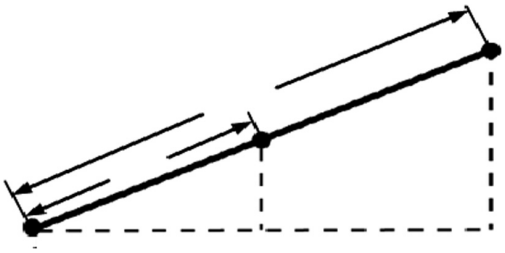


Fig. 4. Similar triangles principle

$$x = \frac{l}{D}(x_{end} - x_{start}) + x_{start} \tag{1}$$

$$y = \frac{l}{D}(y_{end} - y_{start}) + y_{start} \tag{2}$$

There into, l is encryption step (step setting takes the principle of correctly establishing Delaunay triangle network, i.e. established triangles being continuous but disjoint) and D is length of segmental arc N_1N_2. At the end of encryption process of segmental arc, in case that surplus distance is less than encryption step, following methods were applied in the paper to settle the problem: Before encryption, it should make judgement on distance from encryption node to end node: if $N_2P \geq 1.5 \cdot l$, encryption starts; if $N_2P < 1.5 \cdot l$, encryption ends.

(2) Construct Restraint Delaunay Triangle Network

After encrypting node of road and residence data, peripheral contour of road and residence was taken as constraint side (peripheral contour of road and residence should be side of triangulation network) to establish restraint Delaunay triangle network for all data gathered from residence and road. Establishment result as shown in Fig. 5.

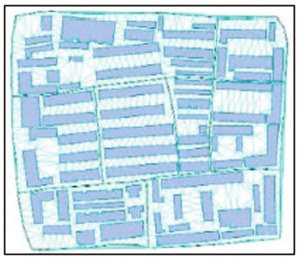


Fig. 5. Constructing constraint Delaunay triangulation network

(3) Extraction of Skeleton Line from Blank Area

After the establishment of restraint Delaunay triangle network, it could extract skeleton line from blank area according to triangle type (Fig. 6), and triangle is classified by quantity and position of road and residence. Concrete principles are as follows:

① Type i triangle: Three vertexes are all on the outline of residence.
② Type ii triangle: One vertex is on the road and the other two vertexes are on different outlines of residence.
③ Type iii triangle: One vertex is on the road and the other two vertexes are on the same outline of residence.
④ Type iv triangle: Two vertexes are on the road and the rest vertex is on the outline of residence.
⑤ Type v triangle: Three vertexes are all on the road.

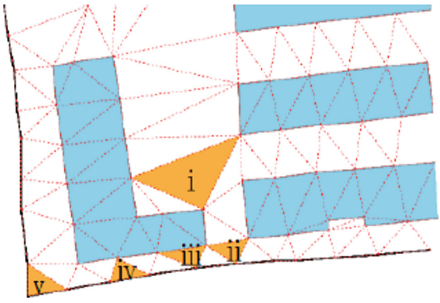


Fig. 6. Classification of triangle [6]

After triangle classification, skeleton lines were extracted from above 5 types of triangles [25]. For type i, it firstly made judgement on whether its three sides were restrained edge or not and then linked the midpoint of non-restrained edge with center of triangle (Fig. 6(a), (b) and (c)); for type ii, it directly linked the sole peak with the midpoint of its opposite edge (Fig. 6(d)). For iii, iv and v, no connection was made (Fig. 7).

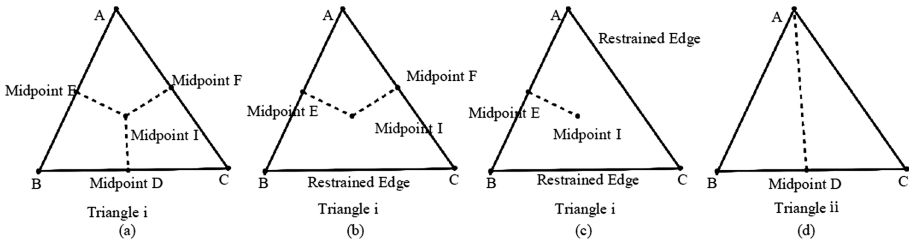


Fig. 7. Extraction of skeleton line in blank area [6]

It could obtain skeleton lines of all triangles through connection by this method from blank area, as shown in Fig. 8; urban skeleton network was formed by combining skeleton line in blank area and road network, as shown in Fig. 9.

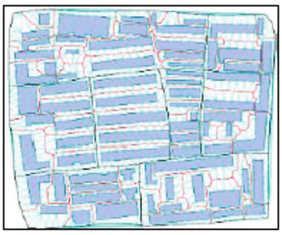


Fig. 8. Extraction result of skeleton lines of blank area

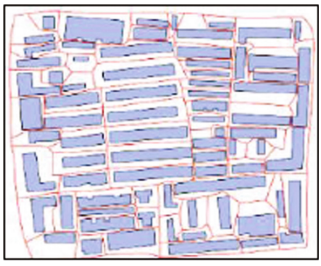
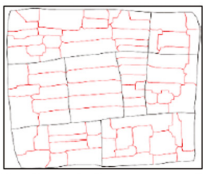
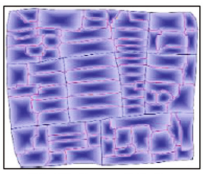


Fig. 9. Urban skeleton line network

(4) Establishment of Urban Skeleton Network Mesh

In road network, "road mesh" refers to the smallest closed block formed by crisscross road network [7]. From the perspective of geometrical features, road mesh is polygon formed by crisscross road essentially. Similarly, in urban skeleton network, the smallest closed block enclosed by crisscross skeleton lines is called "skeleton line mesh", as shown in Fig. 10. Just as road mesh, skeleton mesh is not real geographical object but area element object in order to better understand the architectural features of skeleton network.


(a)Urban Skeleton Network


(b) Skeleton Mesh

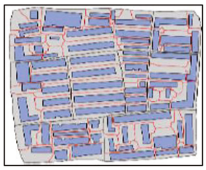

(c) Skeleton Mesh and Residence

Fig. 10. Establishment of skeleton mesh

After establishment of mesh structure of urban skeleton network, it made a summary for characteristics of skeleton mesh, as shown below:

(1) Skeleton mesh is area feature and its outline is composed of urban skeleton lines. For urban skeleton network consists of road (or boundary) and skeleton line in blank area, so outline elements of skeleton mesh include road (or boundary) and skeleton line in blank area;
(2) Each skeleton mesh only includes one residence element, i.e. corresponding relationship between skeleton mesh and involved residence;
(3) Skeleton meshes of common edge are connected with each seamlessly. Especially, there is topological relation between two skeleton meshes on common edge.

4 Correlation Establishment of Road and Residence

After the establishment of skeleton network and its mesh, road (or boundary), skeleton line, skeleton mesh show the topological relationship in Fig. 11.

(1) Correlation between road and skeleton line is equivalent. It means road must be skeleton line and a part of skeleton line is equal to road;
(2) There is incidence relation between skeleton line and skeleton mesh. It means that a skeleton line associates with one or two skeleton meshes and one skeleton mesh associates with three skeleton lines at least;
(3) Skeleton mesh and residence are presented as inclusion relation. It means that one skeleton mesh only contains one residence and vice versa.

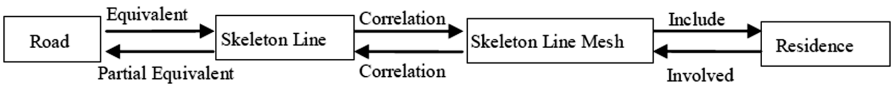
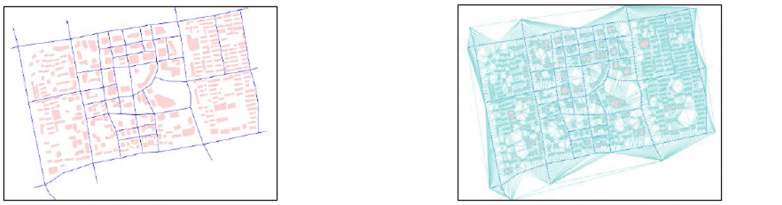


Fig. 11. Correlation diagram of road and residence

It can be concluded from the structure diagram of this topological relation that correlation can be established for road and residence in discrete state without direct relation by associating with skeleton mesh.

5 Experimental Analysis

Road and residence of somewhere in Zhongguancun, Beijing was taken as the experimental data and it was as shown in Fig. 12 after projection transformation and error correction.



(a) Original Data (b) Establishment of Restraint Delany Triangulation Network

Fig. 12. Data pre-processing

Similar triangles were utilized to conduct encryption processing for data node to establish constraint Delany triangulation network and the result was as shown in Fig. 12(b). Afterwards, skeleton lines in blank area were extracted to obtain the combination of skeleton line and road in bland area, i.e. urban skeleton line, as shown in Fig. 13. Topological relation between road and residence was established through road, skeleton line, skeleton mesh and residence, in hope of providing new idea for follow-up collaboration and integration of road and residence.

Fig. 13. Urban skeleton network

6 Conclusion

Advantage of collaboration and integration between road and residence is that it can avoid potential space conflict in the case of single-element integration, provide complete reference information for conflict area and make integration scientific and effective to a certain degree. In the paper, correlation between road and residence is established based on urban skeleton line to link road and residence in a state of separation in map and make road and residence learn from each other and restrain each

other in the process of integration. Finally, it can improve automation and intelligence level of automated generalization of map, in hope of solving conflict between road and residence caused by integration from the root.

Acknowledgment. The work described in this paper was supported by the Project of National Natural Science Foundation of China (Numbers: 41171305; 41571442).

References

1. Wang, P., Doihra, T.: Automatic generalization of road and buildings: ISPRS Congress Istanbul 2004. In: Proceedings of Commission IV, Istanbul (2004)
2. Revell, P., Regnauld, N., Thom, S.: Generalising OS MasterMap, topographic buildings and ITN road centerlines to 1:5000 scale using a spatial hierarchy of agents, triangulation and topology. International Cartographic Conference, A Caruna, Spain (2005)
3. Qian, H., Wu, F., Zhu, K., et al.: A generalization method of street block based on dimension-reducing technique. Acta Geodaetica Cartogr. Sin. **36**(1), 102–107 (2007)
4. Deng, H.: A study of automated cartographic generalization based on design for quality. Information Engineering University, Zhengzhou (2006)
5. He, H., Qian, H., Wang, X., et al.: Avoiding special conflicts in road simplification by using road bends. Acta Geodaetica Cartogr. Sin. **45**(3), 354–361 (2016)
6. Wang, X., Qian, H., He, H., et al.: Matching multi-source areal habitations with skeleton line mesh of blank region. Acta Geodaetica Cartogr. Sin. **44**(8), 927–935 (2015)
7. Hu, Y., Chen, J., Li, Z., et al.: Selective omission of road features based on mesh density for digital map generalization. Acta Geodaetica Cartogr. Sin. **36**(3), 351–357 (2007)
8. Liu, H., Qian, H., Wang, X., et al.: Road networks global matching method using analytic hierarchy process. Geomat. Inf. Sci. Wuhan Univ. **40**(5), 644–650 (2015)
9. Huang, Z., Qian, H., Guo, M., et al.: Matching algorithm of polygon habitations based on their skeleton-lines using fourier transform. Acta Geodaetica Cartogr. Sin. **42**(6), 913–921 (2013)
10. Hua, Y., Wu, S., Zhao, J.: Principal and Method of GIS. Beijing, 30–31 (2000)

A Hybrid Parallel Computing Model to Support Scalable Processing of Big Oceanographic Spatial Data

Miaomiao Song[1(✉)], Wenwen Li[2], Wenqing Li[1], Enxiao Liu[1], and Dingfeng Yu[1]

[1] Institute of Oceanographic Instrument Shandong Academy of Sciences, Qingdao, China
songmiaomiao_2006@126.com, livenson@163.com, xiaoyu_rs@163.com, liuexhit@gmail.com

[2] GeoDa Center for Geospatial Analysis and Computation School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, USA
Wenwen@asu.edu

Abstract. Oceanographic sciences are facing big challenges due to the deluge of big data. As of 2010, the amount of new data stored in the world main countries, led by the US, has grown over 7 exabytes. Although the computer hardware is quickly evolving, with faster processor frequency, multi-core technology, and larger memory, traditional reprocessing paradigm on a single-desktop basis still suffers from significant limitations in its low computational efficiency and scalability. In this paper, we report our effort in developing a hybrid parallel computing model which utilizes Graphic Processing Unit (GPU) to accelerate Hadoop Map Reduce system. In each computing node, the actual reprocessing is offloaded from a CPU to a GPU to further boost up the system performance. We describe the architecture design of the proposed model and the automated task/data assignment on each GPU-enabled compute node. Electronic Navigational Charts in ocean fields involves a huge amount of spatio-temporal data. Reprojection of these data between different coordinate reference systems, which is a computation-intensive task, is selected as the use case. Systematic experiments were conducted to demonstrate the good performance of the proposed model.

Keywords: Parallel computing · Hadoop MapReduce · GPU general computing · Oceanographic spatial data · Coordinate projection

1 Introduction

Oceanographic sciences are facing big challenges due to the deluge of big data. The development of aerospace and aviation technologies and the continuous improvement of the measurement instruments has produced tremendous amount of spatial data which digitally represents continuous or discrete ocean phenomena. As of 2010, the amount of new data stored in the world main countries, led by the US, has grown over 7

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 276–285, 2017.
DOI: 10.1007/978-981-10-3969-0_32

exabytes. In particular, the earth observing data produced per day by NASA'S Earth Science missions has already reached several terabytes [1]. Oceanographic spatial data has also been increasing at a rapid pace, resulting in the increasing complexity of data storing, searching, processing and analysis that extends far beyond the capability of traditional spatial computing technologies [2].

How to efficiently process these physically-distributed big data is an important issue in both GIScience and marine science. In this paper, we report our effort in developing a hybrid parallel computing model to realize scalable processing of big oceanographic spatio-temporal data, by utilizing General Purpose GPU (GPGPU) [3] to accelerate Hadoop MapReduce system [4]. We take the reprojection of massive spatio-temporal data, which are heavily used in depicting Electronic Navigational Charts (ENCs) [5] in ocean fields, as the use case. Based on our proposed parallel computing model, data processing time can be substantially reduced.

The rest of the paper is organized as follows: Sect. 2 introduces the architecture of parallel computing model aggregating GPUs and the Hadoop MapReduce framework. Section 3 introduces the mapping strategy between GPU threads and geospatial entities. Section 4 demonstrates the experiments and results. Section 5 concludes the paper and discusses directions for future research.

2 Architecture of Parallel Computing Model

Figure 1 demonstrates the architecture of the hybrid parallel computing model, which is designed as a dual-parallel model using GPU to accelerate the Hadoop MapReduce distributed computing platform. This proposed architecture consists of three primary components: Hadoop MapReduce distributed computing model, Hadoop Distributed File System (HDFS) [6] and GPU high-performance computing nodes.

The dual-parallel model is deployed in a distributed Hadoop cluster environment. In this model, Hadoop Map Reduce is adopted to aggregate distributed computing resources which work collaboratively in the Master-Slave mode [7]. Basically, it contains two types of computing resources: the Name node and Data node. Name node plays the role of Master and the data node plays the role of Slave. Each data node is built as a hybrid CPU and GPU platform to perform assigned computing tasks. Name node is responsible for maintaining the list of hosts in the cluster, information about segmentation of the computing tasks, data partition, global configuration and tracking jobs, as well as monitoring job status through a job tracker. The GPU-accelerated Hadoop Map Reduce parallel model is integrated into the mapper module located within each data node, so the mapper module contains the control procedures on CPU (Central Processing Unit) and high-speed parallel computation on GPU. Compared with CPU, GPU has more processing units, and is suitable to perform a huge number of simple uniform computational operations, but its capacity of controlling logic processing is much poorer than that of CPU. Therefore, we established the controlling program on CPU for data organization, indexing, data allocation, scheduling of GPU blocks. The GPU module is dedicated to the execution of highly intensive and complex geo-processing, such as the coordinate reprojection in our use case. HDFS plays a very important role on making data shared and accessible in the distributed environment.

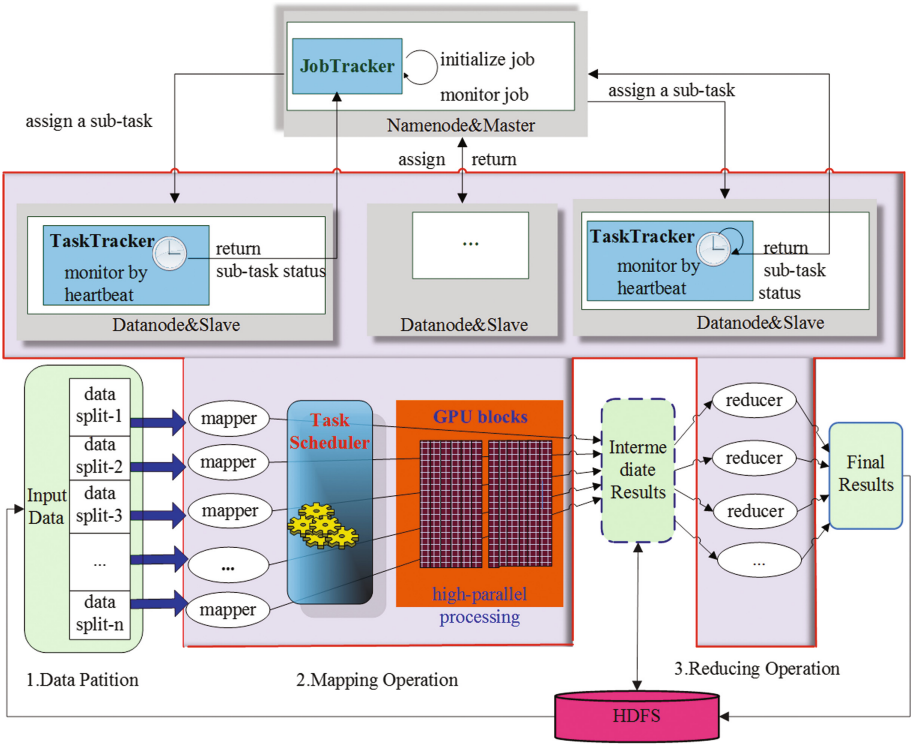


Fig. 1. The architecture of the dual parallel computing model

Every data block is stored in HDFS. All data nodes read original data from HDFS and write intermediate data or final results into HDFS.

In the process of computation, sub-tasks and data blocks will be assigned to each data node. Once computing tasks arrive at a data node, a mapper module or reducer module will be launched to execute computation. Meanwhile, a task tracker at data node polls the status of sub-tasks to name node at the heartbeat frequency. Eventually, according to these real-time status information, name node provides centralized management and scheduling of all computational jobs.

Practically, the dual-parallel model adopts two levels of parallelism, large-scale parallelism and fine-grained parallelism. The former is handled by Hadoop Map Reduce framework, and the latter is handled by GPUs. The workflow for the parallel data processing is briefly summarized as follows:

(a) The client submits a data processing request to the name node.
(b) Name node splits data into sub-blocks and assigns them to each data node.
(c) Every data node processes data blocks in the mapper module in parallel. Inside the mapper module, GPU kernel functions are invoked to implement the projection transformation.

(d) Once the mapper operation is completed on data node, intermediate result is generated and written back into HDFS, Task tracker sends a message indicating the task is finished to name node.
(e) When all mapper operations are finished, the reducer module will be launched to read intermediate data on each data node.
(f) When reducer operations on all data nodes are completed, the final results will be injected into HDFS in the format of Key-Value Pairs (KVP).
(g) Name node returns a prompt of accomplishment to the client.

3 The Mapping Strategy Between GPU Threads and Geospatial Entities

The high parallelism of the above computing model is derived from its capability of handling bulk data with thousands of threads at one time. When calculation is performed in parallel, it is essential to map the index of the threads on each GPU block to one record or a set of oceanographic data records. We utilize two strategies: single-point mapping and vertex-set mapping to map vertexes of oceanographic data into grid Index block index and tread index of the GPU to achieve high parallelism.

3.1 Single-Point Mapping

In the single-point mapping, a point coordinate (i.e. a vertex) is assigned to a GPU thread. This mapping process requires the creation of computing threads to be slightly more than or equal to the sum of vertexes of all geospatial entities. Once determined, the same number of threads should be launched in each GPU block. When the thread index is greater than the vertexes number, calculation is completed and threads are terminated. Single point mapping is suitable for the kind of calculation involving a single point coordinate or a single pixel as the handled object, such as coordinate projection and pixel-based calculation of raster data.

3.2 Vertex-Set Mapping

When the process involves the computation using multiple coordinates, such as the length calculation of a polyline, the single-point mapping approach is not suitable. Therefore, we designed the vertex-set mapping to assign multiple coordinates to one GPU thread.

Figure 2 shows the mapping between GPU threads and geospatial entities. In the application of oceanographic ENCs, the mapping takes place at two stages: coordinate reprojection and length calculation. During the stage of coordinate reprojection, single-point mapping is used and each vertex is handled by one GPU thread. For length calculation, the entire geospatial entity (a polyline or a polygon) is assigned to a GPU thread for length calculation based on the vertex-set mapping.

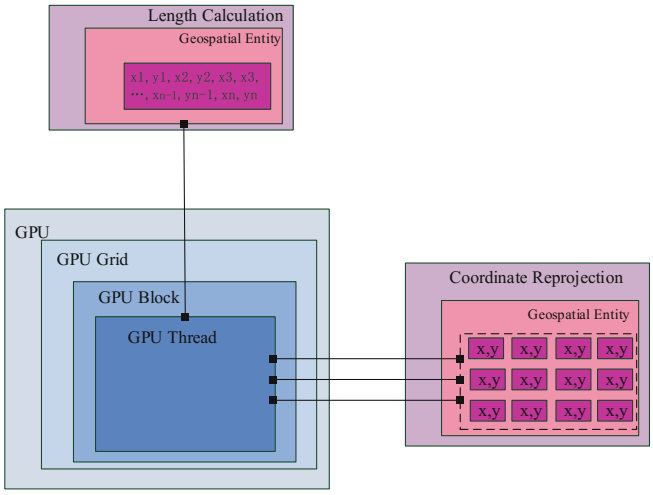


Fig. 2. The mapping between GPU threads and geospatial entities

4 Experiment and Result

Systematic experiments are conducted to demonstrate the good performance of the proposed model and mapping strategy.

4.1 Hardware

The Hadoop cluster contains three personal computers, one as the name node and the other two as data node. The name node has a two-core Intel CPU (i5-2450M) with 2.50 GHz frequency and 8 GB of 1333-MHz RAM. The configuration of one data node is two-core 2.66-GHZ CPUs, 4 GB RAM, NVIDIA GeForce GTX275 GPUs with 1-GB GDDR3 device memory and 96 CUDA Cores. The configuration of the other data node is two-core Intel i3-3240 3.4-GHZ CPUs, 4 GB RAM, NVIDIA GeForce GT620 GPUs with 1-GB DDR3 device memory and 96 CUDA Cores. In our experiment, Hadoop 2.0.5 version is set up for the distributing computing platform, and CUDA 5.5 libraries [8] are utilized as the toolkits for GPU programming. Hadoop and CUDA were deployed on both data nodes. For the name node, only Hadoop is necessary because there is no computation involved in the name node.

4.2 Experiment

According to the principle of "Map Number * Max Map Memory <= the memory of the computer", the maximal number of mappers is designed to be 2, the maximal memory size of the map task is 1 GB, the maximal number of mapper is 4, the maximal memory size of the reduce task is set to be 512 MB. In particular, although each data node has 4G memory, considering some of the additional overhead of system

operations, in our experiments, the maximal heap size of Hadoop is set as 2G. The configuration of Hadoop cluster is shown as follows (Fig. 3).

The purpose of the experiments is to investigate how GPU will improve the processing performance of Hadoop Map Reduce model in geospatial processing, including coordinates reprojection and length calculation.

The experimental data is vector data of coastlines whose feature type is polyline. Six treatments are deployed by doubling data size from 0.5 GB to 16 GB in the experiment. In these six treatments, the number of polylines is 26215, 52430, 104860, 209720, 419440 and 838880, respectively.

```
<configuration>
<property>
<name>mapred.job.tracker</name>
<value>10.1.1.152:9001</value>
</property>
<property>
<name>mapred.map.tasks</name>
<value>7</value>
</property>
<property>
<name>mapred.tasktracker.map.tasks.maximum</name>
<value>2</value>
</property>
<property>
<name>mapred.tasktracker.reduce.tasks.maximum</name>
<value>4</value>
</property>
<property>
<name>mapred.map.java.opts</name>
<value>-Xmx1024m</value>
</property>
<property>
<name>mapred.reduce.java.opts</name>
<value>-Xmx512m</value>
</property>
</configuration>
```

Fig. 3. Hadoop cluster configuration

4.3 Result

In this section, we focus on comparative analysis and evaluation of the computing performance of the native Hadoop Map Reduce Model and GPU-accelerated Hadoop Map Reduce Model for processing big oceanographic ENC data. We make use of four types of metrics, preprocessing time (T_1), processing time (T_2), scheduling time (T_3) and speedup ratio (C_{SR}), to evaluate the performance advantage of the Hadoop Map Reduce model, comparing to two-thread computing model, and the performance advantage of GPU-accelerated Map Reduce model. Preprocessing time, processing time and scheduling time constitute the total computing time. Speedup ratio represented by C_{SR} indicates the reduction of processing time when the computing model is changed from the Hadoop Map Reduce Model to GPU-accelerated Map Reduce Model.

It can be observed that, the total processing time of the GPU-enabled Hadoop processing mode is much lower than the other two modes. And with the increase of the data amount, the time difference becomes larger. Nevertheless, compared to the Hadoop mode, data pretreatment and task scheduling time of GPU-enabled Hadoop mode have no much change at each level of data amount. But the data processing time is significantly reduced. Here, the data processing time includes the data transmission time for copying data from host memory to device memory (Fig. 4).

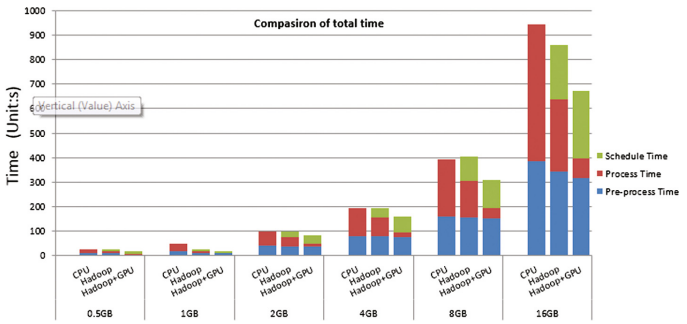


Fig. 4. Comparing total time of data processing in three modes (the single-node two-thread mode, the native Hadoop Map-Reduce mode and the GPU-enabled Hadoop mode)

More specifically, the enhancement of the performance in the GPU Hadoop mode in comparison to the regular Hadoop mode is primarily due to the reduction in processing time. Without GPU acceleration, the proportion of processing data is 35% to 40% (see the left chart in Fig. 5) of the total execution time. With GPU acceleration, the ratio of processing time is reduced to around 12.5% to 14% (see the right chart in Fig. 5). Figure 5 shows the contrast of the computing time ratio with and without GPUs acceleration when the data amount is 4 GB.

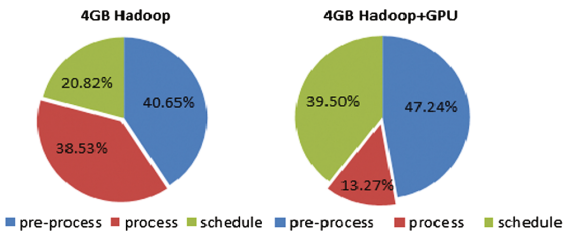


Fig. 5. The ratio of processing time in the total computing time

Figure 6 illustrates the performance optimization results in the computing phrase. It can be observed that the performance and efficiency of Hadoop mode is significantly higher than the dual-core CPU mode, at the same time the efficiency GPU-accelerated Hadoop mode is significantly higher than native Hadoop mode. For example, when

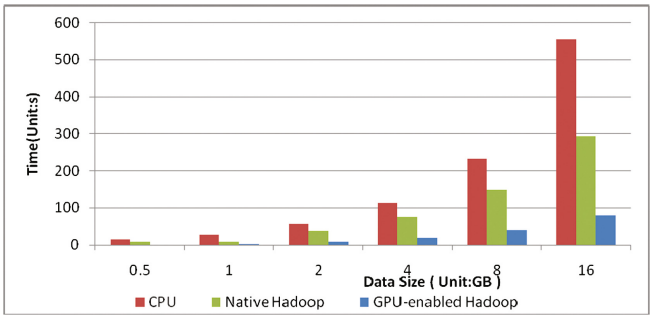


Fig. 6. Comparing data computing time with transferring time under three processing modes.

dealing with 16 GB data (838880 plotlines with 574632800 vertices), the computing time of dual-core CPU is 555.9 s, the computing time of native Hadoop is 295.3 s and the computing time of GPU-accelerated Hadoop is only 81 s. This includes the data transmission time from HDFS to main memory and from main memory to GPU device memory. When the data transmission time is considered, the speedup ratio is shown in Fig. 7, ranging from 3.5 to 4. After removing data transmission time, the speedup ratio in terms of pure computing time can reach up to 600 times (see results in Fig. 8). Therefore, the data transferring time is the primary factor for optimizing performance of GPU and it is also one of main challenges in using GPUs to accelerate the Map Reduce model.

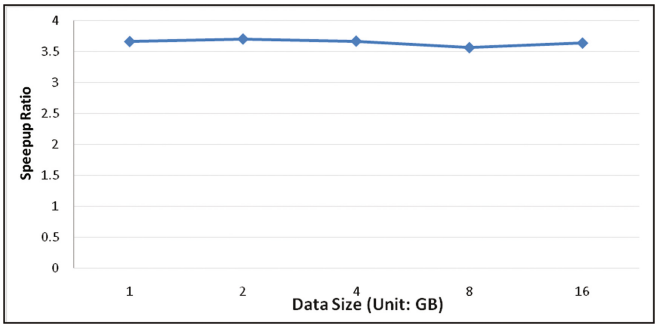


Fig. 7. Speedup ratio of total processing time (data transferring time included)

In summary, implanting high-performance GPU computing module into Hadoop Map Reduce distributed computing framework performs well in computing acceleration. And it presents significant performance improvement on compute-intensive tasks, such as spatial coordinates projection.

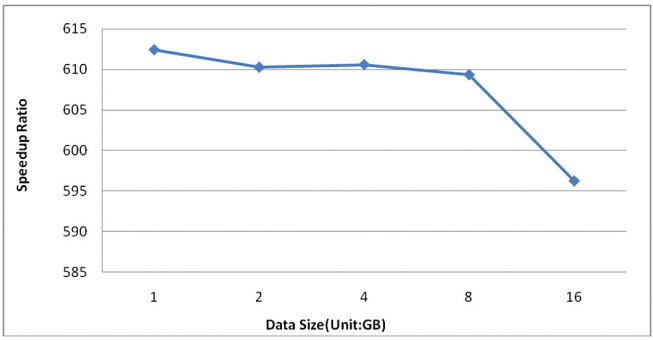


Fig. 8. Speedup ratio of total processing time without considering data transfer time

5 Conclusion

This paper proposes a high performance computing platform for parallel processing of large-scale spatial data. The big oceanographic data processes are first abstracted as Map and Reduce operations to realize cluster-scale parallelism. Then each node in the cluster is accelerated by GPU to achieve fine-grained parallelism. The utilization of GPU and Hadoop Map Reduce model to parallelize geospatial processes ensures that the processing is scalable, while its efficiency is improved and computing time is reduced. In our experiment, the hybrid parallel computing model is adopted to process spatial data of oceanographic ENCs data. This model is also suitable for other types of vector data as well, such as disaster data, urban planning data, land use data, and so on.

Acknowledgment. The research work report in this paper was mainly supported by the Young Scientists Funds (Grant No. 2015QN027) from Shandong Academy of Sciences. It was partially sponsored by the Youth Fund of Natural Science of China (Grant No. 41401435).

References

1. Mitchell, A.E., et al.: NASA's earth observing data and information system-supporting interoperability through a scalable architecture. AGU Fall Meet. Abstr. **1** (2013)
2. Shekhar, S., Gunturi, V., Evans, M.R., Yang, K.: Spatial big-data challenges intersecting mobility and cloud computing. In: Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access. ACM, Scottsdale, pp. 1–6 (2012)
3. Miao, X., Hao, L.: An implementation of GPU accelerated MapReduce: using Hadoop with OpenCL for data- and compute-intensive jobs. In: 2012 International Joint Conference on Service Sciences (IJCSS), pp. 6–11 (2012)
4. Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., Saltz, J.: Hadoop GIS: a high performance spatial data warehousing system over MapReduce. Proc. VLDB Endow. **6**, 1009–1020 (2013)
5. Hecht, H., Berking, B., Buttgenbach, G., et al.: The Electronic Chart: Functions, Potential, and Limitations of a New Marine Navigation System. GITC bv, Lemmer (2006)

6. Shvachko, K., Kuang, H., Radia, S., et al.: The Hadoop distributed file system. In: IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp. 1–10. IEEE (2010)
7. Shao, G., Berman, F., Wolski, R.: Master/slave computing on the grid. In: Proceedings of the 9th Heterogeneous Computing Workshop (HCW 2000), pp. 3–16. IEEE (2000)
8. Bell, N., Hoberock, J.: Thrust: a productivity-oriented library for CUDA. In: GPU Computing Gems Jade Edition, vol. 2, pp. 359–371 (2011)

A Study on Curve Simplification Method Combining Douglas-Peucker with Li-Openshaw

Chengming Li, Pengda Wu, Teng Gu, and Xiaoli Liu(✉)

Chinese Academy of Surveying and Mapping, Beijing, China
{cmli,Wupengda,guteng}@casm.ac.cn, 83391860@qq.com

Abstract. At present, there were two classical approaches commonly used for line simplification. One was Douglas-Peucker (D-P) algorithm and the other was Li-Openshaw (L-O) algorithm. Although the former was able to preferably reserve characteristic bending points of the curve and compress other non-feature points, the simplified result was excessively inflexible and sharp corners were also generated on feature points. As for the latter, not only can the corner of a line be smoothed, instead of becoming over inflexible, but feature point smoothing was also carried out by it for the line. Therefore, based on analyzing characteristics of such two algorithms, an improved algorithm was presented in this first place by means of combining the both together. To be specific, feature points of curves for generalized simplification were figured out by the D-P algorithm, while the L-O algorithm was used to perform curve processing by adjusting radius R of a circle SVO. In the end, real data were applied to carry out experimental verification contrasts for the modified algorithm and the existing two additional algorithms. Experimental results indicated that such a modified algorithm that combined them together exhibited their advantages and had the capability to reserve feature points and simplify other parts simultaneously. Moreover, it could be effectively applied in automated mapping.

Keywords: Douglas-Peucker · Li-Openshaw · Line simplification · Local feature

1 Introduction

Line element that occupies a rather large proportion in map features is mostly used to express extremely important geographic elements such as roads, rivers and contour lines, etc. Hence, simplification related to line element is not only studied in most cases, it has the maximum number of algorithms and the best effects during cartographic generalization. As for the cartographic generalization, the basic thought of simple line element simplification algorithm is to reduce storage of line factors provided that shape features of the line are remained to the greatest extent, such as the common Douglas-Peucker algorithm [1], the Li-Openshaw algorithm, the modified Li-Openshaw algorithm [2] and the arc to chord algorithm [3] etc. Among them, the Douglas-Peucker algorithm is the most well-known line element simplification approach that is able to reserve characteristic bending points of the curve in a better manner and compress its

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 286–294, 2017.
DOI: 10.1007/978-981-10-3969-0_33

other non-feature points despite that the simplification results are over inflexible and sharp corners are caused at feature points. In comparison, the Li-Openshaw algorithm can smooth corners of the line without being excessively inflexible. Besides, feature points of the line are also smoothed by it. Currently, none of them can be completely up to generalization requirements in the process of simplification.

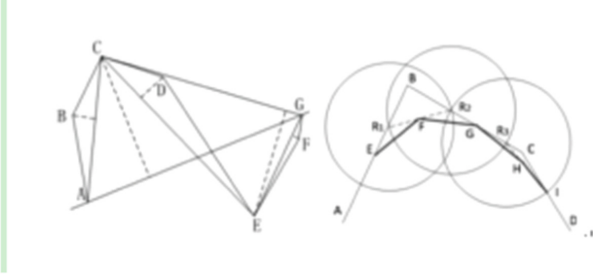
Specific to diversity of the line element and complexity of local shape features related to bending points, a simple curve simplification method was presented in this paper by combining Douglas-Peucker and Li-Openshaw approaches together to guarantee that feature points of the curve can be reserved and other parts of it can also be smoothed and simplified at the time of linear feature generalization.

2 Combination of Douglas-Peucker and Li-Openshaw

2.1 Douglas-Peucker (D-P) Algorithm

As an algorithm emerging at an earlier stage, the D-P algorithm has extensive applications in compression of lines and is also referred to as a very classical algorithm. Compression carried out based on this algorithm is mainly targeted at points on the curve on the premise of reserving main features of the line. In addition to high efficiency, no redundant points can be generated.

Principle of the D-P algorithm is shown in Fig. 1(a). Both ends A and G of a broken line ABCDEFG are connected into a straight line; then, the maximum vertical dimension d from point C on this broken line to segment AG is figured out. If the vertical dimension dc from C to AG is less than the given integrated threshold of distance, the segment AG is seen as an approximate curve of this broken line. However, in the case of dc > threshold, point B is utilized to divide such a broken line into CA and CG. Subsequently, the above procedures are repeated. In the end, all break points are successively connected together into a broken line serving as the approximation of the original broken line. For example, in Fig. 1, ACEF is the broken line after simplification of the original broken line. It can be found that employing the D-P algorithm to simplify segment can remain feature points of the original segment and generalize such an



a) Douglas-Peucker Algorithm b) Li-Openshaw Algorithm

Fig. 1. Douglas-Peucker and Li-Openshaw algorithms

original segment at the same time without smoothing and generalizing those other than feature points. Consequently, it becomes rather difficult for such an algorithm to be widely applied into simplification of geographical line elements.

2.2 Douglas-Peucker (D-P) Algorithm

The Li-Openshaw algorithm is a self-adaptive line synthesis algorithm based on natural laws. Basic thought of this algorithm is that under the circumstance of fixed resolution, a rather large object may turn into a small one with the reduction in measuring scale; and, after it is up to a certain limit, the object can become a point or disappear completely. Algorithm simplification process is as follows:

- The dimension R of the smallest visual object (SVO) of a circle can be estimated in line with the target and the original scales. Where, S_t refers to the simplified target scale required and S_f to the original scale. As for D, it is a parameter of the simplified SVO on the map. In the opinion of Muller [7], the value of D on the map should be taken as 0.4 mm to ensure the minimum visual resolution.

$$R = S_t \times D \times \left(1 - \frac{S_f}{S_t}\right) \tag{1}$$

- Determine the initial position of a circular SVO. Generally, the starting point of a curve for synthesis serves as the first center of the circle, as presented in Fig. 1(b); then, point A and dimension R of the circular SVO are used as the center of a circle and its diameter respectively to draw a circle additionally, and curves intersect at point Q. thus, midpoint of AQ can be selected as the selection point after synthesis.
- Start from point Q to repeat Step 2. Besides, the processing only arrives at a bending point R1 that serves as the center of a circle. Together with a radius denoted by R, another circle is drawn and it intersects with the original broken line at R2. The midpoint F of R1 and R2 is chosen to be the selection point after synthesis.
- Start from point R2 to repeat steps 2 and 3 up until an endpoint D is not terminated within the circle SVO.
- In Fig. 1(b), a broken line ABCD is the original segment without simplification and AEFGHID is a broken line after generalized simplification. It is evident in this figure that L-O algorithm can be adopted to smooth and generalize inflection points of the broken line and make its local features artistic. However, the relevant feature points are also generalized, which cause such an algorithm to be not applicable to line element simplifications during which feature points related should be reserved.

2.3 An Algorithm Combining Douglas-Peucker and Li-Openshaw

Based on the above analysis, it can be found that the individual application of D-P algorithm gives rise to rough and inflexible line element simplification results while the individual employment of the L-O algorithm can lead to circumstances of reserved points in chaos and overall curve shape deformation, etc. Considering that line simplification is aimed at maintaining feature points of line element, guaranteeing shape

features of line element and obtaining smooth and artistic line element simplification results, a simple curve simplification method combining D-P and L-O algorithms together is put forward in this paper. The corresponding modifications to line simplification algorithms are as follows.

- As for a curve for generalized simplification, D-P algorithm is adopted to figure out an inflection point with a vertical dimension larger than the distance threshold on the broken line, as shown in Fig. 3. B, C, D and F are all inflection points meeting the above condition.
- Dependent on target and original scales, Eq. (1) is used to estimate the dimension R of SVO.
- Starting from a circular SCO with a center of the circle of A and a radius of R, L-O algorithm is utilized to simplify the broken line. At a gathering point of inflection points (e.g., the inflection point B given in Fig. 3), it can be processed as follows. Point B is used as the center of a circle to draw a circle O with a radius of R1 (R1 = R). If inflection points (C & D) before/after the point B are inside the circle O, C and D are noted down for processing as the circumstances may require. In the case that both of them should be processed, the radius of SVO should be changed (e.g., 1/2 R) by regarding point B as the center of another circle that should be drawn again, followed by smoothing and generalization of C and B. If not required, smoothing can be ignored. At a point where inflection points are sparsely distributed (e.g., inflection point F in Fig. 3), it should be processed as follows. Point F and radius R1 (R1 = R) are used as the center of a circle and its radius respectively to draw a circle O that should be smoothed subsequently.
- Relevant processing is carried out in order up until I reaching the terminal ends. In Fig. 2, ABCEFI is the original curve while ABCDEGHI is the curve simplified.

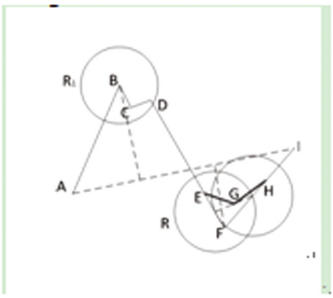


Fig. 2. Modified algorithm combining Douglas-Peucker with Li-Openshaw

3 Analysis and Evaluation of Algorithm

3.1 Simplification Results of the Modified Algorithm

Data generalization simplification of map roads and river networks for a city in South China is adopted as the research case in this paper. Moreover, dependent on WJ-III

map workstation on the NewMap platform developed by the Chinese Academy of Surveying and Mapping (CASM), the modified line simplification method proposed here is inserted to make a comprehensive choice for a map whose scale is changed into 1:100000 from 1:10000 initially. At the same time, comparative analysis is also conducted between this algorithm and Douglas-Peucker and Li-Openshaw algorithms. In Figs. 3 and 4, data related to roads and rivers are intercepted, so are results of their simplifications based on such 3 algorithms on the premise of identical parameters.

a) Douglas-Peucker b) Li-OpenShaw c) The Modified Algorithm

Fig. 3. A comparison of three algorithms for 1:10000 scale road line simplification results (full lines are after the simplification)

In line with the generalized simplification thoughts of this algorithm, its simplification results become more similar to the original curve if compared to the additional two algorithms. Visually, the results conform to those that can be obtained according to the initial thoughts. In Fig. 3(a), original feature points are reserved by the D-P algorithm at the time of road simplification and the relevant line elements are also compressed in a better manner, despite that it fails to be strongly applicable to cartographic generalization as smoothing processing is required by geographical line elements in most cases. As for Fig. 3(b), road lines before and after simplification based on the L-O algorithm are compared. In this figure, it is clear that this algorithm succeeds in smoothing filtering in the process of road line generalization simplification. However, during practical cartographic generalization, not only smoothing is required, it should also be guaranteed that no processing or minor processing is conducted for the original feature bending. In Fig. 3(c), the modified algorithm that combines D-P and L-O algorithms together is used to draw a comparative graph for road lines before and after generalization. According to this figure, their structures after generalized simplification are superior to simplification results of the above two figures. In detail, not only is smoothing performed for road line element simplification, but the original feature bending is also reserved together with local maximum value points precisely. Hence, the applicability of such a modified algorithm in cartographic generation is substantially improved.

With regard to Fig. 4, they are data simplification results for 1:10000 river networks in a city of South China. By comparing such results achieved based on three algorithms, it is verified again that the modified algorithm presented here combines advantages of two classical simplification algorithms and reserves macro-bending features.

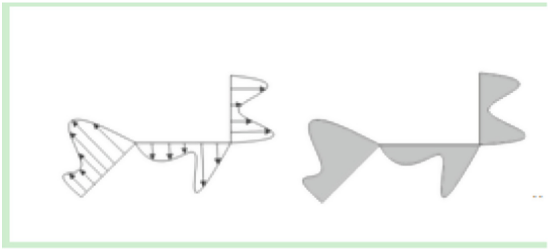
a) Douglas-Peucker b) Li-Openshaw c) The Modified Algorithm

Fig. 4. A comparison of three algorithms for 1:10000 scale river network simplification results (full lines are after the simplification)

3.2 Result Evaluation

In terms of cartographic generalization for line elements, influences of algorithm simplification on line element precision are mainly embodied in overall and local displacements of curves. Therefore, evaluation indexes such as vector displacement, area displacement, displacement standard deviation and position error, etc. are given to compare and assess the modified algorithm and another two algorithms of Douglas-Peucker and Li-Openshaw.

Both vector displacement and area displacement are two evaluation indexes presented by White for geographical line element simplification [4]. The former refers to the position deviation of corresponding points on the simplified and the original curves; while, the latter is the area of a part enclosed by the simplified and the original curves [5, 6]. In this paper, the average offset value and the area deformation value are respectively taken for the vector displacement and the area displacement to carry out the relevant evaluations, as shown in Fig. 5. Moreover, both values are quantitative indexes.



a） Vector Displacement b） Area Displacement

Fig. 5. Evaluation parameters

Three algorithms are adopted to perform generalized simplification for road lines and river networks separately. After simple comparisons based on simplification results (see Table 1), it is found that the modified algorithm is not only better than another two algorithms in terms of visual effects, but far less than and superior to the them as far as the average offset values and area deformation values of their vectors are concerned. Such a result indicates that the modified algorithm is more appropriate to maintain shape features of the line element.

Table 1. Vector and area displacement evaluation results of roads and river network data

	Road simplification		River network simplification	
Algorithm	Average offset value (m)	Area deformation value (m^2)	Average offset value (m)	Area deformation value (m^2)
Douglas-Peucker	4.772	14400.256	16.651	33191.355
Li-Openshaw	1.911	5657.401	4.431	8538.953
The modified algorithm	1.146	3558.345	1.991	4080.557

Muller puts forward a concept of Standardized Measure of Displacement (SMD) to calculate displacement [7]. The corresponding computing formula is the Eq. (2) below.

$$SMD(\%) = 100(1 - (S - D)/S) \tag{2}$$

Where, S refers to the distance from a point with the maximum displacement on the original curve to the connecting line of its start and end points after the algorithm simplification, and D is the displacement value of this point before/after simplification. Evaluations based on SMD are primarily targeted at the local maximum value. Therefore, displacement position error is further employed in this paper to evaluate the global displacement before and after simplification. The displacement position is the ratio between the original curve length and the area enclosed by intersections of curves before and after simplification. It can be calculated according to Eq. (3) below [8–10].

$$\delta = \frac{\Delta s}{L} \tag{3}$$

Where, Δs is the area enclosed by intersections of curves before and after simplification; and L is the length of the original curve.

Table 2 shows displacement standard deviation and position error evaluation results of data related to roads and river networks based on the modified algorithm as well as D-P and L-O algorithms. From Table 2, it can be seen that the modified algorithm is provided with a preferable performance and occupies an optimal position in terms of displacement standard deviation and position error if compared with another two algorithms. Especially for the index of position error, it is particularly excellent considering that its value is the least one. Such phenomena indicate that the modified algorithm is able to validly maintain the overall shape features of road and river network data.

Table 2. Displacement standard deviation and position error evaluation results of roads and river network

	Road simplification		River network simplification	
Algorithm	Displacement standard deviation (%)	Position error (m)	Displacement standard deviation (%)	Position error (m)
Douglas-Peucker	33.536	4.490	11.934	15.796
Li-Openshaw	18.745	1.764	24.743	4.063
The modified algorithm	19.766	1.101	11.841	1.942

4 Conclusions and Expectations

Considering that line simplification is aimed at maintaining feature points of line element, guaranteeing shape features of line element and obtaining smooth and artistic line element simplification results, a simple curve simplification method combining D-P and L-O algorithms together is put forward in this paper. On one hand, such 3 algorithms are used to simplify 2 types of line element objects; on the other hand, 4 quantitative indexes are adopted to evaluate and analyze algorithm simplification effects. Experimental results signify that the modified algorithm presented in this paper is able to maintain local maximum value points of the curve in the process of 1:10000 road and river network data generalized into a 1:100000 scale. In addition, synthetic data are also smoothed so that the overall shape of the curve can be reserved preferably and a synthesis result better than those obtained based on the original algorithm is also achieved.

At present, only generalized simplification of simple linear elements is realized in this paper regardless of adjacent intersections among dense line groups and complex line elements. Moreover, impacts of diverse scales before and after the corresponding generalization on the modified algorithm are also not taken into account. These problems need to be further studied.

Acknowledgment. *Sponsor acknowledgments:* Special Scientific Research Fund of Surveying and Mapping Geographic Information Public Welfare Profession (201512027); National SciTech Support Plan (2015BAJ06B01); Basic Scientific Research Business Expense Project of the Chinese Academy of Surveying & Mapping (7771606).

References

1. Douglas, D.H., Peucker, T.K.: Algorithms for the reduction of the number of points required to represent a digitized line or caricature. Can. Cartogr. **10**(2), 112–122 (1973)
2. Li, Z.L., Openshaw, S.: Automatic synthesis algorithm for line elements based on objective laws of nature. Transl. Wuhan Univ. (1) 49–58 (1994)
3. Nako, B., Mitropoulos, V.: Local length ratio as a measure of critical point detection for line simplification. In: The Symposium of the 5th ICA Workshop on Progress in Automated Map Generalization, pp. 28–30 (2003)

4. White, E.: Assessment of line generalization algorithms using characteristics points. Am. Cartogr. **12**(1), 17–27 (1985)
5. Liu, H., Fan, Z., Xu, Z., Deng, M.: Arc to chord algorithm modification and evaluation for curve simplification. Geogr. Geo-Inf. Sci. **27**(1), 45–48 (2011)
6. Deng, M., Chen, J., Li, Z., et al.: An improved local measure method for the importance of vertices in curve simplification. Geogr. Geo-Inf. Sci. **25**(1), 40–43 (2009)
7. Muller, J.C.: Fractal and automated line generalization. Cartogr. J. **24**(1), 27–34 (1987)
8. Joao, E.M.: Causes and Consequences of Map Generalization. Taylor and Francis, London (1998)
9. Zhu, K., Wu, F., Wang, H., et al.: Li-Openshaw algorithm modification and evaluations. Acta Geodaet. Cartogr. Sin. **36**(4), 450–456 (2007)
10. Wu, F., Zhu, K.: Evaluation on geometric accuracy of line element simplification algorithm. Geomat. Inf. Sci. Wuhan Univ. **33**(6), 600–603 (2008)

Applications of Geo-informatics in Resource Management and Sustainable Ecosystem

A Mobile Services Collaborative Recommendation Algorithm Based on Location-Aware Hidden Markov Model

Mingjun Xin(✉), Shunxiang Li, Liyuan Zhou, and Guobing Zou

School of Computer Engineering and Science,
Shanghai University, Shanghai 200444, China
xinmj@shu.edu.cn

Abstract. Nowadays, location based services (LBS) has become one of the most popular applications with the rapid development of mobile Internet environment. More and more research is focused on discovering the required services among massive information according to the personalized behavior. In this paper, a collaborative filtering (CF) recommendation algorithm is presented based on the Location-aware Hidden Markov Model (LHMM). This approach includes three main stages. First, it clusters users by making a pattern similarity calculation of their historical check-in data. Then, it establishes the location-aware transfer matrix so as to get the next most likely service. Furthermore, it integrates the generated LHMM, user's score and interest migration into the traditional CF algorithm to generate a final recommendation list. The LHMM-based CF algorithm mixes the geographic factors and personalized behavior and experimental results show that it has more accuracy than other state-of-the-arts algorithms.

Keywords: Behavior prediction · LBS · LHMM · Collaborative recommendation

1 Introduction

With the rapid development of mobile Internet and spatial information processing technology, user's behavioral prediction stimulates considerable research interests. Collaborative filtering (CF) is an effective recommendation algorithm. But when it applies to the behavioral prediction, it has limitation [1]. Users' next action is greatly depends on their former choice, which are always not considered. Traditional CF algorithm is not as good as Hidden Markov Model (HMM) under LBS environment, which is frequently used to deal with states transition and predict the probability of services-to-service transfer.

As for HMM strategies, Blasiak and Rangwala [2] applied HMM to the classification, which completed the sequence classification by combining Baum-Welch, Gibbs sampling and change function together. Hamada et al. [3] used a modified BP-AR-HMM algorithm to predict user's driving behavior under multi-time series. Mathew and Raposo [4] completed the prediction of user's next location through putting labeled triangle into HMM learning model.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 297–306, 2017.
DOI: 10.1007/978-981-10-3969-0_34

Through mapping the geographic information and service categories, the location-aware HMM (LHMM) is presented in this paper. This model gives the occurring probability of each service, along with the most likely occurred area. Then, the CF-Behavior prediction algorithm combining LHMM and CF is proposed. It both considers the location and personalization factor. What's more, it can reduce the dimensions of similarity calculation.

2 Behavioral Sequence Prediction

2.1 Behavioral Prediction and Recommendation Framework

As shown in Fig. 1, the whole recommendation framework is divided into three key modules: similar behavior cluster, series forecast and CF-behavior prediction.

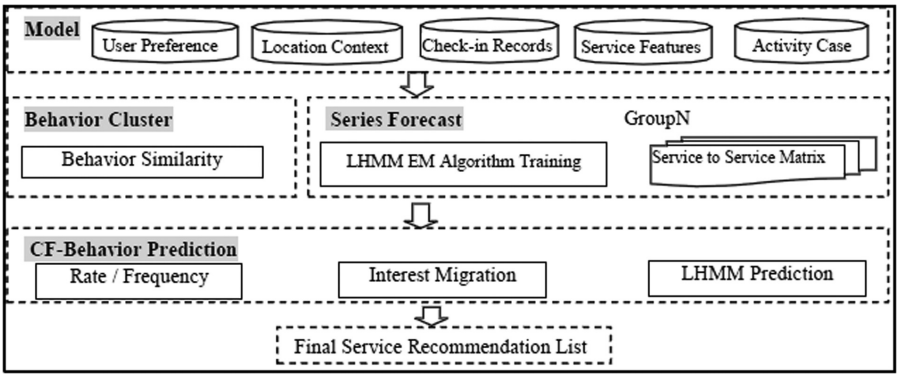


Fig. 1. The BP model to predict behavior and recommend services

(1) *Behavior cluster*. This module is design to group users who enjoy similar life rhythm together. It will generate the top-k users who have similar life pattern.
(2) *Series forecast*. This module trains similar user's check-ins to obtain initial probability matrix, transition probability matrix and emission probability matrix for LHMM model.
(3) *CF-behavior prediction*. This module provides more detailed recommendations, which combines the users' rate, visit frequency and interest migration.

2.2 Behavioral Sequence Model and Similarity Calculation

In order to reduce the sparseness calculate the transfer probability between two time states for different user groups, the check-in data is then to be divided into six different stages, namely {1–5, 6–10, 11–13, 14–16, 17–19, 20–24}.

Definition 1 (Check-in): Check-in (*CK*) indicates that a user *U* check in at shop *S* at a certain time *T*. User rates shop *S* with *R*, R ∈ {0, 10, 20, 30, 40, 50}.

CK = (userId, userName, user City, time, shopId, star, comment)

Definition 2 (Shop): Shop *S* represents the place where user participates in an activity with check-in.

Shop = (shopId, shopName, address, city, district, area, category, subcat, lat, lon)

Definition 3 (Score Function): *x* and *y* are two check-in record. σ(x, y) represents the output through function σ. The bonus points are designed as follow:

$$\sigma = \begin{cases} \sigma + 4x.district = y.district \\ \sigma + 2x.service = y.service \\ \sigma + 1x.area = y.area \\ \sigma - 3x \neq y \end{cases} \quad (1)$$

Definition 4 (Sequence Similarity): Given two sequences $S = S_1 \text{ -> } \ldots \text{ -> } S\text{-}_n$, $T = T_1 \text{ -> } \ldots \text{ -> } T\text{-}_n$. $|S|$ denotes the length of the sequence *S*. The similarity between two behavioral sequences is calculated by Eq. 2 as below:

$$\text{Score} = \sum_{i=1}^{m} \sigma(S_i, T_i) \text{ where } m = |S| = |T| \quad (2)$$

Definition 5 (Behavior Similarity): $A(U_1, U2)$ shows the highest similarity after one-to-one sequences comparison between two users. Based on ClustalW [5] sequences match theory, the sequences similarity one-on-one between two different groups can be calculated.

3 Behavioral Prediction Based Collaborative Recommendation Algorithm

3.1 LHMM (Location-Aware Hidden Markov Model) Generation

A set of hidden states $S = \{S_1, S_2 \ldots S_M\}$ represents user's current location, along with a set of observations $O = \{O_1, O_2 \ldots O_N\}$ which represents the activities participated by the users. Figure 2 shows an illustration of relationship between hidden states and observations.

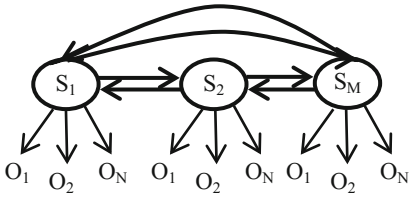


Fig. 2. The relationships between hidden states and observations of LHMM

There are three important parameters defined as follows: The initial probability matrix indicates the probability of each hidden state $S_i \in S$. Transition probability matrix indicates the probability from hidden state S_i to hidden state S_j. Emission probability matrix indicates the probability of observed $O_i \in O$ under a given state S_i.

Expectation-maximization algorithm (EM) can be used to search a set of LHMM parameters. Baum-Welch [6], also known as forward-backward algorithm, is the most widely used method for solving HMM learning. The basic idea is using a random initialization λ, assuming that this λ is the optimal solution.

What's more, in view of occurrence frequency of the sequence, we improve the Eqs. 8–10 listed in paper [7] with weight factor, so that periodic sequences can be better handled in new model. The improved formula is shown below.

$$\mathrm{L(i)} = \sum_{\mathrm{O_k=O_1}}^{\mathrm{O_n}} \gamma_{\mathrm{i,O_k}}(1) * \mathrm{C(O_k)}\overline{\pi_{\mathrm{i}}} = \frac{\mathrm{L(i)}}{\sum_{\mathrm{i=1}}^{|\mathrm{S}|} \mathrm{L(i)}} \tag{3}$$

$$\mathrm{M(i,j)} = \sum_{\mathrm{O=O_1}}^{\mathrm{O_n}} \sum_{\mathrm{t=1}}^{\mathrm{L-1}} \varepsilon_{\mathrm{i,j,O_k}}(\mathrm{t}) * \mathrm{C(O_k)}\overline{\mathrm{A_{i,j}}} = \frac{\mathrm{M(i,j)}}{\sum_{\mathrm{j=1}}^{|\mathrm{S}|} \mathrm{M(i,j)}} \tag{4}$$

$$\mathrm{N(i,E_p)} = \sum_{\mathrm{O=O_1}}^{\mathrm{O_n}} \sum_{\mathrm{t=1}}^{\mathrm{L}} \delta_{\mathrm{o_t,o_k}} \gamma_{\mathrm{i}}(\mathrm{t}) * \mathrm{C(O_k)}\overline{\mathrm{B_i(E_p)}} = \frac{\mathrm{N(i,E_p)}}{\sum_{\mathrm{E_p \in O}} \mathrm{N(i,E_p)}} \tag{5}$$

Where $\mathrm{C}(O_k)$ represents the weight of observed sequence O_k. γ_i represents the probability of generating sequence Z under the state i. $\varepsilon_{i,j}$ shows the probability of generating sequence Z during the transition from state i to state j. The original equation multiplied by the weight $\mathrm{C}(O_k)$, *L, M, N* formula is obtained which help specify the certain occurred frequency of sequences.

3.2 Next Service and Location Prediction

After the three important parameters of LHMM are generated, it is time to predict user's future possible behavior. Assuming that $\mathrm{CK} = \{\mathrm{O_1,O_2 \ldots O_t}\}$ is a check-in activity sequence, where O_i represents the check-in activity at time i, corresponding to the check-in place sequence $\mathrm{S} = \{\mathrm{S_1,S_2 \ldots S_t}\}$. Now, in order to derive the activity O_{t+1} at time $t + 1$, it can analyze the most probable hidden state S_{t+1} at time $t + 1$ by applying Eqs. 6–7 in this paper.

$$\mathrm{p(O_{t+1}|O_{1..t})} = \sum_{\mathrm{S_{t+1}}} \mathrm{p(O_{t+1}|S_{t+1})} \sum_{\mathrm{S_t}} \mathrm{p(S_{t+1}, S_t|O_{1..t})} \tag{6}$$

$$\sum_{\mathrm{S_t}} \mathrm{p(S_{t+1}, S_t|O_{1..t})} = \frac{1}{\sum_{\mathrm{S_t}} \alpha(\mathrm{S_t})} \sum_{\mathrm{S_t}} \mathrm{p(S_{t+1}|S_t)} \alpha(\mathrm{S_t}) \tag{7}$$

The $p(O_{t+1}|S_{t+1})$ represents the occurrence probability of observation O at time $t + 1$ under given hidden state S_{t+1}. Moreover, $p(S_{t+1}|S_t)$ shows the state transition

probability from time t to time $t + 1$. $\alpha(S_t)$ represents the probability of observing $O_{1\ldots t}$ at time t under hidden state S.

Each calculated pair $<S_i, O_j>$ forms an array of service-location probability. By sorting this array, the most probable activity and its corresponding location will be easily worked out with the LHMM which user may take in the next period.

3.3 CF-Behavior Prediction Algorithm (LHCF)

If it is just stopped at previous module, the recommendations within a group are the same result. In fact, although users share similar routine of the day, it doesn't imply that they enjoy similar interests. For instance, users A and B usually have lunch during 11:00–13:00 at District X. User A prefers restaurant a while user B prefers b. To solve this problem, the CF-Behavior prediction based on LHMM model (LHCF) is proposed.

Combining with the output from the previous sections, the large matrix can be easily divided into several sub-matrixes at space (S_i) and category (C_i) level, shown as Fig. 3.

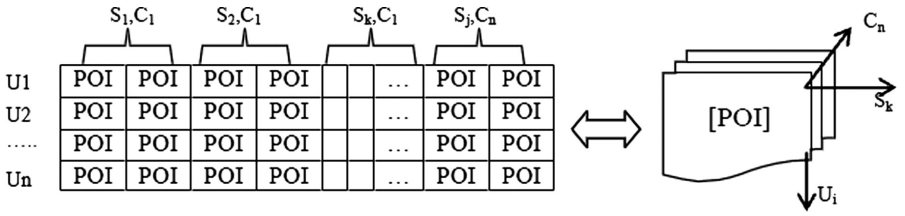


Fig. 3. Divide dataset into subset by space and category information

According to the above dataset, user's preferences can be affected by three important factors: score, visiting frequency and the time user visited. Therefore, a time transfer function is added into final formula, so that users can be clustered in a more proper way. Equations 8 and 9 are the definition of user's point of interest (POI):

$$\mathrm{POI}(U_i, S_j) = \left(a^* \frac{\mathrm{avg(score)}}{\mathrm{maxscore}} + b^* \frac{\mathrm{count}(s_j)}{\sum_{s_k \in s} \mathrm{count}(s_k)} \right) * t(U_i, S_j) \tag{8}$$

$$t(U_i, S_j) = 1/\left(\frac{\mathrm{currenDate} - \mathrm{maxDate}}{7}\right) \tag{9}$$

Where t function is an interest migration function, the more frequently a user visit a shop, the higher score it will be. Attributes a and b are two fit parameters in order to calculate the POI in a more flexible way. For check in records without user's rate, an average score will be assigned according to user's historical records.

The cosine similarity calculation formula is applied to calculate the similarity between the different mobile users.

$$\mathrm{Sim}(U_i, U_j) = \frac{\sum_{s_k \in s}(\mathrm{POI}(U_i, S_k) - \overline{\mathrm{POI}(U_i)})(\mathrm{POI}(U_j, S_k) - \overline{\mathrm{POI}(U_j)})}{\sqrt{(\mathrm{POI}(U_i, S_k) - \overline{\mathrm{POI}(U_i)})^2}\sqrt{(\mathrm{POI}(U_j, S_k) - \overline{\mathrm{POI}(U_j)})^2}} \quad (10)$$

$\mathrm{POI}(U_i, S_k)$ means the points of interest of user *i* for service *k*. $\overline{\mathrm{POI}(U_i)}$ presents the average points of interest the user *i* for all service categories.

Algorithm *CF-Behavior prediction*

Input: User: U, prevObservList<location,activity>, HMM parameters
Output: ProbList<location,activity>, RecommendList
Algorithm:
pastObserv = buildPastObserv(prevObservList,time) //predict t+1 service through <$O_1 \ldots O_t$>
stateList = getSortedProbHiddenState(pastObserv); // probability calculation of next state
for stateI in stateList do // For each predicted state, calculating the probability of next observation sequence
observNext<<state,observe>,prob> = insertAndCalculateProb(pastObserv,stateI);
getTop5Prob(observNext);// Probability values are sorted and selected the top five combinations for <region,activity> in <state,observe> do// Iteratively predicted next combination
for userU in users do
for CkR in userU.checkinRecords do
if userU.checkinRecord.region = region and checkinRecord.categories = activity
POI[userU][userU.CkR.shop] = calculatePOI();
UserSim = calculateCosSim(POI) // Calculate the cosine similarity
topSimUserList[] = topUserSim(userI,5) // Get nearest 5 users
RecommendList<region,activity> = findHigherPOIShop(topSimUserList[])
// From the similar users group, find out recommendations users might like shops

Based on the service sequences which users have participated in, a numeric probability list is calculated, indicating the likelihood of each possible service and its corresponding occurred places. After the most possible service is determined, the cosine similarity between two users is calculated based on user's historical rating behavior, visit frequency and interest shift. Finally the recommendation list is gotten according to the improved CF-behavior prediction algorithm.

4 Experimental Evaluation

4.1 Data Analysis

Dianping website (http://www.dianping.com) is a famous leading third-party website that provides detail business information, consumer reviews and other O2O trading services. Through its open API, our dataset has nearly 6000 shops in Shanghai, along with 60,000 check-ins records from 3000 distinct users from 2010 to 2014, shown as Fig. 4.

The 5-fold cross-validation method is used to in this paper. The check-ins is divided into 5 subsets. Every time, one of the 5 subsets is used as the test set S_{test} and the other 4subsets are put together to form training set $S_{training}$. The detail is as follow:

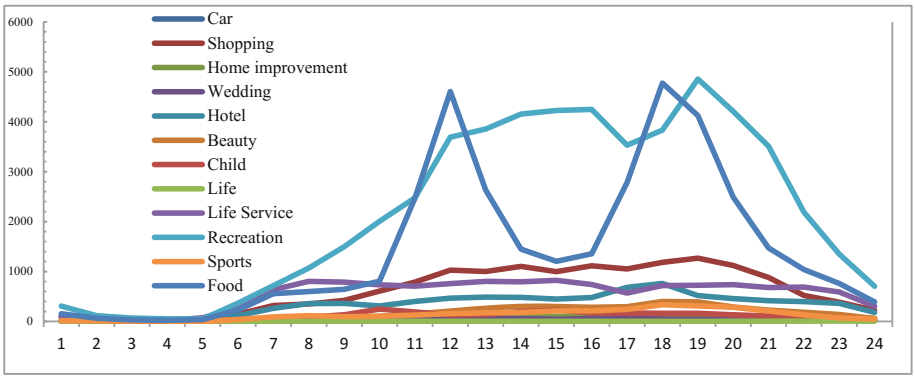


Fig. 4. Different services check-in frequency chart

4.2 Behavioral Prediction Evaluation

The purpose of this experiment is to calculate the prediction accuracy on TopK-LHMM algorithm compared with other traditional methods.

To demonstrate the prediction accuracy of the algorithm, the concept of N-hit is introduced. If the biggest probability value is the one user participates in reality, so this scenario is defined as 1-Hit. All in all, N-hit means that the *n*th value in the prediction array is match with the next step that happens in reality. Obviously, the smaller *n* is, the higher accuracy the algorithm indicates (Table 1).

Table 1. Behavioral sequence similarity calculation

User Check-In	A:-,-,100102,100102,-,-	A:-,-,100105,-,100105,090103
B:-,-,020104,-,020204,-	2(phase3) = 2	2(phase3)−3(phase5) = -1
B:-,-,-,020203,-,-	−3(phase4) = −3	0
C:-,-,-,100110,130110,-	6(phase4) = 6	2(phase5) = 2
C:-,-,-,100203,100105,09020	4(phase4) = 6	7(phase5) + 5(phase6) = 12

Figure 5 shows the experiment result of different nearest *k* value. The performance is very close when *k* is 5 or 10, but when *k* is bigger than 10, the prediction of users' next action will drop. The user behavior is similar, not exactly the same. Expanding the number of similar users will reduce the accuracy of the model prediction. Thus, we assign *k* as 10 in follow-up experiments.

As the Fig. 6 shows, the accuracy of action prediction for the first two hit is over 50%, which has better impact than traditional HMM. If Top-K feature adds into LHMM, the new model helps improve additional 5% better accuracy on the 1-hit to 3-hit in average.

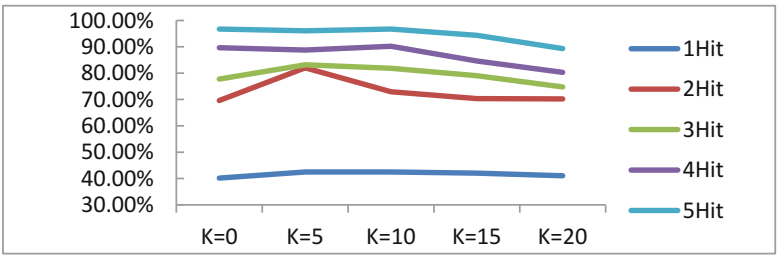


Fig. 5. Experiment on nearest K value

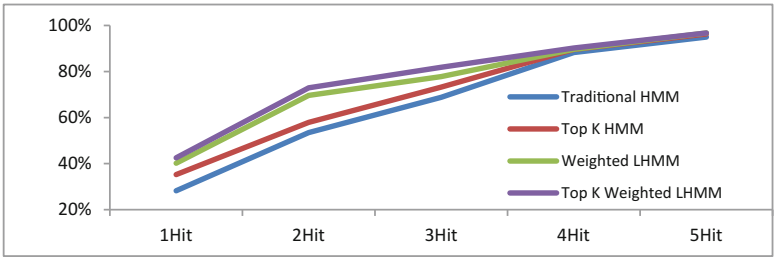


Fig. 6. Behavioral prediction comparison

4.3 Interests Recommendation Evaluation

Different from traditional collaborative filtering, the LHCF use the output <service, area, probability> from LHMM. For evaluating the recommendation algorithms, we compare ours with the following five state-of-the-arts recommendation methods. Location-based Collaborative Filtering (LCF), User interest and Proximity (UP) [8], User interest and geographical influences (UG) [9], Spatio-Temporal Collaborative Filtering (STCF) [10], Location-aware Hidden Markov Model (LHMM):

It is the evaluation indicators *hit@N* and strategies in papers [11, 12] that we use to evaluate the effectiveness of different recommendation algorithms.

$$hit@N = \frac{|S_{success}|}{|S_{test}|} \tag{11}$$

The N represents the number of recommended services. The $S_{success}$ represents the number of success in S_{test}. The S_{test} is a test case set. As for an individual test case $(u,\ s,\ l) \in S_{test}$, the u represents user, s represents service, l represents location.

Firstly, we simulate user's current temporal and spatial properties, which are close to the test check-in. Secondly, different algorithm works out its Top-N recommendation list. Finally, if a Top-N recommendation list includes the testing service, it is successful and the number of $S_{success}$ increases one. The *hit@N* can reflect the quality of recommendation algorithm.

Figure 7 shows the performance of each algorithm. The experiment shows that LHCF algorithm is greatly superior to the others, where it takes time, space and user interest into consideration.

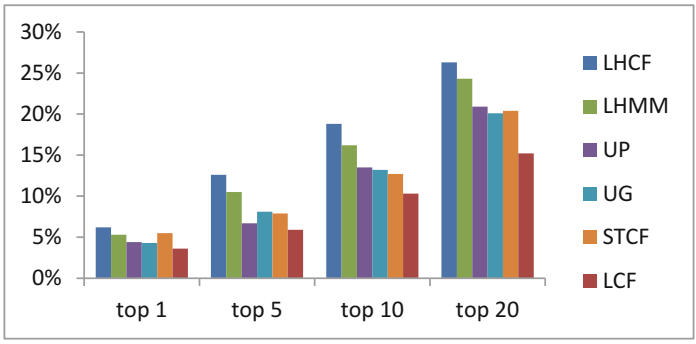


Fig. 7. Comparison experiments

5 Conclusions

In this paper, a user behavior prediction and recommendation framework for location based services is proposed under mobile Internet environment. Based on the users' activity behavior sequences clustering module for location aware mobile services, a Top K-LHMM algorithm is proposed and implemented to do a better prediction for different kind of services and regions under a certain user's status. Under the extensive experiments designed in this paper, the improved system gives us more accurate results. Especially it overcomes the weakness of perception of location and time context in traditional collaborative filtering algorithm and obtains a high efficiency on mobile service prediction. The improved system has strong scalability to adapt to different services recommendation environment. In the future, our research work is mainly focused on how to extend the framework in several directions.

Acknowledgments. This work is partially supported by National Natural Science Foundation of China (61074135, 61303096, 71101086) and Shanghai Leading Academic Discipline Project (J50103). We also would like to show our great appreciations to all of our hard working fellows in the projects above.

References

1. Natarajan, N., Shin, D.: Which app will you use next: collaborative filtering with interactional context. In: Proceedings of the 7th ACM Conference on Recommender Systems, pp. 201–208 (2013)
2. Blasiak, S., Rangwala, H.: A hidden Markov model variant for sequence classification. In: IJCAI 2011, Proceedings of the International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, pp. 1192–1197. DBLP, July 2011

3. Hamada, R., Kubo, T., Ikeda, K.: Towards prediction of driving behavior via basic pattern discovery with BP-AR-HMM. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2805–2809 (2013)
4. Mathew, W., Raposo, R.: Predicting future locations with hidden Markov models. In: ACM Conference on Ubiquitous Computing, pp. 911–918. ACM (2012)
5. Thompson, J.D., Gibson, T.J.: Multiple sequence alignment using ClustalW and ClustalX. Curr. Protoc. Bioinf (2002). UNIT 2.3
6. Welch, L.R.: Hidden Markov models and the Baum-Welch algorithm. IEEE Inf. Theory Soc. Newslett. **53**(4), 10–13 (2003)
7. Zukerman, I., Albrecht, D.,W., Nicholson, A.,E.: Predicting users' requests on the WWW. In: Kay, J. (ed.) UM99 User Modeling. CICMS, vol. 407, pp. 275–284. Springer, Heidelberg (1999). doi:10.1007/978-3-7091-2490-1_27
8. Ference, G., Ye, M., Lee, W.C.: Location recommendation for out-of-town users in location-based social networks. In: ACM International Conference on Information and Knowledge Management, pp. 721–726. ACM (2013)
9. Ye, M., Yin, P., Lee, W.C., et al.: Exploiting geographical influence for collaborative point-of-interest recommendation. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 325–334. ACM (2011)
10. Yuan, Q., Cong, G., Ma, Z., et al.: Time-aware point-of-interest recommendation. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 363–372. ACM (2013)
11. Koren, Y.: Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, pp. 426–434, August 2008
12. Yin, H., Cui, B., Li, J., et al.: Challenging the Long Tail Recommendation. Proc. VLDB Endow. **5**(9), 896–907 (2012)

3D Visualization Analysis of Longtan Reservoir-Induced Earthquakes and Active Faults

Zhengqiang Long[1(✉)], Hong Yao[1], Shuangqing Liu[2], and Xuejun Sun[1]

[1] Earthquake Bureau of the Guangxi Zhuang Autonomous Region, Nanning 5300252, China
longzhengqiang@l26.com
[2] Earthquake Administration of Tianjin Municipality, Tianjin, China

Abstract. This article collected geological formation, fault occurrence, precise positioning of medium and small earthquakes, focal mechanism solution, and other data of Longtan Reservoir area, built a 3D geologic model on ARCGIS platform, conducted a 3D visualization analysis on the relationships between reservoir earthquakes and active faults and between focal mechanism of medium and small earthquakes and fault occurrence and discussed 3D spatial relationship between focal depth of reservoir-induced earthquakes and geological formation. Results showed that (1) Using focal longitude and latitude, magnitude, focal depth and other parameters from precise positioning results of small earthquakes of 5 earthquake clusters, a 3D spatial model was built between reservoir earthquakes and faults and 3D buffer zones were set up within 6 km radius and 20 km depth of each fault, to analyze the relationship between faults and earthquake distribution, indicating that Gaoxu-Bamao Fault (F2), Fengting-Xialao Fault (F3), Maer-Lalang Fault (F4), Changli-Banan Fault (F8) and Longfeng-Bala Fault (F9) were controlled seismic faults of various earthquake clusters. (2) Using focal longitude and latitude, magnitude, focal depth, strike, dip, slip, T-axis dip, P-axis dip and other parameters of 120 focal mechanism solutions derived from waveform data of earthquakes with magnitudes greater than M_L 2.0 in the reservoir area, a 3D focal mechanism and fault model was built, to calculate focal mechanism type and parameters of 5 earthquakes clusters, suggesting that focal mechanisms in the reservoir area were mainly thrust. Strike and dips of 5 earthquake clusters were basically identical with fault occurrence of main faults through this cluster; (3) Using geological formation, fault occurrence and precise positioning results of medium and small earthquakes, a 3D strata and earthquake model was set up, the relationship between focal depth and geological formation was analyzed visually, suggesting that brittle strata within 5–13 km under the reservoir area were prone to brittle rupture and thereby induce seismicity.

Keywords: Reservoir-induced earthquake · Fault · 3D geologic model · Focal mechanism solution · 3D visualization

Fund Project: This work was supported by the science-technology plan of Guangxi (Project number: 12426002-1, 12426001-2, GXJ2011002, XH12035).

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 307–320, 2017.
DOI: 10.1007/978-981-10-3969-0_35

1 Introduction

The relationship between earthquakes and active faults is one of the most important research content in the field of seismogeology. According to statistics, about 85%–95% of earthquakes with magnitudes greater than 6 are caused by faulting. Not only the relationships between active tectonics and earthquakes were consistent in spatial distribution, but also the movement features of active faults and co-seismic ruptures produced along faults in great earthquakes were uniform [1]. In recent years, many scholars at home and abroad [2–10] have studied the relationship between earthquakes and active faults in different areas and tried to find out seismic features of all fault zones, as well as the deep tectonic changes of main faults. Most of the predecessors' work adopted 2-Dimension Geographic Information System (2D-GIS), such as fault buffer zone analysis, seismicity overlay analysis and fault section plotting. Earthquakes and active faults lie in 3D geological space and earthquakes are distributed unevenly in 2D space. If this kind of uneven distribution is mapped to 2D space, earthquake mapping produced in the same active fault may be mapped to different faults and mislead our understanding of the whole data field. Moreover, it is susceptible to distortion of geologic information, complex operation, difficulty in revision and update, etc. when using 2D plane method.

With the rapid development of computer and information technology, in order to solve 3D problems in the field of geoscience, the Canadian Houlding [11] first put forward 3-Dimension Geoscience Modeling System (3DGMS), whose meaning was to apply modern spatial information theory to study the information processing, data organization, spatial modeling and numeric expression of geoscience and its environment and to apply 3D visualization to represent geological formation and its environment. In the field of seismogeology, 3DGMS not only describes the distribution of physical parameters of complex geologic body underground and spatial relationship between tectonic elements and earthquakes, but also carries out a spatial query and visualization analysis based on this model and further finds out hidden phenomena and laws from the seemingly chaotic massive data of earthquakes and faults, to provide a basis for scientific decision-making [12]. Therefore, scholars began to seek correlations between earthquakes and faults in 3D space [13–16] and raised some problems worthy to be discussed: first of all, geological formation, fault occurrence, precise positioning of earthquakes, focal mechanism and other data were seldom applied to the study of 3D earthquakes and faults. Secondly, a more intuitive and stereo 3D geological model hadn't been set up to analyze the relationship between earthquakes and faults. In this paper, on the basis of predecessors' work, with Longtan Reservoir as a target area, using geological formation, fault occurrence, precise positioning of medium and small earthquakes, focal mechanism and other data, a 3D geological model was built on ARCGIS platform, a 3D visualization analysis was conducted on the relationships between the focal mechanism of medium and small earthquakes and fault occurrence and between earthquakes and active faults and the spatial relationship between focal depth and geological formation was discussed.

2 Geological Structure of Longtan Reservoir Area and Establishment of a 3D Geological Model

Longtan Reservoir area stretches across Tian'e County in Guangxi and Luodian County in Guizhou and is situated in a slope from the southern margin of Yunnan-Guizhou Plateau to mountains and hills in Northwest Guangxi. According to tectono-sedimentary evolution history, stratal sedimentary environment and development features in this area, strata in the study area (24.8°–25.5° N; 106.5°–107.3° E) were divided into two parts: basement and caprock. Basement strata were formed in Sibao Orogeny and Xuefeng Orogeny during Meso-Neoproterozoic period. They are bathyal-abyssal facies clastic rocks, mingled with multilayer pillow-shaped spilite keratophyre and pyroclastic rocks, generally subject to regional metamorphism, and belong to epimetamorphic greenschist facies. Sedimentary caprocks are divided into three major stratigraphic sequences, namely: (1) Sinian- Early Paleozoic geosynclinal sedimentary sequence, mainly Sinian glacial-marine facies pebbly sand mudstone and Cambrian–Silurian carbonatites. (2) Late Paleozoic paraplatform sedimentary sequence. Among them, Lower-Middle Devonian Series are composed of sandstone, mudstone, argillaceous limestone and other clastic rocks and carbonatites. Upper Devonian-Lower Permian Series are continuous carbonatites sediments. Upper Permian Series are siliceous limestone, siliceous shale mingled with sandstone, shale and other clastic rocks and carbonatites. (3) Meso-Cenozoic regenerated geosynclinal- epicontinental active belt basin-type stratigraphic sequence. Among them, Lower-Middle Triassic Series are sandstone, mudstone, shale, argillaceous limestone and other quasi-flysch sediments. Upper Triassic-Palaeogene Systems are thick bedded conglomerate, pebbly sandstone, shale and other continental facies sediments. Quaternary System is alluvium, diluvium and eluvium composed of loose substances [17, 18]. The study area mainly presents Devonian, Carboniferous, Permian, Triassic, Palaeogene and Quaternary sedimentary strata (Fig. 1). Based on the above analysis, a 3D strata model of reservoir area was set up. The ground surface of model adopted SRTM 90 m elevation data (http://srtm.csi.cgiar.org/SELECTION/inputCoord.asp). The surface

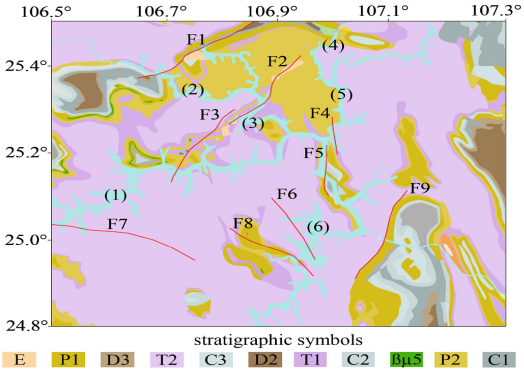


Fig. 1. Tectonic map of Longtan reservoir area

stretches down to −20 km. The deep strata are divided into four horizontal strata, corresponding to 1 basement and 3 sets of sedimentary caprocks in the study area respectively. Among them, the epimetamorphic rock strata of basement are 8 km thick. Sinian- Lower Palaeozoic Erathem strata, which are mainly compose of carbonatites, are 4 km thick. Upper Paleozoic Erathem strata, which are mainly compose of carbonatites, are 4 km thick. Triassic-Quaternary quasi-flysch and clastic rock sedimentary strata extend from ground surface to an altitude of −4 km.

In geotectonics, the reservoir and adjacent areas are located within Youjiang fold belt, a second-order tectonic unit of South China fold system. Indosinian-Yanshan Orogeny lays the basic tectonic framework in this area. Fault structures in different directions during the Neotectonic Period also presented different degrees of differential activities [19]. According to field geological landform and seismic geological surveies, within the scope of study area, four groups of faults, i.e., NW, NNW, NE and quasi-SN are developed (Fig. 1). Most of them belong to fault structure formed with Indosinian folds and are limited to two wings or core of the fold. The dips are abrupt, the stretches are not far and the scales are small. The main structure near the reservoir area is Longfeng-Bala Fault (F9), with a Tian'e box-like anticline, and its west wing is NNE-striking. This fault was developed between Triassic and Permian Systems. The length inside is about 25 km. It is a normal fault tilting to the west at a high angle. In the southwest of reservoir area, there is NNW-striking Wangmo-Luoxi Fault (F7) and NW-striking Changli-Banan Fault (F8) and Dangming-Guahua Fault (F6). Among them, Wangmo-Luoxi Fault shows a left-stepping, oblique-slip distribution and is composed of multiple secondary small faults and basement fractured zones, about 28 km long inside. Changli-Banan Fault is a dense cleavage belt and gentle fold deformation belt developed in Permian limestones and tends to be SW-striking, with reversed strike-slip movement properties. Dangming-Guahua Fault is developed in Middle Triassic sand mudstone strata and also has thrust strike-slip movement properties. In the northwest of reservoir area, there is NE-striking Luodian-Wangmo Fault (F1), Gaoxu- Bamao Fault (F2) and Fengting-Xialao Fault (F3). Among them, the northeast section of Luodian-Wangmo Fault is a line of demarcation between Paleozoic and Mesozoic Erathems. The southwest section is inserted into Mesozoic strata. The length inside is about 26 km. Gaoxu-Bamao Fault and Fengting-Xialao Fault divides Permian and Triassic Systems. Fractured zones develop, up to 30–100 m wide. The section has a relaxed wave-like shape, predominantly dextral strike-slip movement. In the due north of the dam, main structures include Ma'er-Lalang Fault (F4) and Daheng-Daliang Fault (F5), which have a Daliang anticline and the east and west wings are quasi-SN-striking. 2 faults have opposite fracture tendency, but both of them have normal strike-slip movement properties. According to the fault activities revealed in tectonic landform and geologic section and the obtained chronological evidence, expect that F1, F5 and F7 are pre-Quaternary active faults, other faults had different degrees of activities during Early-Middle Pleistocene. Fractured zones develop. The fissure and karst fissure springs are in linear distribution along fractured zones [20, 21]. Combined with fault occurrence parameters (Table 1), a 3D fracture model of reservoir area was set up (Fig. 2b). Each fault in the model extended from ground surface to an altitude of −20 km. The main basis of fault occurrence parameters included existing geological

Table 1. 3D modeling parameters of reservoir faults

Fault name	Strike φ /°	Dip δ /°	Slip γ /°
Luodian-Wangmo Fault (F1)	55	90	180
Gaoxu -Bamao Fault (F2)	30	65	−150
Fengting-Xialao Fault (F3)	210	73	−150
Maer-Lalang Fault (F4)	353	70	−35
Daheng-Daliang Fault (F5)	178	75	−35
Dangming-Guihua Fault (F6)	155	75	40
Wangmo-Luoxi Fault (F7)	110	90	0
Changli-Banan Fault (F8)	125	70	35
Longfeng-Bala Fault (F9)	203	65	-55

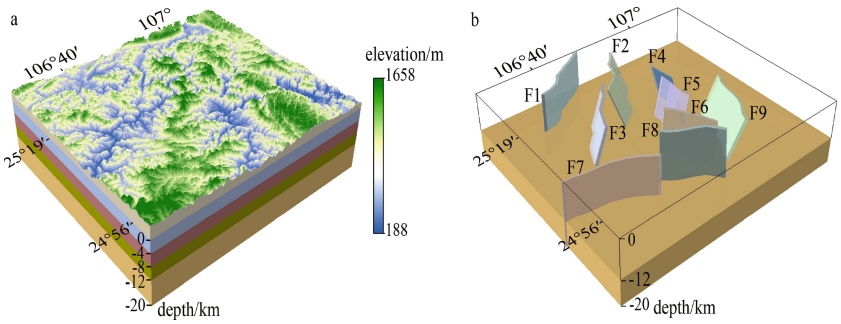


Fig. 2. 3D geologic model of Longtan reservoir area (a) 3D stratigraphic model of Longtan reservoir area; (b) 3D Fracture model of Longtan reservoir area

data [17, 18] and field seismogeologcial survey data. The author also referred to Chen et al.'s [22] focal mechanism findings, to determine comprehensively.

(1) Hongshui River; (2) Meng River; (3) Youla River; (4) Caodu River; (5) Niu River; (6) Buliu River; F1: Luodian-Wangmo Fault; F2: Gaoxu-Bamao Fault; F3: Fengting-Xialao Fault; F4: Maer-Lalang Fault; F5: Daheng-Daliang Fault; F6: Dangming-Guihua Fault; F7: Wangmo-Luoxi Fault; F8: Changli-Banan Fault; F9: Longfeng-Bala Fault.

3 3D Visualization Analysis of Longtan Reservoir-Induced Earthquakes and Active Faults

According to historical records, from the 19th Century to 1969, a total of 7 earthquakes with magnitudes greater than Ms3.0 happened in Longtan Reservoir Area. The biggest was Leye Earthquake with a magnitude of 6½ on June 8, 1875. The epicenter of this earthquake was about 56 km away from the damsite [23]. From 1970 when earthquakes were first recorded with instrument to 2002, all regional earthquakes were basically distributed in the periphery of reservoir area. In 1983, an earthquake swarm

happened in Mazhuang Township, Leye County. The biggest earthquake was magnitude M_S4.6. Earthquakes during the period from 2003 to 2005 were concentrated around the dam. As Longtan Reservoir has been dammed and constructed since 2003, it was speculated that these earthquakes were likely to result from blasting during the dam construction or the construction process of dam. Since there were no seismic stations around the reservoir, such a speculation needed to be confirmed with more surveys [24]. On September 30, 2006, Longtan Reservoir began to impound water. With the leap of water level, small earthquakes began to happen in the reservoir frequently. As of May 26, 2013, a total of 3682 earthquakes with magnitudes greater than M_L0.0 were recorded, including 837 earthquakes with M_L1.0–1.9, 128 earthquakes with M_L2.0–2.9, 8 earthquakes with M_L3.0–3.9 and 3 earthquakes with M_L4.0–4.9, i.e., M_L4.0 Luotuo Earthquake on March 7, 2007, M_L4.5 Tian'e Earthquake on July 17, 2007 and M_L4.8 Luotuo Earthquake on September 18, 2010. Zhou et al. [25] used waveform data recorded by Longtan Reservoir Digital Telemetry Seismic Network (including 12 fixed seismic stations and 1 relay station), adopted the double difference algorithm and waveform cross-correlation technology to precisely position earthquakes happening in the reservoir from September 30, 2006 to May 26, 2006. A total of 3074 precise positioned earthquakes were harvested, accounting for 83.5% of total number. From the precise positioning results, the maximum horizontal error was 76.6 m. The maximum vertical error was 88.8 m, obviously superior to positioning results of the Network. In this paper, according to the precise positioning results of medium and small earthquakes in Longtan Reservoir, an epicenter distribution map of earthquakes in reservoir area was drawn. It can be seen from Fig. 3 that after Longtan Reservoir impounded water, seismicity is obviously clustered and mainly distributed in deep water zones flooded by five reservoirs, Luotuo (Cluster I), Bamao (Cluster II), Lalang (Cluster III), Bashou (Cluster IV) and Buliu River (Cluster V) after impoundment. Among them: Cluster I is distributed in Hongshui River from Luotuo to Xialao. F3 goes through this cluster along NNE. The predominant seismic strike is NW. Among 3 earthquakes greater than M_L4.0 happening in the reservoir area after impoundment, 2 went through this cluster. The biggest was an M_L4.8 earthquake on September 18, 2010, which also the largest earthquake since impoundment. Cluster II is distributed around Bamao in the intersection between F2 and Youla River, predominately NW-striking. There are numerous small earthquakes. The biggest was an M_L3.2 earthquake on June 20, 2008 and an M_L3.2 earthquake on October 7, 2010. Cluster III is distributed in Lazhong-Nasha. F4 goes through this cluster and is flooded by Niuhe Waters after impoundment. This cluster was predominately NW-striking. There are also many small earthquakes. The biggest was an M_L3.0 earthquake on January 25, 2010. Cluster IV is distributed in Ladang-Pojie, near F9 and Longtan Dam and nearest to the damsite. The biggest was an M_L4.5 earthquake on July 17, 2007. Since impoundment, earthquakes have been very active in this area. The earthquake strikes are consistent with F9. Cluster V is distributed in the broad west bank of Nayi, Buliu River, south of the dam, near the intersection between F8 and Buliu River, north of an M_L4.8 earthquake in Mazhuang Township, Leye County on December 5, 1983. The distribution of this cluster is consistent with F8 strike. Whether in terms of frequency or intensity, it is the weakest cluster among the five clusters. The biggest was an M_L2.8 earthquake on June 21, 2007.

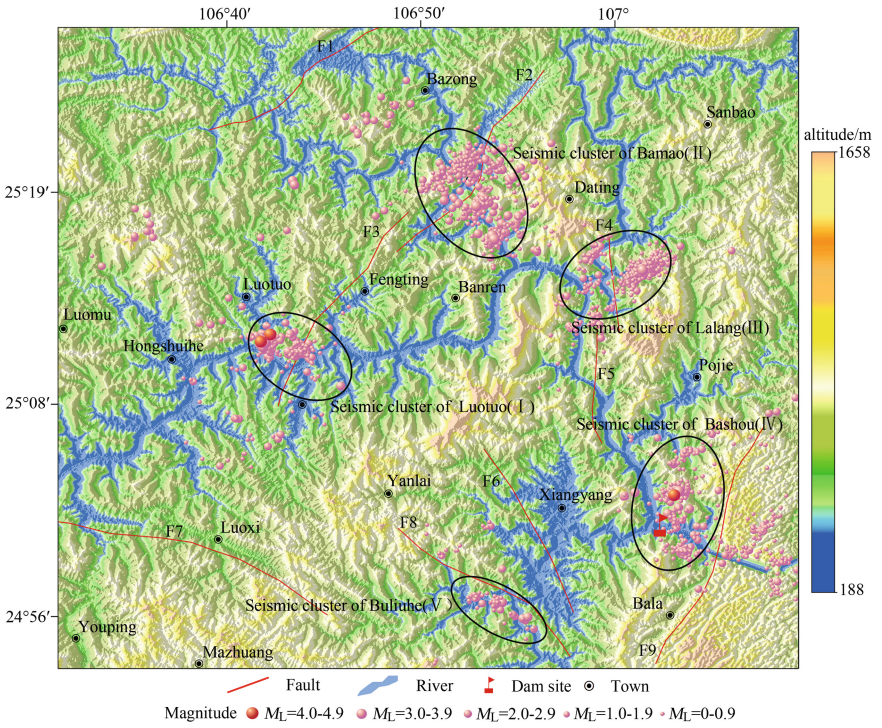


Fig. 3. Epicenter distribution of earthquakes in Longtan reservoir

Statistical results show that about 80% of historical earthquakes with different magnitudes happened 5 km around the fracture buffer zone. Earthquakes recorded with instrument also showed such a proportion [3]. But when analyzing the relationship between earthquakes and faults in 2D space, there was a problem of mapping earthquakes produced by the same active fault to different faults. To analyze 3D distribution relationship between earthquakes and faults intuitively, this paper used focal longitude and latitude, magnitude and focal depth parameters in precise positioning results of small earthquakes in five clusters (Table 2), combined with 3D fracture data of reservoir area and set up a 3D visualization model between earthquakes and faults in the reservoir area (Fig. 4a). From this model, we can see clearly that 3D faults of F3, F2, F4, F9 and F8 go through the center of Clusters I–V. For further quantitative analysis of their relationship, after fully considering the smaller earthquake magnitude in the reservoir area, 3D fault buffer zones were set up within 6 km radius and 20 km depth of each fault (Fig. 4b). Statistical analysis results between 3D fault buffer zones and earthquakes showed that (Table 3) the number of F3, F2, F4, F9 and F8 earthquakes in buffer zones accounted for more than 90% of all clusters and further confirmed that these 5 clusters were controlled seismic faults among all clusters.

Table 2. 3D modeling parameters of reservoir earthquakes

Object_id	Latitude	Longitude	Magnitude	Depth
1	25.240075	106.986699	0.6	7.733
2	25.229869	106.980384	0.8	6.114
3	25.033048	107.053509	0.2	6.984
4	24.996031	107.092338	1.1	6.514
5	25.023142	107.046358	0.4	6.912
6	25.036085	107.047666	0.4	6.025
7	25.1781	106.702753	0.8	7.758
8	25.035416	107.046966	0.9	5.595
9	24.955766	106.901619	1.1	5.166
10	25.328416	106.90835	1.8	6.365
……	……	……	……	……
3074	25.344231	106.882155	1.7	8.335

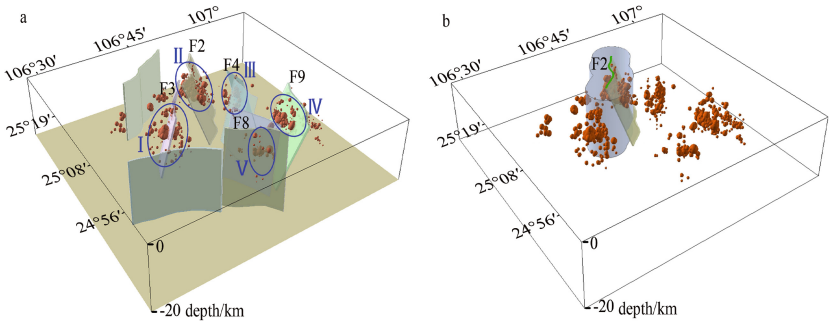


Fig. 4. 3D visualization analysis between Longtan reservoir-induced earthquakes and faults (a) 3D spatial distribution between earthquakes and faults; (b) Spatial distribution between F2 3D buffer zone and earthquakes

Table 3. Statistical results between 3D fault buffer zones and earthquakes

Fault name	3D buffer zone/km	Earthquakes within the buffer zone (percentage in the total number of clusters)
Gaoxu-Bamao Fault (F2)	Radius 6 Depth 20	612 (95%)
Fengting-Xialao Fault (F3)		293 (97%)
Maer-Lalang Fault (F4)		977 (93%)
Changli-Banan Fault (F8)		919 (91%)
Longfeng-Bala Fault (F9)		97 (96%)

4 Relationship Between Focal Mechanism of Medium and Small Earthquakes in Longtan Reservoir Area and Fault Occurrence

Based on studies on the precise positioning of medium and small earthquakes in Longtan Reservoir area, Yan et al. [26] selected waveform data of earthquakes with magnitudes greater than M_L2.0 in the reservoir area recorded by Longtan Reservoir Digital Telemetry Seismic Network and solved focal mechanism solutions of earthquakes one by one, using FOCMEC program. During solution, they set the upper limit of contradictions in first motion symbols as 0. The upper limit of contradictions in amplitude ratio was less than or equal to 1. A total of 120 focal mechanism solutions of earthquakes were obtained. Using focal longitude and latitude, magnitude, focal depth, strike, dip, slip, T-axis dip, P-axis dip and other parameters in 120 focal mechanism solutions of earthquakes, this paper drew 3D spatial focal mechanism solutions (Fig. 5) and conducted a statistical analysis on the focal mechanism type and parameters of 5 clusters. Results showed that Cluster I was dominated by thrust earthquakes. The nodal plane strike was predominately SSW and SWW. The dip was generally abrupt around 40°–70°, basically consistent with F3 fault occurrence parameters through this cluster. In Cluster II, strike-slip and thrust earthquakes accounted for a large proportion. The nodal plane strike was predominately NNE-SSW. The dip was very abrupt around 50°–80°, roughly consistent with F2 strike and dip through this cluster. In Cluster II, both thrust and strike-slip earthquakes accounted for a large proportion. The nodal plane

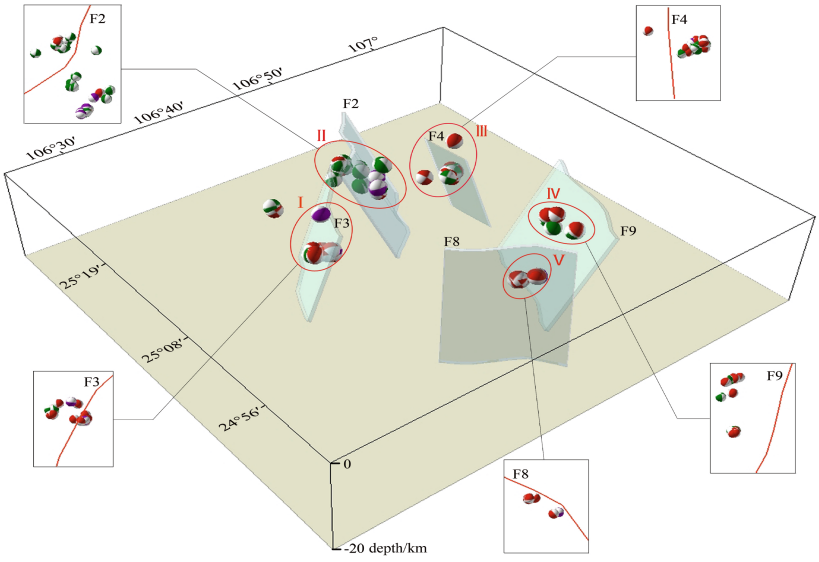


Fig. 5. 3D focal mechanism solutions of ML ≥ 2.0 earthquakes and distribution of active faults in Longtan reservoir area. (Red focal mechanism solution stands for thrust earthquake. Green focal mechanism solution stands for strike-slip earthquake. Purple focal mechanism solution stands normal earthquake.) (Color figure online)

strike was predominately quasi-SN. The dip range was wide around 50°–70°, similar to F4 strike and dip through this cluster. Cluster IV was dominated by thrust earthquakes. The nodal plane strike was predominately quasi-SN and NNE-SSW. The dip was generally abrupt around 50°–70°, roughly consistent with F9 fault occurrence parameters through this cluster. Cluster V was also dominated by thrust earthquakes. The nodal plane strike was scatter, predominately NW. The dip range fell into two intervals: 20°–40° and 60°–90°, similar to F8 fault occurrence parameters through this cluster. To sum up, after Longtan Reservoir impounded water, the focal mechanisms of earthquakes with magnitudes greater than M_L2.0 happened in the reservoir area were dominated by thrust earthquakes. Dips and strikes of focal mechanisms of 5 clusters were all roughly the same as fault occurrence of main faults through these clusters (Tables 4 and 5).

Table 4. 3D modeling parameters of focal mechanism solutions of ML $\geq$ 2.0 earthquakes in the reservoir area

Object_id	Latitude	Longitude	Strike	Dip	Rake	Magnitude	Depth	T_plunge	P_plunge
1	25.1935	106.7041	240.58	46.03	153.26	4	7.4	46.05	15.19
2	25.1772	106.7287	267.79	44.81	135.19	2.7	8.3	58.52	8.65
3	25.1801	106.7303	264.28	39.67	122.41	2.2	8.3	67.73	9.39
4	25.1884	106.7001	14.98	45.21	97.05	2.2	7.6	85	0
5	25.1826	106.6978	61.4	52.24	129.23	3.2	7.4	60	0
6	25.1826	106.6993	58.81	61.98	112.79	2.5	7.7	65.19	14.08
7	25.3228	106.7248	231.83	35.53	126.06	2	8.1	65.18	14.08
8	25.3265	106.7231	278.79	41.02	−168.31	2.8	7.5	26.07	38.86
9	25.1772	106.7257	142.18	28.9	122.38	2.2	7.3	65.18	19.29
10	25.1878	106.7545	262.93	45.22	−175.02	2.2	7.5	27.04	32.79
11	25.1837	106.7288	167.12	48.36	117.24	2.4	8.4	70	0
12	25.1806	106.7285	179.2	22.27	154.49	2.3	8.5	50.33	32.62
13	25.1798	106.7275	125.36	50.15	123.4	2.7	8.3	65	0
14	25.1800	106.7286	253.52	17.97	124.27	2.4	8.4	58.53	29.49
15	25.1941	106.7212	161.27	41.02	105.34	2.2	8.3	78.83	4.93
16	25.1941	106.7212	277.11	80.15	−100.15	2.2	0	34.39	53.78
17	25.1796	106.7296	85	77.76	125.94	2	8.5	45.19	24.18
18	25.1874	106.6965	79.28	66.07	140.67	4.5	7.6	44.14	7.05
19	25.1866	106.6987	254.66	41.02	168.31	2	7.6	38.86	26.07
20	25.1877	106.6966	190	60	90	2.2	8	75	15
……	……	……	……	……	……	……	……	……	……
120	24.9625	106.8973	60.14	33.23	118.18	2.3	6.4	68.91	14.48

5 3D Distribution Relationship Between Longtan Reservoir-Induced Earthquakes Focal Depth and Geological Formation

From the precise positioning results of medium and small earthquakes in Longtan Reservoir area after impoundment, focal depth ranges from Clusters I to V were 0.3–11.3 km, 0.1–10.2 km, 0.2–11.7 km, 0.6–12.7 km and 0.8–9.4 km respectively. Predominant depths were 6–9 km, 4–9 km, 5–9 km, 5–9 km, 6–8 km and 4–9 km

Table 5. Statistical results of focal mechanism solutions

Cluster no.	Number of ML $\geq$ 2.0 earthquakes	Number of earthquakes obtaining focal mechanism solutions	Nodal plane strike predominant distribution	Dip	Earthquake type	Faults through this cluster
I	32	26	SSW,SWW	40°–70°	Mostly thrust earthquakes	F3
II	41	35	NNE,SSW	50°–80°	Largely thrust and strike-slip earthquakes	F2
III	34	33	SN	50°–70°	Largely thrust and strike-slip earthquakes	F4
IV	23	19	SN, NNE-SSW	50°–70°	Mostly thrust earthquakes	F9
V	9	7	NW	20°–40°, 60°–90°	Mostly thrust earthquakes	F8

respectively [25]. Although focal depth ranges and predominant depths of different clusters were slightly different, about 90% earthquakes happened within 5–13 km underground (Fig. 6). As far as rock mechanics are concerned, the sediments above 5 km under Longtan Reservoir were mainly Triassic-Quaternary quasi-flysch, clastic

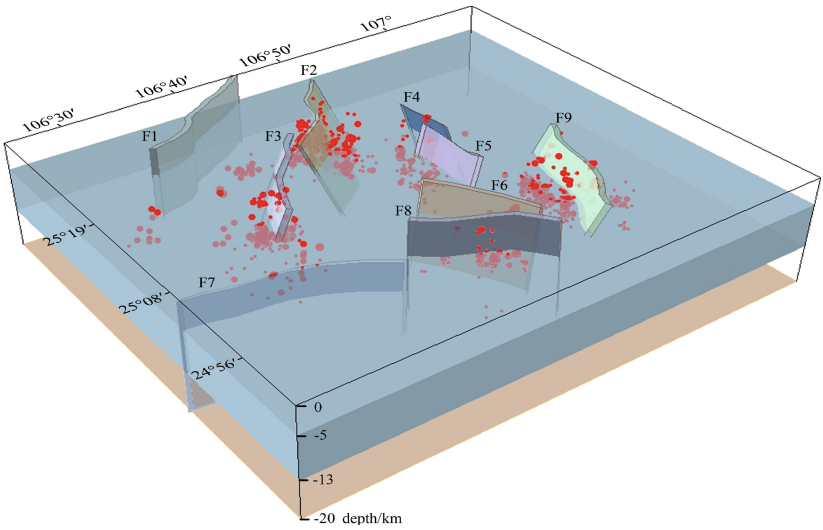


Fig. 6. 3D spatial distribution between focal depths of Longtan reservoir-induced earthquakes and Strata (90% earthquakes in the reservoir area occurred within 5–13 km light blue strata underground.) (Color figure online)

rocks and other ductile rock masses. Within 5–13 km under the reservoir area were mainly brittle rock masses dominated by Sinian-Upper Paleozoic carbonatites. It was easier for brittle rock masses to gather elastic strain energy than ductile rock masses. The in-situ crustal stress was higher than that of ductile rock masses. Especially pre-existing faults and other structural planes in brittle rock masses reached a critical state most easily during the accumulation of tectonic stress. In this way, a small stress disturbance may trigger brittle fracture and release large seismic strain energy. Furthermore, from Zhou et al.'s [25] numerical simulation analysis results of Longtan Reservoir, we know that during the impoundment of Longtan Reservoir, Sinian-Upper Paleozoic strata dominated by carbonatites 5–13 km underground became a principal position where additional head pressure of surface water body diffused to deep strata. To sum up, rock masses 5–13 km under Longtan Reservoir area were prone to brittle failure and thereby induce seismicity.

6 Conclusion

Taking Longtan Reservoir as a target area, using geological formation, fault occurrence, precise positioning of medium and small earthquakes, focal mechanism solution and other data of the reservoir area, this paper sets up a 3D geological model on ARCGIS platform, conducts a 3D spatial analysis on relationships between earthquakes and active faults and between focal mechanism solutions medium and small earthquakes and fault occurrence and discusses 3D spatial relationship between focal depth and geological formation. From the analysis results, taking considering geological formation, precise positioning of small earthquakes, focal mechanism and other seismogeological factors into full consideration, 3D geological models can reflect spatial relationship between earthquakes and faults more truthfully and visually. This provides an important channel for the quantitative and qualitative evaluation of correlation between earthquakes and faults. However, there is still much room for improvements in 3D modeling, for example, by collecting detailed geological profile data of reservoir area, we can build a stratigraphic model using irregular triangle net (TIN), thereby improving the spatial form of stratum surface to different degrees, to make it closer to the real form. Therefore, in future research work, we need to further combine with geophysical prospecting data to make comprehensive analyses and obtain more reliable and intuitive 3D analysis results.

References

1. Deng, Q.D., Zhang, P.Z., Ran, Y.K., et al.: Active tectonics and earthquake activities in China. Earth Sci. Front. **10**(suppl.), 66–73 (2003)
2. Shi, S.Z., Song, L.J., Yang, J.L., et al.: Design and application of geographic information system of active faults in northern Tianshan mountains. Inland Earthq. **12**(4), 367–370 (1998)
3. Qu, C.Y., Ye, H.: Analysis of correlation between fault and earthquake using GIS. J. Seismol. Res. **23**(1), 72–75 (2000)

4. Peng, Z.Z., Zhao, A.P., Hu, C.E., et al.: Research on the relationship based on GIS between the distribution of the active faults and the seismicity in Jiangxi. S. China J. Seismol. **22**(4), 9–18 (2002)
5. Wu, N.F., Zhou, Z.Y., Lao, Q.Y., et al.: GIS-based study on the relationship between earthquakes and active faults in Shanghai and its adjacent offshore region. Geotectonica Metallogenia **28**(3), 248–253 (2004)
6. Zhou, J.: Research on the relationship based on GIS between the distribution of the active faults and the seismicity in Chuandian, Institute of Geology, China Earthquake Administration, Beijing, pp. 1–77 (2005)
7. Wang, J.Y., Ma, B.Q., Qin, L.: GIS-based study on the relationship between earthquakes and active faults in the southern Yellow Sea. Northwest. Seismol. J. **30**(4), 400–404 (2008)
8. Liu, F., Zhang, J.S., Huang, X.N., et al.: Research on the relationship between active faults and earthquakes in the junction area of the China North-South Seismic Belt and Central Orogenic Belt based on GIS. Earthq. Res. China **25**(4), 394–404 (2009)
9. Yu, S.L.: Research on the relation based on ARCGIS between the seismicity and active faults zong in Xiamen and near area. Taiyuan University of Technology, Shanxi Taiyuan, pp. 1–71 (2011)
10. Cui, J.: Study on the relationship between earthquakes and active faults in Yinchuan basin based on GIS. J. Seismol. Res. **37**(3), 385–389 (2014)
11. Houlding, S.W.: 3D Geoscience Modeling Computer Techniques for Geological Characterization. Springer, Heidelberg (1994)
12. Liu, Y.J., Zhao, S.X.: Structure modeling and 3D visualization of active fault. Institute of Crustal Dynamics, CEA. The collection of theses of tectonic and crustal stress, no. 19, pp. 93–101. Beijing (2006)
13. Wu, L.X., Shi, W.Z.: Christopher Gold: Spatial modeling technologies for 3D GIS and 3D GMS. Geogr. Geo-Inf. Sci. **19**(1), 5–11 (2003)
14. Yao, X., Wang, C.C.: Study of 3D model and visibility for mine-bed based on ArcScene. Eng. Geol. Comput. Appl. **41**(1), 15–18 (2006)
15. Zhang, M., Li, Z.H., Liu, H.F., et al.: An ArcGIS based 3D model of Taiyuan Graben Basin in Quaternary. Technol. Earthq. Disaster Prev. **2**(3), 243–248 (2007)
16. Li, Z.H., Liu, H.F., Zhang, M., et al.: 3D visualization and modeling of spatial relationship between earthquakes and active faults. Seismol. Geol. **35**(3), 565–575 (2013)
17. Zhuang, G.: Autonomous Region Bureau of Geology: Nandan Geological Map. Geological Publishing House, Beijing (1968)
18. Zhuang, G.: Autonomous Region Bureau of Geology: Leye Geological Map. Geological Publishing House, Beijing (1972)
19. Li, W.Q.: The relationship between the characteristics of neotectonic regionalization and earthquakes in Guangxi. S. China Seismol. J. **9**(4), 22–26 (1989)
20. Xiang, H.F., Zhou, Q.: Review report of ground motion parameters of red river Longtan Hydropower Station in Guangxi. Institute of Geology, China Seismological Bureau (2006)
21. Guo, P.L., Yao, H., Yuan, Y.: Analysis on potential seismic risk in Longtan Reservoir. Earthq. Res. Plateau **18**(4), 17–23 (2006)
22. Chen, H.L., Zhao, C.P., Xiu, J.G., et al.: Study on the characteristics of focal mechanisms of reservoir induced earthquakes and stress field in the Longtan reservoir area. Seismol. Geol. **31**(4), 686–698 (2009)
23. Shi, S.P., Yu, X.Q., Long, Z.Q., et al.: Analysis on the nature of earthquake in Longtan area after the reservoir storage water. Seismol. Geomagn. Obs. Res. **31**(3), 40–45 (2010)
24. Chen, H.L., Zhao, C.P., Xiu, J.G., et al.: Study on precise relocation of Longtan reservoir earthquakes and its seismic activity. Chin. J. Geophys. **52**(8), 2035–2043 (2009)

25. Zhou, B., Sun, F., Yan, C.H., et al.: 3D-poreelastic finite element numerical simulation of Longtan reservoir-induced seismicity. Chin. J. Geophys. **57**(9), 2846–2868 (2014)
26. Yan, C.H., Zhou, B., Lu, L.J., et al.: Research on focal mechanism of moderate and small earthquakes occurred after reservoir recharge in Longtan reservoir region. Chin. J. Geophys. **58**(11), 4207–4222 (2015)

Identification and Characterization of Geological Hazards in a Coal Mining Area Using Remote Sensing

Jin Liu(✉)

Geological Environmental Center of Shanxi Province, Taiyuan 030024, China
sxliujin2014@163.com

Abstract. Multi-source datasets, including GF-1 remote sensing image, Digital Elevation Model (DEM), basic geographic information, were used to analyze the geological hazards in Liliu coal mining area located in the western Shanxi Province, China. A total of six geological hazards were identified and characterized including collapse, landslide, unstable slope, debris flow, ground subsidence and ground fissure. A combination method with object-oriented and man-machine interactive way was used to identify the geological hazards, and then the results were validated by field survey for obtaining their spatial distribution and incidence features in the study area. The results show that a total of 1096 geological disasters are found, in which the number of unstable slope is 39, landslides is 420, collapse is 316, debris flow is 3, ground subsidence is 212, and ground fissure is 106. Furthermore, the intensity of geological hazards was analyzed and the relationship between geological hazards and landforms was also investigated by the GIS spatial analysis for presenting the harmfulness of disaster points to human lives. The intensity of disaster points are medium and large and they mainly developed in the slope form 5° to 35°, with more distribution in the western and southwestern slopes. The disasters have more influences on the roads and farmlands compared with the rivers.

Keywords: Remote sensing · Coal mining area · Geological disasters · GF-1 · Object-oriented classification

1 Introduction

Coal is one of the most important energy sources in the world. It has to be dug out due to the storage condition in the ground. In recent years, the coal mining areas have been developed rapidly within the driving forces of industry, agriculture, tourism, transportation, urban construction, etc. The frequent human activities of mining and infrastructure construction such as road, housing, have a huge impact on the geological environment and the region's economic and sustainable development. Various geological disasters have been induced because of intensive human activities. Great losses have been cased in human lives and properties. It is of practical significance of providing an early warning and preventing from geological disasters in a coal mining area.

Consequently, it is of great significance to assist in the promotion of disaster prevention by analyzing the characteristics and causes of geological disasters for a coal

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 321–329, 2017.
DOI: 10.1007/978-981-10-3969-0_36

mining area. In recent years, some progresses have been made on information extraction of different geological disasters based on the 3S technology [1–3], etc., which has been approved to be a reliable technical measurement for quickly identifying geological disasters. Nevertheless, the development characteristics and causes of geological disasters are mainly investigated after an earthquake in previous studies [5–7]. Conversely, corresponding studies on the feature extraction of geological disasters have not been fully conducted for a coal mining area [8–10], etc.

The development of remote sensing has greatly facilitated the interpretation and analysis of geological disasters. In our study, a typical disaster-occurrence area - Liliu coal mining area was used as the study area and a combination method with object-oriented and man-machine interactive way was used to identify the geological hazards. A total of six types of geological hazards were quantitatively identified and analyzed. More specifically, the disasters included collapse, landslide, unstable slope, debris flow, ground subsidence and ground fissure, with an interpretation accuracy of 1:50000 and an area of 820 km^2.

2 Materials and Methods

2.1 Study Area

Liliu coal mining area is located in the western Shanxi Province between 110° 45′ to 111° 15′ eastern longitudes and 37° 20′ and 37° 30′ northern latitudes. It belongs to the loess landform with plenty of ravines and sparse vegetation, which provides a good development foundation for the slope geological disaster. The region has rich coal resources and well-developed mining industry, which leads to land subsidence and ground fissures hazards. A large-area subsidence sliding zones have been formed due to the subsidence induced by mountain slumping. At the same time, excessive exploitation of coal resources have promoted the region's rapid economic development, but slope cutting of various construction engineering have increased the number of unstable slope, which is prone to causing slope geological disasters.

2.2 Data Sources and Analysis Software

GF-1 remote sensing image and 1:50000 topographic map were taken as the primary data sources and Map GIS was used as the image processing and spatial analysis platform. After preprocessing the image, the orthophoto map of the study area was obtained, and meanwhile the 25 m resolution DEM was generated by 1:50000 topographic map.

2.3 Land Cover Classification Schemes

(1) Man-machine classification
In general, there are mainly four steps to carry out the man-machine classification of geological disasters (Fig. 1).

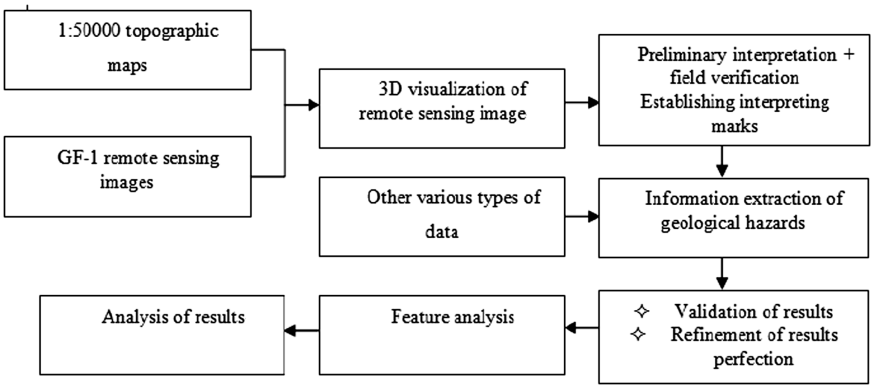


Fig. 1. Overall technical workflow of man-machine classification

(1) On the Map GIS platform, orthophoto map was overlaid by several geographical elements, e.g. roads, residents, rivers, note information, etc., for providing harmfulness background data in analyzing the disaster points;
(2) Three dimensional (3D) terrain visualization model was generated by the Digital Elevation Model (DEM) and Digital Orthophoto Map (DOM) after finishing the registration, which was an extremely useful auxiliary perspective for identifying geological disasters rapidly;
(3) Based on the above work, the specific steps of identifying geological disasters included preliminary interpretation, field verification, amendments for improving interpretation signs, detailed interpretation, field verification and refinement of results;
(4) Buffer analysis based on the GIS spatial analysis function was used to analyze and evaluate the characteristics of development and spatial distribution of disaster points.

(2) Object-oriented disaster classification

The ENVI EX commercial software developed by ITT (ENVI EX) was used to perform object-oriented disaster segmentation (Fig. 2). There are mainly five necessary steps to complete the classification, including choosing the scale parameter; merging the object primitives; refine the objects using a threshold value for just one band of the image and it is an optional step; extraction of the attributes; object classification based on rules or examples. The first step of the process was to choose the appropriate scale for the segmentation process. Several trials were performed with different scale values ranging from 20 to 65 and the best segmentation was obtained with a scale value of 30. The second step was the objects merge and like for the first step, several trials were made. The best merge values for a scale parameter of 35 for our study area was 90.

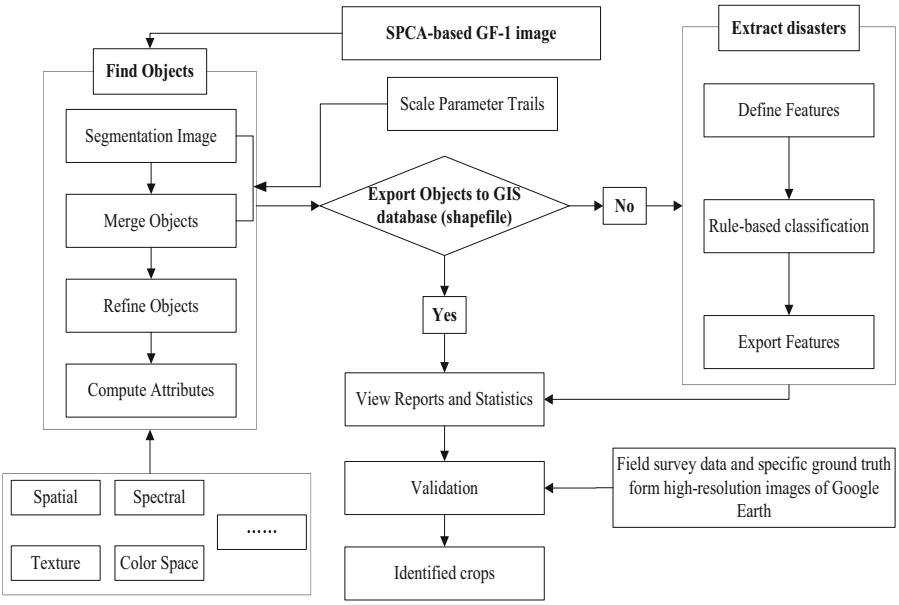


Fig. 2. Schematic representation of the disaster identification from the object-oriented segmentation process by using the feature extraction module (ENVI EX)

3 Results and Analysis

3.1 Identification of Geological Disasters

The preliminary interpretation was firstly performed and a total of 533 disaster points were identified. They specifically included 44 unstable slopes, 54 landslides, 77 collapses, 219 ground subsidence, 96 ground fissures and 2 debris flows. Furthermore, the disaster-intensive areas were selected for the field validation. The length of verification route was 127 km and 42 geological disasters were investigated by positioning the location using a GPS (Figs. 3, 4 and 5). The verified disasters accounted for 7.88% of

Fig. 3. Image characteristics of landslide and corresponding photograph

Fig. 4. Image characteristics of collapse and corresponding photograph

Fig. 5. Image characteristics of unstable slope and corresponding photograph

the total disaster points of preliminary interpretation. More specifically, the validated geological disasters included 20 landslides, 14 collapses, 2 unstable slopes, 12 ground subsidence and 6 ground fissures. Object-oriented classification was further carried out after the field verification. The number of geological hazards and hidden danger points were 1096, in which there were 39 unstable slopes accounting for 3.56%; 316 collapses accounting for 28.83%, 420 landslides accounting for 38.32%; 3 debris flows accounting for 0.27%, 212 surface collapses accounting for 19.34%; and 106 ground fissures accounting for 9.67%.

3.2 Characteristics of Geological Disasters

Intensity of Geological Disasters. The intensity of geological disasters was classified based on the Monitoring Specification of Rockfall, Landslide and Debris flow and Basic Requirements of County (City) Geological Disaster Investigation and Regionalization (Table 1).

Table 1. Intensity statistics of geological disasters in Liliu coal mining area

Intensity	Disaster types						
	Landslide	Collapse	Surface collapse	Ground fissures	Unstable slope	Debris flow	Total
Over-size	13	2	0	0	3	0	18
Large-size	232	148	0	0	21	2	403
Medium-size	164	161	49	0	14	1	389
Small-size	11	5	163	106	1	0	286
Total	420	316	212	106	39	3	1096

We can find that there are mainly large- and medium-sized disasters. The number are 403 and 389 and they account for 36.77% and 35.49%, respectively. Conversely, the number of over-sized disaster is only 18, and all of them are slope geological disasters in which landslide disaster (13 points) is the most serious. The reasons are that the earth's surface is covered with loess and severe underground mining activities are being performed, which is also consistent with the actual situation in the study area.

Relationship of Geological Hazards and Geomorphic Features. Overlay analysis and buffer analysis were taken on the interpretation disasters in the study area, and the spatial relationship between disaster points and slope and aspect were obtained (Tables 2 and 3).

Table 2. Relationship of geological disasters and slope in Liliu coal mining area

Slope	Disaster types						
	Collapse	Landslide	Unstable slope	Debris flow	Surface collapse	Ground fissures	Total
0–5°	17	29	1	1	13	10	71
5°–15°	152	208	22	1	100	45	528
15°–35°	136	161	14	1	89	46	447
35°–55°	10	20	2	0	10	5	47
>55°	1	1	0	0	0	0	2

Table 3. Relationship of geological disasters and aspect in Liliu coal mining area

Aspect	Disaster types						
	Collapse	Landslide	Unstable slope	Debris flow	Surface collapse	Ground fissures	Total
North	47	60	5	0	31	13	156
Northeast	37	61	4	0	30	12	144
East	26	36	5	0	25	15	107
Southeast	23	28	5	1	12	5	74
South	28	38	5	0	13	8	92
Southwest	53	78	2	1	30	16	180
West	56	70	6	1	33	18	184
Northwest	46	48	7	0	38	19	158

It is obvious that disasters are mainly developed within the slope from 5° to 35°, a total of 975 disaster points accounting for 88.96% of the total hazards, which dominates the most compared with other slope segments (Table 2). Thus, the slope section can be considered as the high-risk area in the remote sensing based interpretation of geological disasters.

As can be seen in Table 3, generally, disasters in different aspects are basically distributed evenly, but there are a little more in the west and southwest are compared to others.

Hazardous Characteristics of Geological Disasters. Disaster points are prone to being located within a certain distance from residential areas, roads and rivers, which determine the number of hazards affected by the three places of intensive human activities (Tables 4, 5 and 6).

Table 4. Relationship of geological disasters and roads in Liliu coal mining area

Distance to roads (m)	Disaster types						
	Collapse	Landslide	Unstable slope	Debris flow	Surface collapse	Ground fissures	Total
0–50	25	37	5	3	23	20	113
50–100	65	55	6	0	43	18	187
>100	226	328	28	0	146	68	796

Table 5. Relationship of geological disasters and rivers in Liliu coal mining area

Distance to rivers (m)	Disaster types						
	Collapse	Landslide	Unstable slope	Debris flow	Surface collapse	Ground fissures	Total
0–50	8	6	0	0	3	0	17
50–100	24	27	1	0	13	0	65
>100	284	387	38	3	196	106	1014

Table 6. Relationship of geological disasters and residential areas in Liliu coal mining area

Distance to Residential areas(m)	Disaster types						
	Collapse	Landslide	Unstable slope	Debris flow	Surface collapse	Ground fissures	Total
0–500	97	158	12	0	76	30	373
500–1000	185	220	26	3	119	67	620
1000–1500	30	37	1	0	16	9	93
>1500	4	5	0	0	1	0	10

According to the Table 4, a total of 113 disaster points are located within 50 m on both sides of roads, accounting for 10.31% of the total disaster points, and a total of

796 points are located within 100 m away from roads, accounting for 72.63%. It shows that the disaster points of the study area have an impact on roads.

As shown in Table 5, a total of 17 disaster points are located within 50 meters on both sides of rivers, accounting for 1.55% of the total disaster points, and a total 1014 points are located within 100 m away from rivers, accounting for 92.52%. It shows disaster points of the study area have also little impact on rivers.

It shows that a total of 373 disaster points are located within 500 m from residential areas, accounting for 34.03% of the total disaster points, and most points are located within the scope of 500–1000 m from residential areas, accounting for 56.57%. Therefore, we should monitor geological disasters more often around the residential distribution.

4 Conclusion

A total 1096 geological disasters and hidden points were identified eventually in Liliu coal mining area through GF-1 remote sensing image and field investigation. Several conclusions can be drawn as follows:

(1) According to the statistics, the intensity of disasters is mainly the large- and medium-sized hazards, accounting for 36.77% and 35.49% of total disaster points, respectively. Then, small-sized hazards accounts for 26.09%. Conversely, over-sized disaster points are only 18 and accounts for 1.64%.
(2) The spatial analysis between hazards and geomorphic factors are performed. We find that the hazard points are mainly concentrated in the slope from 5° to 35°, and a little more are distributed on the west and southwest in aspect. Combination of slope and aspect can be used to extract geological disasters by overlaying high-resolution, real-time remote sensing image.
(3) The buffer analysis are also carried out between hazards and road, river and residential areas in the study area. It shows that disasters have more impact on roads and less impact on rivers. The number of hazards surrounding residential areas within 1000 m account for 90.6%, which can have an impact on the surrounding cultivated land and residential areas.

References

1. Li, C.Z., Nie, H.F., Wang, J., et al.: A remote sensing study of characteristics of geological disasters in a mine. Remote Sens. Land Resour. **1**, 45–48 (2005)
2. Zhang, M.M., Li, J., Xue, Y.A.: Mining geological disaster monitoring in South Suburb of Datong based on 3S technology. Saf. Coal Min. **43**, 203–205 (2012)
3. Lu, X.J., Shi, Z.C., Shang, W.T., et al.: The method and application of multi-dimension interpretation for landslides using high resolution remote sensing image. J. Image Graph. **19**, 141–149 (2014)
4. Tong, L.Q., Guo, Z.C.: A study of remote sensing image features of typical landslides. Remote Sens. Land Resour. **25**, 86–92 (2013)

5. Huang, R.Q., Wang, Y.S., Pei, X.J., et al.: Characteristics of co-seismic landslides triggered by the Lushan Ms7.0 Earthquake on the 20th of April, Sichuan Province, China. J. Southwest Jiaotong Univ. **48**, 581–589 (2013)
6. Fang, C.G., Yang, X., Yang, W.N., et al.: Spatial feature analysis of geo-hazard based on RS and GIStechnology after Wenchuan earthquake. Appl. Res. Comput. **30**, 291–294 (2013)
7. Li, L.J., Yao, X., Zhang, Y.S., et al.: RS-based extraction and distribution characteristics of geo-hazards triggered by Wenchuan earthquake in Mianyuan River Basin. J. Eng. Geol. **22**, 46–55 (2014)
8. Xue, Y.A., Zhang, M.M., Li, J., et al.: Research of 3S technology in monitoring geological disasters in coal-mining area. Disaster Adv. **5**, 427–432 (2012)
9. Lü, Y.Q., Liu, H.F.: Developmental characteristic and cause reasons of geological disaster of compacting density in Fengmaoding. Coal Technol. **30**, 164–166 (2011)
10. Hua, X.Q., Huang, J.J., Miao, S.X., et al.: Distribution and causes of Geo-hazards in Xuzhou. J. Geol. Hazards Environ. Preserv. **26**, 74–80 (2015)
11. XY, Li, M.H., Wang, D.W., et al.: Characteristics and genetic mechanism of coal mining geo-hazards. Coal Technol. **35**, 137–139 (2016)

Monitoring Landslides Using Multi-frequency SAR Data in Danba County, Sichuan Province, China

Yansheng Ding[1], Jie Dong[2(✉)], Lu Zhang[2], Mingsheng Liao[2,3], and Yang Zhou[4]

[1] Southwest China Branch of State Grid Corporation of China, Chengdu 610041, China
[2] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
dongjie@whu.edu.cn
[3] Collaborative Innovation Center for Geospatial Technology, Wuhan University, Wuhan 430079, China
[4] Beijing North-Star Technology Development Co., Ltd., Beijing 100120, China

Abstract. Danba County, located at the northwestern Sichuan, is one of the areas prone to severe landslides in China. The landslides in this area present great threaten upon the local public safety and the famous heritage architectures. Therefore, monitoring landslide is of great importance for the sustainable developments in Danba. In this paper, InSAR techniques is employed to investigate typical landslide activities, based on the multi-frequency SAR images acquired from C-band Envisat, X-band TerraSAR-X, and L-band ALOS-1/2 satellites. Firstly, differential InSAR (D-InSAR) is used to recognize known landslides and find potential unstable slopes in a region scale. Then, for a specific landslide, advanced multi-temporal InSAR method is exploited to characterize its surface deformation by obtaining time-series displacement on coherent targets. Furthermore, the triggering factors are discussed based on the deformation results and on-site surveys.

Keywords: Landslides · InSAR · Multi-frequency · PSInSAR · Deformation monitoring

1 Introduction

Disasters, caused by landslide, rock fall, debris flow, ground fissure, etc., are one of the significant natural catastrophes, threatening and influencing the socio-economic conditions around the world. China is one of the countries that suffer heavily from such geo-hazards. And most landslide sites are located in the mountainous and valley areas of western China due to several factors such as rough terrain, vulnerable geological

This work was financially supported by 2015 science and technology project of SGCC Southwest China Branch and the National Key Basic Research Program of China (Grant No. 2013CB733205).

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 330–338, 2017.
DOI: 10.1007/978-981-10-3969-0_37

environment, and complicated meteorological condition. Landslides in these areas are usually close to human settlements, and thus pose great threatens upon public safety and social economy in local and vicinity regions. Furthermore, landslides may also have significant impacts on the construction of large infrastructures like highway, railway, airport, dock, power grid, oil pipeline, etc. by increasing the financial and time costs as well as risk of damage. Therefore, detecting and then monitoring landslides is of great importance for the sustainable developments in these areas.

Radar remote sensing techniques, such as differential InSAR (D-InSAR), have already proven its potential for remotely monitoring unstable slopes, with its wide coverage and sub-centimeter accuracy [1]. However, landslide monitoring with D-InSAR is often limited by inaccurate external DEM, geometrical and temporal decorrelation and atmospheric phase screen (APS). In consequence, these limits make it very difficult to interpret the landslides quantitatively, especially in mountainous areas with steep slopes and dense vegetation [2].

In the last decade, multiple InSAR methods were developed to overcome these limits, such as Permanent/Persistent Scatterer SAR Interferometry (PSI) [3–6]. PSI technique exploits persistent scatterers (PSs) exhibiting high phase stability in a stack of interferograms generated with the same master image. These PSs, mainly corresponding to buildings and exposed rocks, generally exist in urban area. Reliable results can be obtained by only focusing on the stable PS points [7, 8].

In this paper, we pay concentrate on Danba County, Sichuan Province covered by steep mountains. There are many unstable slopes in this region. Multiple-frequency SAR data, X-band TerraSAR, C-band Envisat, and L-band ALOS-1/2, is collected for our experiments.

We implement the mature D-InSAR technique to detect both known landslides and potential unstable slopes and get an initial information for the landslides, such as type, location, coverage, and so on.

In our project, D-InSAR method is first implemented to locate the unstable landslides and provide some initial knowledge for the landslides, such as location and coverage. The advanced multi-temporal InSAR methods are then used for further analysis [9–13]. Both qualitative and quantitative evaluations of the results are carried out together.

2 Methodology

2.1 Differential InSAR

D-InSAR technology is an effective tool to map landslides, with its wide coverage and sub-centimeter accuracy [8]. With two SAR images, the differential interferometric phase after removing the topography phase component can be expressed as follows:

$$\phi_{\text{diff}} = \phi_{\text{def}} + \Delta\phi_{\text{topo}} + \phi_{\text{atm}} + \phi_{\text{n}} \tag{1}$$

where the deformation phase ϕ_{def} responses the LOS displacement of the ground target between the two SAR acquisitions, $\Delta\phi_{\text{topo}}$ is the residual topography phase due to the

external DEM error, the atmospheric phase component ϕ_{atm} represents the phase delay caused by the different atmosphere phase screen in repeat-pass InSAR, and the ϕ_{n} is the random noise phase possibly induced by the thermal noise or data processing errors.

2.2 Persistent Scatterer in SAR

In the case of landslide monitoring in rural regions, such as mountainous areas, D-InSAR method is often limited by inaccurate external DEM, spatial and temporal decorrelations and atmospheric phase screen (APS). In order to overcome these limitations, advanced persistent scatterer InSAR (PS-InSAR) is developed. The basic concept of PS-InSAR is to identify stable point-like targets from a stack of SAR images. These points are minimally affected by spatial and temporal decorrelations [3–6].

A stack of interferograms with a common master image are formed from multiple SAR images. As amplitude dispersion index is a good representation of phase stability, it is employed to select PS candidates. The amplitude dispersion value is defined as:

$$D_A = \frac{\sigma_A}{\mu_A} \tag{2}$$

where σ_A and μ_A are the standard deviation and mean of the calibrated time series amplitude values, respectively. Then, based on the different characteristics of the components in Eq. (1), the differential phase is filtered spatially and temporally to obtain the final deformation term.

3 Study Area and SAR Data

3.1 Study Area

Danba County is located in Ganzi Canton, Sichuan Province with latitude N30.7–31.2 and longitude E101.5–102.0. This region, belonging to Minshan-Qionglai alp, western Sichuan alp and plateau areas, has a complicated terrain with altitude ranging from 1700 m to 5520 m. As the key forest district in southwest China, Danba has various dense forest vegetations, including approximately 46% composed of short shrubs, as shown in Fig. 1. It also owns rich water resource with hundreds of rivers and alpine lakes, and the famous one is Dadu River. Besides, the rainfall is moderate and concentrated in the summer.

There are many classical metamorphic and strongly tectonic deformations in these complicated geology backgrounds. Because of its plateau monsoon climate along with the strong polymetamorphism and special tectonics, rock fall, landslides and debris flow break out frequently and widely. The frequent human activities, road construction, building expansion, and irrigation, also cause new landslides or reactivate some old landslides. The most dangerous and urgent issue is that many villages and provincial road are just located on the body or underneath of the landslides.

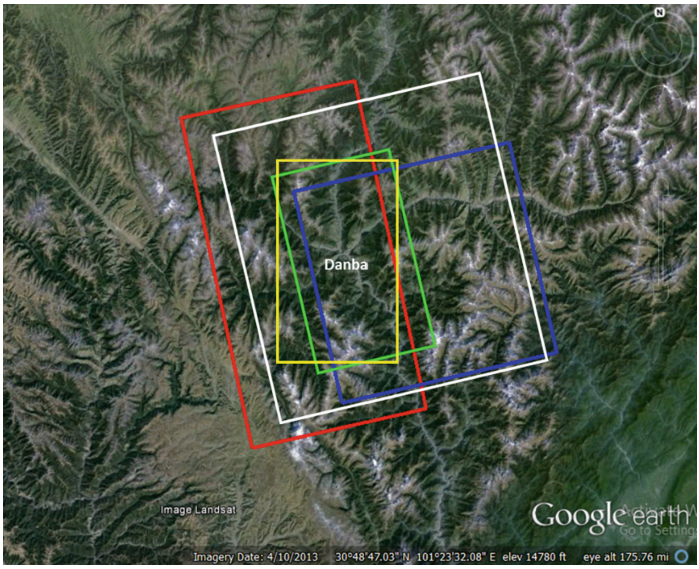


Fig. 1. Location of our research area on Google earth and the images coverages. The green and yellow rectangles indicates the ascending and descending orbits of TerraSAR-X respectively, the red rectangle donates the ascending orbit of Envisat ASAR, the white and blue ones represent the ascending orbits of ALOS-1 and ALOS-2 separately. (Color figure online)

3.2 SAR Data

In our experiments, multi-frequency SAR datasets from three satellites, with different spatial resolutions and coverages, are considered.

- X-band SAR images: 4 TerraSAR-X stripmap images acquired in 2016: 2 ascending and 2 descending;
- C-band SAR images: 9 Envisat ASAR images from ascending orbit acquired from August, 2007 to June, 2008;
- L-band SAR images: 19 ALOS-1 level 1.0 raw images from ascending orbit acquired between December, 2016 and January, 2011 and 3 ALOS-2 level 1.1 SLC images also from ascending orbit acquired in 2015.

The coverages of these satellite images are superposed on the Google earth in Fig. 1. For the ALOS-1 raw images consisting of fine-beam dual-polarization (FBD) and single-polarization (FBS) modes, the FBD raw data has to be oversampled to the pixel spacing of FBS, and then they are focused to generate single-look complex SAR images for interferometric processing. As mentioned before, reference DEM must be used to remove the topographic phase in the D-InSAR process. Here, 90 m-resolution SRTM is used as reference DEM in our study.

4 Real Data Analysis

To inspect the landslides in Danba County, differential interferograms are first formed for all the three bands data. Figure 2 shows the differential interferograms for two pairs of TerraSAR-X images. We can find that the phases are noisy and coherences are very low. This is because the shorter wavelength of X-band has weak penetration to the covering vegetation. In addition, the ascending has a higher coherence than the descending one due to the shorter spatial baselines.

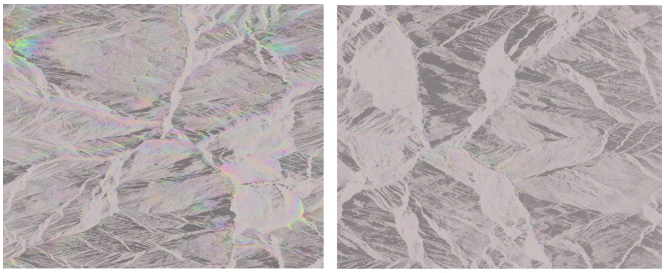


Fig. 2. Differential interferograms for TerraSAR-X. (a) Ascending: date: 20160417–20160509, temporal interval: 22 days, perpendicular baseline: 160 m; (b) Descending: date: 20160330–20160421, temporal interval: 22 days, perpendicular baseline: 225 m.

On the other hand, as shown in Fig. 3, higher coherences are obtained for C- and L- band images, because of the longer wavelengths. Some small phase fringes marked by white dashed curves are both found in 2008 and 2015. These phase fringes indicate the appearance of unstable slopes. Although D-InSAR method is able to detect unstable slopes, it is cannot give the deformation quantity and velocity.

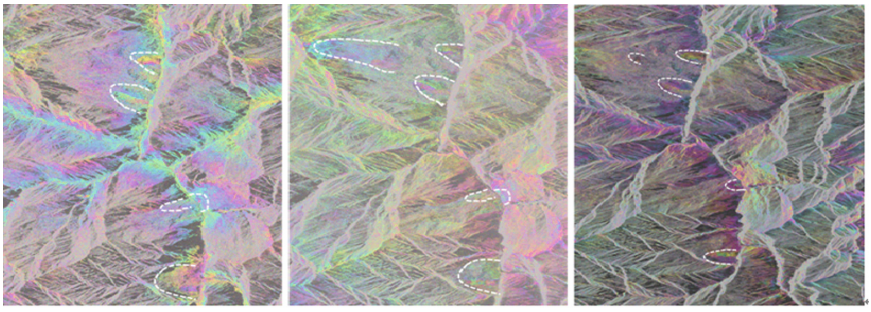


Fig. 3. Differential interferograms for differential satellite data in Danba County, white curves indicate the unstable slopes. (a) Envisat: date: 20080208–20080314, temporal interval: 36 days, perpendicular baseline: 95 m; (b) ALOS-1: date: 20080627–20090630, temporal interval: 369 days, perpendicular baseline: 346 m; (c) ALOS-2: date: 20150731–20151218, temporal interval: 138 days, perpendicular baseline: 276 m. (Color figure online)

With the multiple images provided by Envisat and ALOS-1, PS-InSAR technique is deployed to operate time series processing and analysis. We use StaMPS program to process the data and obtained the deformation velocity for the two datasets. The temporal-spatial baselines are shown in Fig. 4. The red point indicates the master image, with the black points indicating the slave images. The deformation density maps not only reveal the geographical locations of landslides, but also give their corresponding deformation magnitude. The found unstable slopes, i.e. the areas with red or blue PS points, show fine consistency in the two data stacks.

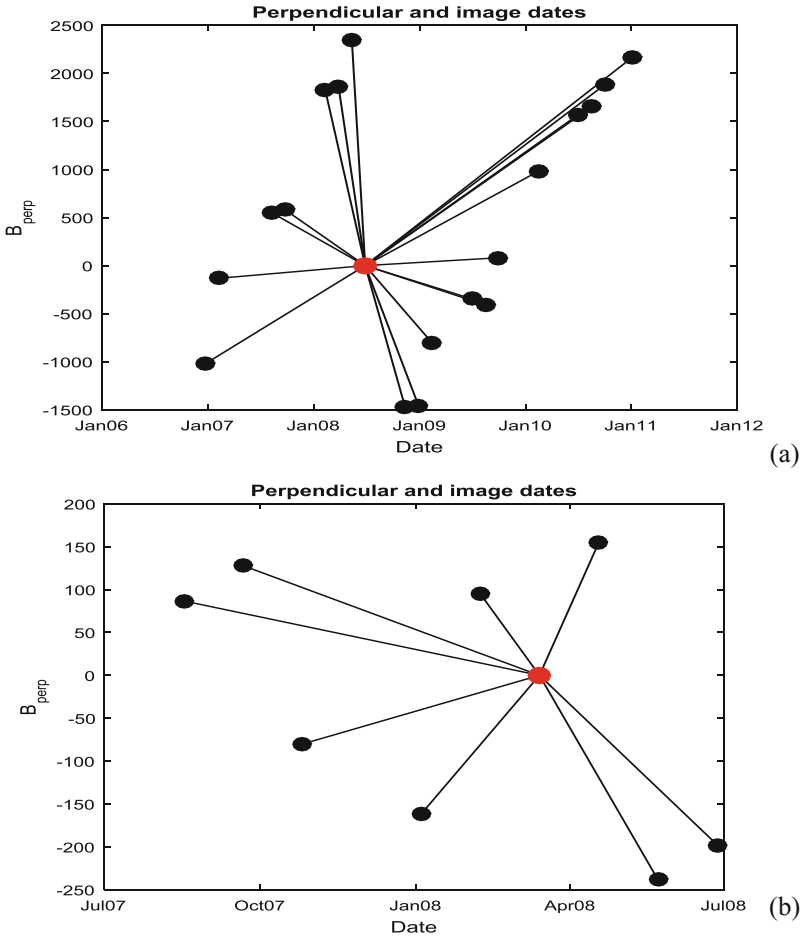


Fig. 4. Temporal-spatial baselines for Envisat and ALOS-1 datasets. (a) ALOS-1; (b) Envisat. (Color figure online)

The PS points of the ALOS-1 data stack is uniformly distributed with very high density. However, the PS point density for Envisat is sparser over the whole area. The reason behinds this phenomenon is that the L-band sensor is able to penetrate the

vegetation and bush. So, in terms of landslides monitoring in rural areas, especially in complex mountainous environment, longer radar wavelength is appreciated to achieving high PS point density.

For the Jiaju landslide, the deformation velocity in the LOS was about 55 mm/y for ALOS-1, but near 70 mm/y for Envisat. The difference may be caused by the different incidence angles and the temporal coverages. The large slope, where the famous Jiaju Tibetan residence located on, were unstable in several parts. The displacement in the bottom was mainly induced by the river degradation, excavation for building roads, and village extension. The wide-range displacement in the top of slope, with max altitude approximately 4390 m, was not detected before. The deformation triggering factors need to be further studied.

The Suopo landslide moved toward to the satellite with LOS deformation velocity about 65 mm/y for ALOS-1 and 50 mm/y for Envisat. The landslide is primarily caused by the local human activities, such as irrigation and road construction. It worth mentioning that many historic Stone Towers on the southern part of Suopo slope, suffer heavily from the local landslide and some of them even became slant.

For the other unstable slopes, with similar deformation velocity to Jiaju landslide, present a large threat to the provincial road S303 under them. From the on-site survey, many cracks were found in the road, which may be ascribed to the extrusion force the sliding slopes (Fig. 5).

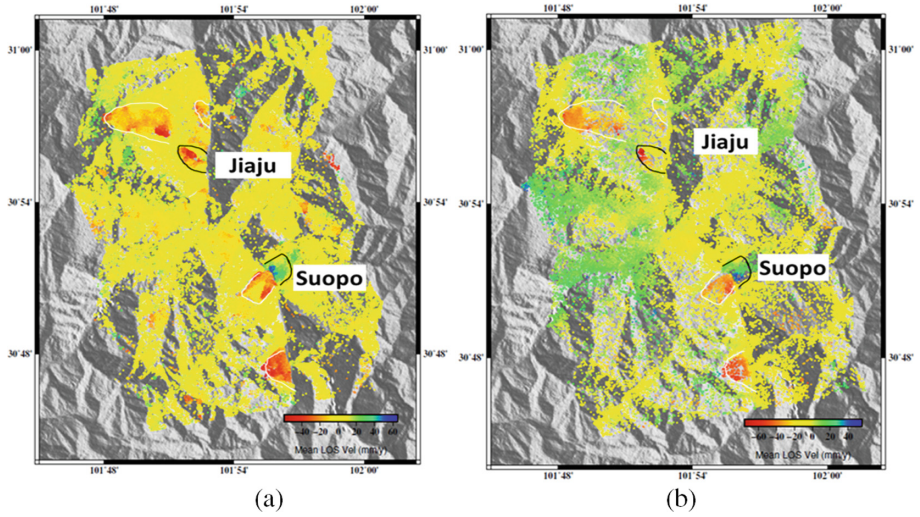


Fig. 5. Deformation velocities processed by StaMPS/PSInSAR in Danba County. (a) ALOS-1, the image acquired on 27th June, 2008 is selected as master, and the perpendicular baseline varies from 82 m to 2350 m, (b) Envisat, the image acquired on 14th March, 2008 is selected as master, and the perpendicular baseline varies from 80 m to 239 m. The PS points of the ALOS-1 data stack is uniformly distributed with very high density. However, the PS point density for Envisat is sparser over the whole area. A consistent results are expected in both data set.

5 Conclusions

In this study, several known landslides and potential unstable slopes were studied in Danba County by using InSAR techniques. The Traditional D-InSAR method was able to detect unstable slopes in a wide region, but restricted to geometrical and temporal decorrelations. The advanced InSAR technique, PS-InSAR, can retrieve the deformation trends of landslides on the stable PS measurement points. In addition, SAR images with longer wavelength are more suitable for landslide monitoring in mountainous areas.

Acknowledgments. The authors thank DLR for providing the TerraSAR-X datasets through the General AO project (GEO0606), ESA for providing Envisat ASAR data through the Dragon-3 project (ID 10569), and JAXA for providing the ALOS-1/2 datasets through ALOS RA4 project (No. 1247 and 1440).

References

1. Crosetto, M., Crippa, B., Biescas, E., Monserrat, O., Agudo, M., Fernández, P.: Land deformation measurement using SAR interferometry: state-of-the-art. Photogrammetrie Fernerkundung Geoinformation **2005**, 497 (2005)
2. Liu, P., Li, Z., Hoey, T., Kincal, C., Zhang, J., Zeng, Q., et al.: Using advanced InSAR time series techniques to monitor landslide movements in Badong of the Three Gorges region, China. Int. J. Appl. Earth Obs. Geoinf. **21**, 253–264 (2013)
3. Ferretti, A., Prati, C., Rocca, F.: Permanent scatterers in SAR interferometry. IEEE Trans. Geosci. Remote Sens. **39**, 8–20 (2001)
4. Ferretti, A., Prati, C., Rocca, F.: Nonlinear subsidence rate estimation using permanent scatterers in differential SAR interferometry. IEEE Trans. Geosci. Remote Sens. **38**, 2202–2212 (2000)
5. Hooper, A., Segall, P., Zebker, H.: Persistent scatterer InSAR for crustal deformation analysis, with application to Volcán Alcedo, Galápagos. J. Geophys. Res. **112**, B07407-1 (2007)
6. Hooper, A., Bekaert, D., Spaans, K., Arıkan, M.: Recent advances in SAR interferometry time series analysis for measuring crustal deformation. Tectonophysics **514**, 1–13 (2012)
7. Colesanti, C., Ferretti, A., Novali, F., Prati, C., Rocca, F.: SAR monitoring of progressive and seasonal ground deformation using the permanent scatterers technique. IEEE Trans. Geosci. Remote Sens. **41**, 1685–1701 (2003)
8. Hilley, G.E., Roland, B., Alessandro, F., Fabrizio, N., Fabio, R.: Dynamics of slow-moving landslides from permanent scatterer analysis. Science **304**, 1952–1955 (2004)
9. Shi, X., Zhang, L., Liao, M., Balz, T.: Deformation monitoring of slow-moving landslide with L- and C-band SAR interferometry. Remote Sens. Lett. **5**, 951–960 (2014)
10. Tantianuparp, P., Shi, X., Zhang, L., Balz, T., Liao, M.: Characterization of landslide deformations in Three Gorges area using multiple InSAR data stacks. Remote Sens. **5**, 2704–2719 (2013)
11. Cascini, L., Fornaro, G., Peduto, D.: Advanced low- and full-resolution DInSAR map generation for slow-moving landslide analysis at different scales. Eng. Geol. **112**, 29–42 (2010)

12. Handwerger, A.L., Roering, J.J., Schmidt, D.A.: Controls on the seasonal deformation of slow-moving landslides. Earth Planet. Sci. Lett. **377–378**, 239–247 (2013)
13. Herrera, G., Gutiérrez, F., García-Davalillo, J.C., Guerrero, J., Notti, D., Galve, J.P., et al.: Multi-sensor advanced DInSAR monitoring of very slow landslides: the Tena Valley case study (Central Spanish Pyrenees). Remote Sens. Environ. **128**, 31–43 (2013)

Modeling the Avian Influenza H5N1 Virus Infection in Human and Analyzing Its Evolution

Ping Zhang(✉)

Geo-Exploration Science and Technology College,
Jilin University, Changchun 130061, China
zp@jlu.edu.cn

Abstract. Here, after identifying the HPAI H5N1 gene data corresponding to the sites of HPAI H5N1 outbreaks in poultry and wild birds, the outbreak locations were set as the sources of human infection and a patch-based SEIR Cellular Automata (CA) epidemic model was run to simulate human responses to the HPAI H5N1 virus. HPAI H5N1 viruses from poultry and wild birds that were capable of infecting humans were identified and, through reconstruction of the phylogenetic trees with estimation of the evolutionary distances, the evolution of the HPAI H5N1 virus capable of infecting humans transmitted through poultry and wild birds was analyzed. HPAI H5N1 transmission between poultry and humans in China was modeled in different human population density scenarios from 2004–2009. The results showed that different human density distributions had little effect on the number of human cases of HPAI H5N1 and that poultry was the main source of infection.

Keywords: Patch-based SEIR CA epidemic model · Phylogenetic tree · Human density distribution · H5N1

1 Introduction

The first human case of highly pathogenic avian influenza (HPAI) subtype H5N1 infection was observed in Hong Kong 1997 [1]. Since then, the cumulative number of confirmed human cases of HPAI subtype H5N1 infection reported to the World Health Organization as of 2014 has reached 667 worldwide [2]. Overall, the 47 HPAI subtype H5N1 human cases in China have shown 60% mortality. More than 80% of the total HPAI subtype H5N1 human cases have been reported in avian influenza endemic areas, indicating hotspots for bird-to-human transmission [3]. Further, the hemagglutinin (HA) genes of 15 human isolates from southern China belong to sub-clade 2.3.4. These isolates are closely related to each other and to HPAI subtype H5N1 poultry viruses isolated from the same geographic location at the same time, suggesting that infected domestic poultry are the source of human infection [4, 5]. However, the HA gene of one human isolate from the Xinjiang Autonomous Region in northern China was found to be more closely related to viruses found in Qinghai-Lake-like migratory birds [6]. In addition, HPAI subtype H5N1 has been found to cause family case clusters

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 339–352, 2017.
DOI: 10.1007/978-981-10-3969-0_38

that may have involved human-to-human transmission [7]. For this reason, many experts expect a pandemic due to a mutant avian influenza virus capable of fast transmission among humans [8]. Therefore it is important and urgent to study the infection of HPAI H5N1 virus in humans to refine the understanding of HPAI subtype H5N1 virus evolution for better control and prevention of HPAI subtype H5N1 virus transmission.

Evolution is the reason why HPAI subtype H5N1 viruses can infect humans. Since 1997, the avian influenza A/Goose/Guangdong/1/96 (Gs/GD virus) lineage has undergone frequent gene rearrangement with different avian influenza viruses that were circulating in the region, and this has generated many different virus strains. Their antigenic shifts (or phenotypes) have been recorded through influenza virus surveillance in China [9, 10]. The evolution of HPAI subtype H5N1 viruses in recent years has been associated with increasing virulence, improving ability to cross the species barrier, and expanding host range, which, besides poultry and wild birds, has also including a broad range of mammalian species, including humans [11–15]. As such, HPAI subtype H5N1 virus evolution of its genetic and antigenic properties has allowed a diversity of strains to emerge in endemic areas with altered receptor specificity, including a new H5 sublineage that has shown enhanced binding affinity to the human-type receptor [3]. The persistent introduction of HPAI subtype H5N1 virus to humans raises the possibility of HPAI becoming a human pandemic virus, either as a purely avian virus adapting to more efficient human transmission, or through reassortment with existing human influenza strains [16, 17].

Progress in mathematical analysis and modeling is of fundamental importance to collective understanding of viral infection because mathematical models can help determine and quantify critical parameters, then thresholds in the relationships of those parameters, even if the relationships are nonlinear and not consistent with simple reasoning [18]. In recent years, a few studies have been performed on the avian-human influenza epidemic model [19–22] but research on this topic is still at the elementary stage. Most of these models were based on the standard SIR (Kermack-McKendrik) epidemic model and were constructed using both birds and humans. However, the lack of direct interaction between wild birds and humans creates an epidemiological bottleneck. Infectious strains must first pass through domestic birds before they can reach humans [23]. None of the previous attempts to model avian influenza virus transmission from domestic birds (poultry) to humans has revealed the mechanism by which the avian influenza virus might move from wild birds to humans.

Improvements in bioinformatic methods and epidemiological analysis, together with profound expansion of the GenBank database of avian influenza viral genome sequences, have provided an unprecedented opportunity to investigate longstanding questions in avian influenza evolution and epidemiology [18, 24–27]. However, even though the evolution of influenza viruses has been modeled in human and avian populations, these models are not practical for use in China because no avian influenza gene data derived from human sampling are available for 2004–2009 [28–33]. In the present analysis we analyzed the evolution of the avian influenza gene from wild birds and poultry using simulated infection in humans. This analysis has identified avian influenza strains in wild birds and poultry capable of infecting humans, and their evolution was analyzed using bioinformatics. Further, the present analysis simulated

HPAI subtype H5N1 virus transmission among wild birds, poultry, and humans under different human density distributions to evaluate the effects of human population density distribution on the number of human cases of HPAI subtype H5N1. These data shed new light on the evolution of the HPAI subtype H5N1 virus in wild birds and poultry that could infect humans.

2 Methods

2.1 Data Source

Cases of HPAI subtype H5N1 virus in wild birds and poultry in China reported from January 2004 to March 2011 were provided by the World Organization for Animal Health [34]. Human cases of HPAI subtype H5N1 infection in China with the associated HA sequence data reported from January 2003 to January 2012 were provided by the World Health Organization [35]. Basic geographic data were provided by Data Sharing Infrastructure of Earth System Science [36]. Data regarding the population distribution of China were extracted from the LandScan Global Population Database at 30 arc seconds and provided by the LandScan Global Population Projectin ESRI Grid format [37]. National Population Sample Survey Data for the year 2004–2010 were provided by the China Statistics Yearbook [38]. Data regarding HPAI subtype H5N1 hemagglutinin (HA) gene sequences in wild birds and poultry in China were collected between 2004–2009. HA sequence entries with nucleotide length not less than 1600 base pairs (bp) were downloaded from GenBank through the National Center for Biotechnology Information [39].

2.2 Population Density Surface in China

In demographics and ecology, the natural human population growth rate (NHPGR) is the rate at which the number of individuals in a population increases naturally in a given time period as a fraction of the initial population. This usually involves the number of births minus the number of deaths over a year [40]. In this work, the total population for each province in China in 2008 was calculated based on the population distribution in China in 2008 derived from LandScan within each of China's provincial boundaries. Using the natural human population growth rate for each province, the total population of each province from 2004–2007 and in 2009 was computed and then population surface density maps for the country were constructed for these periods using unit cell sizes of 10 km $\times$ 10 km (Fig. 1).

2.3 HPAI H5N1 HA Gene Sequence Data Allocation

Data regarding the documented cases of HPAI subtype H5N1 virus in wild birds and poultry in China were collected at the county level with exact geographical location information, but the HPAI subtype H5N1 hemagglutinin (HA) gene sequence data were collected at the province level without exact geographical location information. Five

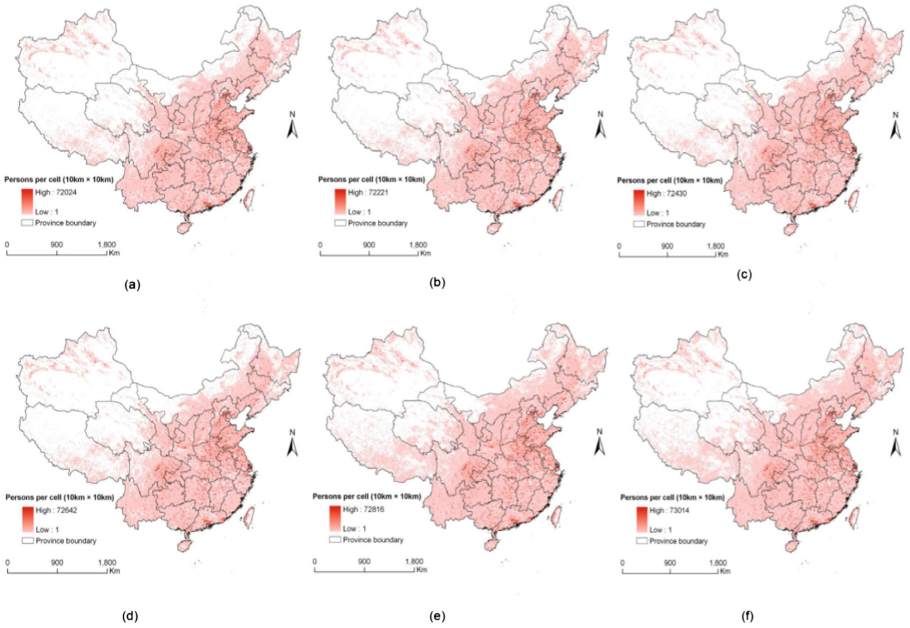


Fig. 1. Population density surfaces in China 2004–2009. (a) In the year of 2004; (b) In the year of 2005; (c) In the year of 2006; (d) In the year of 2007; (e) In the year of 2008; (f) In the year of 2009.

basic criteria were used here to label and group the HPAI subtype H5N1 hemagglutinin (HA) gene sequence data from each HPAI subtype H5N1 associated with a specific outbreak site: location, species, time, bird migration flyway, and the nearest distance. This means that HPAI subtype H5N1 hemagglutinin (HA) gene sequence data were associated with particular HPAI subtype H5N1 outbreak sites depending on whether they were obtained from the same species at the same location (outbreak or isolation) and at the same time (outbreak or isolation). Here, the criterion "species" refers to either of two hosts of HPAI subtype H5N1: wild birds or poultry. If a given HPAI subtype H5N1 outbreak in a particular province did not have any corresponding gene data, then gene data associated with outbreaks involving the same species in the same time period from the nearest province would be attributed to that province. In addition, gene data was attributed to specific HPAI subtype H5N1 outbreaks in wild birds according to whether they were along the same bird migration flyways.

2.4 Patch-Based SEIR CA Epidemic Model

In this work, a patch-based SEIR model was combined with a cellular automata (CA) model. For this patch-based SEIR model, S_i, E_i, I_i, R_i denoted the number of susceptible, exposed, infectious, and recovered individuals in patch i, respectively, for $i = 1 \ldots n$. The total population of patch i was $N_i = S_i + E_i + I_i + R_i$. The birth and natural death rate constant d was here assumed to be the same in each patch, so that the

total population of each patch remained constant [41]. The average latent period $1/\varepsilon$ and the average infectious period $1/\gamma$ were here assumed to be the same in each patch [41]. This patch-based SEIR model can be written as follows:

$$\begin{aligned} S_i^{'} &= dN_i - dS_i - \beta S_i \\ E_i^{'} &= \beta S_i - (d+\varepsilon)E_i \\ I_i^{'} &= \varepsilon E_i - (d+\gamma)I_i \\ R_i^{'} &= \gamma I_i - dR_i \end{aligned} \tag{1}$$

where, d is the natural growth rate, β is the contact rate that includes a mass-action transmission in the patch i and contact between any two patches, $1/\varepsilon$ is the average latent period, $1/\gamma$ is the average infectious period, and $S_i^{'}, E_i^{'}, I_i^{'}, R_i^{'}$ are the susceptible, exposed, infectious, and recovered populations of patch i, respectively, after time t.

Here, the time interval was set at one day, and natural growth rate was not considered for any patch, so the Eq. (1) should be written as follows:

$$\begin{aligned} S_i^{'} &= N_i - \beta S_i \\ E_i^{'} &= \beta S_i - \varepsilon E_i \\ I_i^{'} &= \varepsilon E_i - \gamma I_i \\ R_i^{'} &= \gamma I_i \end{aligned} \tag{2}$$

where, $S_i^{'}, E_i^{'}, I_i^{'}, R_i^{'}$, S_i, E_i, I_i, R_i, and $N_i, \beta, \varepsilon, \gamma$ have the same meaning as in Eq. (1).

The probability that a given patch occupied by human hosts will be exposed on day t, it is given by the following equation, which was adapted from an equation in a previous work [28]:

$$\rho_{p,i}(t) = N_{p,i} * [1 - \exp(-(1 + \varepsilon_p \sin(t/365))(\beta_w/N_{p,i})] \tag{3}$$

where, β_W is a contact rate between all human hosts in a patch (p, i), and $N_{p,i}$ is the total population of a patch (p, i). $\rho_{p,i}(t)$ is the probability of a patch (p, i), and seasonal variation in contact rates is characterized by ε_p, which is set to $\varepsilon_p = +0.25$ for $1 < p \le M/2$, and $\varepsilon_p = -0.25$ for $M/2 < p \le M$. M is the total number of rows in the human population surface density data in China for the period 2004–2009.

In this work, the patch-based SEIR model has been combined with CA model to represent the avian influenza transmission spatially. The CA epidemic model used here was adapted from one used in two previous works [43, 44]. The state $C_{p,i}^t$ of the patch (p, i) at time t is given as follows:

$$C_{p,i}^t = \left\{ P_{p,i}^t, ENF_{p,i}^t, INF_{p,i}^t, IMF_{p,i}^t \right\} \tag{4}$$

where, $ENF_{p,i}^t$ is an exposed flag, $INF_{p,i}^t$ is an infectious flag and $IMF_{p,i}^t$ is an immune flag. The value of $ENF_{p,i}^t$ indicates whether some or all of the individuals located in the patch (p, i) had been exposed to the avian influenza strain at time t. If $ENF_{p,i}^t = 1$, then the patch (p, i) is one of the given set of exposed patches, remaining exposed for latent period t_{ep}. After the latent period, the exposed patches could become infectious if the

value of $INF^t_{p,i} = 1$. They would remain infected for an infectious period t_{in} and then become immune. If $ENF^t_{p,i} = 0$, then none of the individuals located in the patch (p, i) had been exposed to the avian influenza strain. $IMF^t_{p,i}$ indicates whether the human population located in the patch (p, i) is immune to the avian influenza or not. The immune human population loses its immunity after time t_{im}, at which point it becomes susceptible to the avian influenza again. Accordingly, the value of $IMF^t_{p,i}$ changes from 1 to 0. $P^t_{p,i}$ represents the proportion of the individuals in the patch (p, i) infected with the avian influenza virus when $INF^t_{p,i} = 1$ at time t:

$$P^t_{p,i} = \frac{I^t_{p,i}}{E^t_{p,i}} \tag{5}$$

where, $I^t_{p,i}$ is the infected portion of the population in the patch (p, i) and $E^t_{p,i}$ is the total exposed population in the patch (p, i). $I^t_{p,i} \leq E^t_{p,i}$. In this way, $P^t_{p,i}$ may take any value between 0 and 1 but may not exceed 1. When $INF^t_{p,i} = 0, P^t_{p,i} = 0$.

In a slight modification of the CA model proposed by Sirakoulis et al., the transition from an exposed state to an infected state is achieved as follows [42]:

$$P^{t+1}_{p,i} = P^t_{p,i} + E^t_{p,i}(k(P^t_{p-1,i}, P^t_{p,i-1}, P^t_{p,i+1}, P^t_{p+1,i}) + l(P^t_{p-1,i-1}, P^t_{p-1,i+1}, P^t_{p+1,i-1}, P^t_{p+1,i+1})) \tag{6}$$

where, $E^t_{p,i}$ is the proportion of the population that is exposed. The updated state $P^{t+1}_{p,i}$ of the central patch (p, i) is a weighted function of the present state of the patch and that of its neighbors. In the present analysis, the adjacent nearest neighbors of the patch (p, i), i.e., the neighbors that had a common side with the patch (p, i) and the diagonal adjacent neighbors were grouped. The effect of the adjacent nearest neighbors was multiplied by k, and the effect of the diagonal adjacent neighbors by l. It was here assumed that the patch (p, i) would be infected more quickly by an infected adjacent nearest neighbor than by an infected diagonal adjacent neighbor because of the more extensive contact within the population in the patch (p, i) and the adjacent nearest neighbor patch. In this way, k is always greater than l. The sites of epidemic sources were chosen randomly, taking into account the poultry and human populations in each patch.

2.5 Phylogenetic Tree and Evolutionary Distance

The true phylogenetic tree is often not known even after phylogenetic analysis, and it is difficult to determine whether phylogenetic trees with different forms of reconstruction are correct or not [44]. Therefore, it is better to reconstruct the phylogenetic trees using different methods and to compare their differences. Here, three methods were used to reconstruct phylogenetic trees: distance methods, parsimony methods, and likelihood methods [44]. Distance methods are also called distance-matrix methods. Pairwise distances of sequences for taxa were computed if the data did not exist in distance form

[45]. Parsimony methods involve the maximum parsimony (MP) method of phylogenetic tree reconstruction using nucleotide sequence data for a site-by-site analysis [46]. Of the likelihood methods we used the maximum likelihood (ML) method where the ML value is computed for as many topologies as possible and the topology that shows the highest ML value is chosen as the final tree [47].

Before reconstructing the phylogenetic trees, multiple sequence alignments were performed to provide a quantitative measure for sequence similarity. A multiple sequence alignment is a 2D table, in which the rows represent individual sequences, and the columns the residue positions [48]. Here the Clustal program was used. It aligns sequences in pairs, following the branching order of a family tree [48]. Similar sequences are aligned first, and more distantly related sequences are added later [48]. Once pairwise alignment scores for each sequence relative to all others are calculated, they are used to cluster the sequences into groups, which are then aligned against each other to generate the final multiple alignment [48].

Evolutionary distance, also called general genetic distance, has been used to measure genetic divergence estimating the divergence time between two taxa, i.e., the time that has passed since those populations existed as a single population [40, 49]. Smaller evolutionary distances indicate a close genetic relationship and larger evolutionary distances indicate a more distant genetic relationship [40]. In the case of amino acid sequence data, as in this case, the evolutionary distance was measured using the number of amino acid substitutions [49].

3 Results

3.1 Population Surface Density Changes in China from 2004–2009

The highest number of individuals per cell in China increased from 72,024 individuals in 2004 to 73,014 in 2009 (Fig. 1). This showed that, not only China's total population, but also its population density were increasing steadily. Areas that already had high population density did not change much during this period. The areas that increased in population were distributed mainly in the middle, southern, and northeastern parts of China. Population distribution was more toward the north and west in 2009 than in 2004. Some areas that had been completely devoid of people in 2004 were populated in 2009 (Fig. 1). In this way, population distribution in China became more extensive from 2004–2009, and areas unoccupied by humans became less numerous.

3.2 Avian Influenza H5N1 Virus Infection in Humans in China from 2004–2009

In the patch-based SEIR CA epidemic model, the avian influenza H5N1 human infection source (seed) was chosen stochastically from outbreak sites associated with HPAI H5N1 HA gene sequence data. Three sources sites were chosen in this way. The contact rate βw in Eq. (3) was set as 1.5 based on the WHO report [50]. The spatial extent of infection was determined using the limited duration of avian influenza (infectious period of 4 days in the present model). The model was repeated 100 times and

the average number of H5N1 human cases over the 100 simulations were plotted over time for any 6 sequential days during each year from 2004–2009.

The average number of avian influenza H5N1 human cases projected to occur during any six days following infection in the years 2004–2009 showed a similar progression over time for each years analysis (Fig. 2) where the case numbers increased rapidly from the first day, peaked on the second day, dropped slightly on day 3, then plateaued. The only difference among these plots was the maximum number of cases on the second day, where both 2006 and 2009 were slightly higher than in other years. These data suggested that differences in human population density and population distribution have a small effect on avian influenza H5N1 infection in humans.

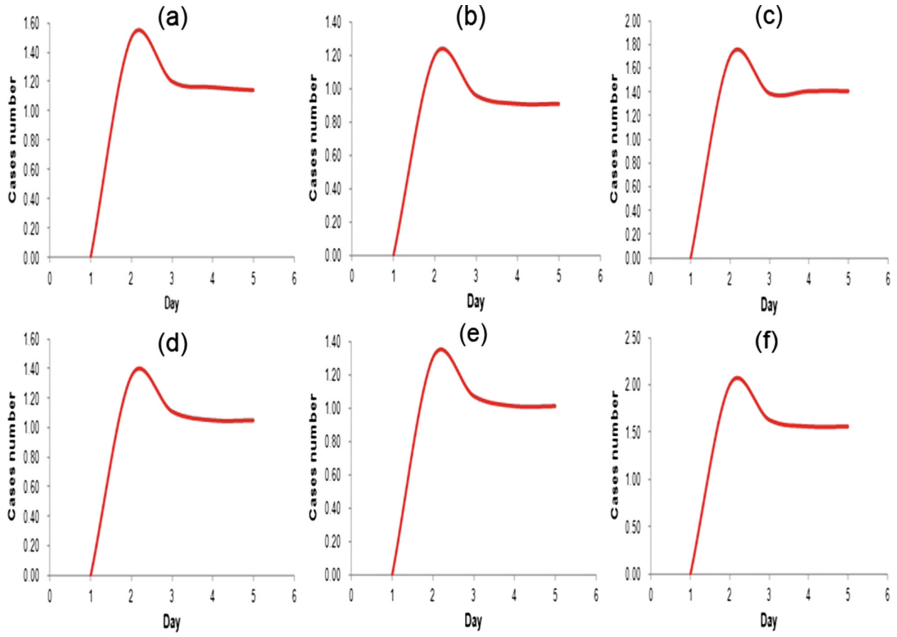


Fig. 2. Average human cases number through any six days in China 2004–2009 obtained using Eqs. (2)–(6). The lines in the figs. (a)–(f) representing the years 2004–2009 denote how the average human cases number of infecting avian influenza H5N1 changes from the first day when the infection starts.

3.3 Evolutionary Analysis of Avian Influenza H5N1 Virus in China 2004–2009

For the evolutionary relationship of HPAI H5N1 virus strains that could infect humans in China between 2004–2009, phylogenetic trees that included the first strain of A/goose/Guangdong/1/1996, were reconstructed with the distance-matrix method, the maximum likelihood (ML) method, and the maximum parsimony (MP) method. Since phylogenetic trees created by the ML and MP methods were similar to that created by

the distance-matrix method, only the result created by the distance-matrix method showed (Fig. 3). From the perspective of geographic distribution, most sublineages in this phylogenetic tree comprise sequences from geographically adjacent areas (Fig. 3).

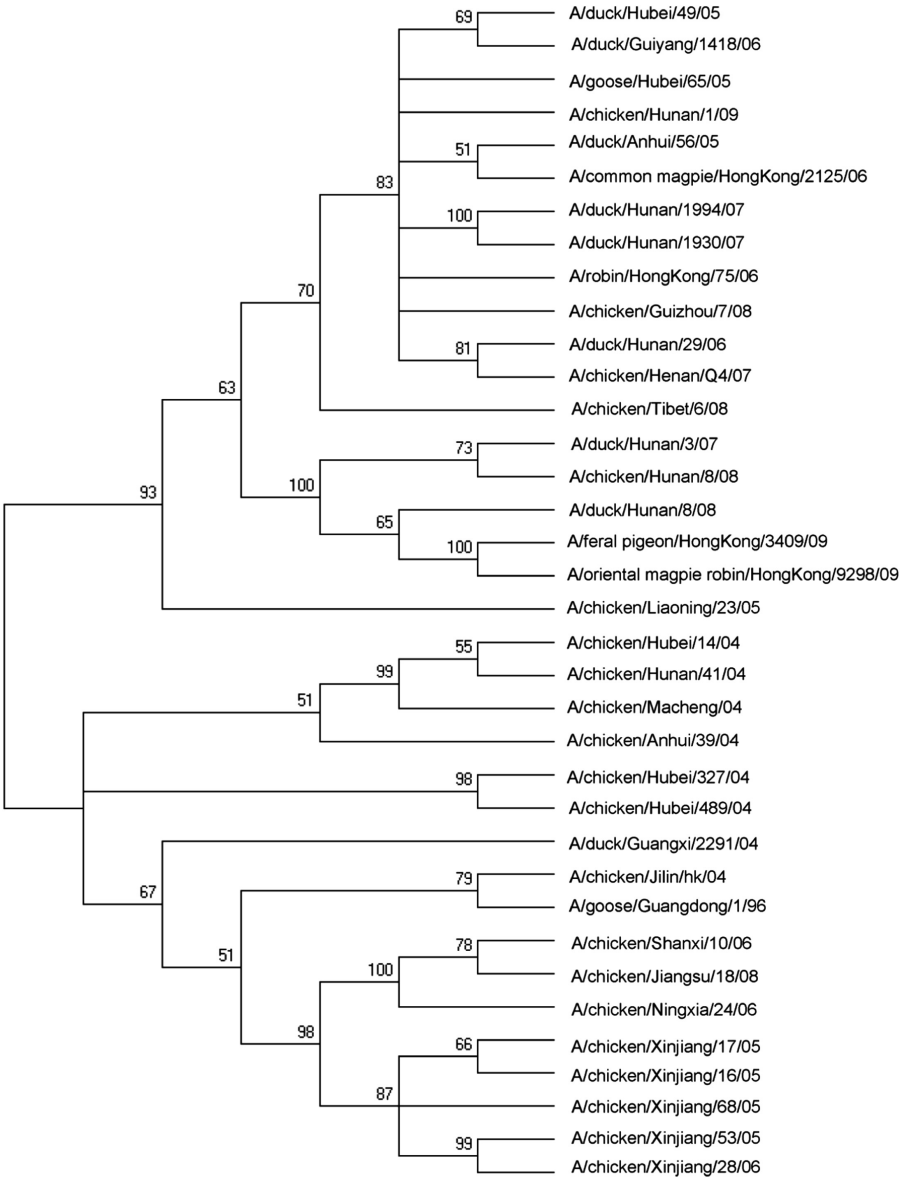


Fig. 3. Phylogenetic relationships of avian influenza H5N1 strains that could infect humans in China. Phylogenetic tree based on HA genes of virus from poultry and wild birds isolated in China were reconstructed with the maximum likelihood (ML) method.

According to the branch positions, HPAI H5N1 virus strains from the middle of China were similar to those in the south of China and different from those in northwestern China (Fig. 3). This accounts for the close evolutionary relationship of HPAI H5N1 virus strains from the middle of China and those from the south. The three HPAI H5N1 virus strains similar to the strain of A/goose/Guangdong/1/1996 were from three very different geographic areas: northwestern, northeastern, and middle China (Fig. 3). This strengthens the position that the sublineage of A/goose/Guangdong/1/1996 continued to be prevalent in China from 2004–2009.

Evolutionary distances were measured using the average evolutionary distance between the A/goose/Guangdong/1/1996 and the other strains capable of infecting humans in each year from 2004–2009 based on nucleotide substitutions and amino acid substitutions, respectively. Evolutionary distance based on nucleotide substitutions in Fig. 4(a) and amino acid substitutions in Fig. 4(b) for the HPAI H5N1 virus strains capable of infecting humans in China 2004–2009 increased steadily overall. However, in 2006, evolutionary distance based on aminoacid substitutions shrank suddenly in Fig. 4(b).

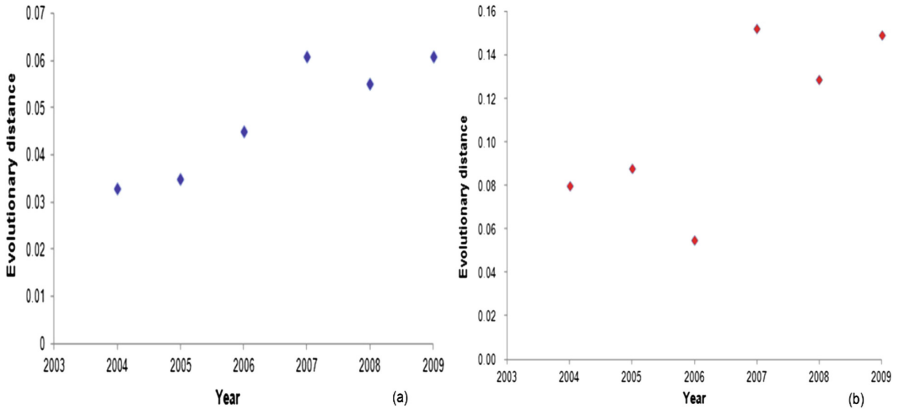


Fig. 4. Evolutionary distance change for the HPAI H5N1 virus strains infecting humans in China 2004–2009. (a) based on nucleotide substitutions; (b) based on amino acid substitutions.

4 Discussion

The central reason for avian influenza virus infection of humans is the mutation and evolution of the avian influenza virus. Here, a model of circulating avian influenza strains for forthcoming seasons is proposed for the prediction of avian-human epidemics. The circulating avian influenza strain in the next season is likely to be different from current strain because of the fast mutation rate of the avian influenza virus [19]. This means that newly susceptible S comes primarily from the recovered class R because of only partial immunity to the mutant strains, instead of coming from births [19]. For this reason, it is better to change the current SEIR-type model to an

SEIRS-type model as some researchers have described [32]. As such, it may be a good idea to combine both models.

A problem that cannot be avoided in this analysis is that it only considers poultry (domestic birds) distribution in the SEIR CA epidemic, not distribution through wild birds because the data for wild birds are unavailable. However, wild birds can carry the avian influenza virus without displaying any symptoms [19]. Wild birds do not normally come into direct contact with humans, but they are known to come into contact with poultry. In this way, wild birds can infect humans through poultry. For this reason, it is necessary to consider the effects of wild birds in models of this kind. It might be a best to include a constant input to the model accounting for the contact between wild birds and poultry [19]. To support this proposed modification to the model, most of the HPAI H5N1 outbreaks in the analysis came from poultry that carried the HPAI H5N1 HA gene suggesting that poultry were the main source of infection probably because poultry come into direct contact with humans far more often than wild birds.

The CA epidemic model is well defined in the mathematical sense, but it has serious shortcomings. In particular, it provides only a limited representation of infectious disease transmission and there is no allowance for action-at-a-distance in a formal CA [43, 51]. For this reason, the CA epidemic model was chosen here due to data limitations. In the absence of such limitations, network-based epidemic models may be superior. The network-based epidemic model has two advantages over the CA epidemic model. It is more flexible and can take bird migration and contact between wild birds and poultry into account. Some researchers have examined the global spatio-temporal distribution of avian influenza cases in both wild birds and poultry and have found that the network of outbreaks, and the links between them, form a scale-free network [52]. This suggests that a more desirable representation of avian influenza virus transmission could be provided if the network-based epidemic model was used.

Comprehensive understanding of the evolution of the avian influenza virus requires a far broader analysis of whole genome sequences from a wider range of subtypes, host species, and geographical areas, including tropical regions. It also requires the development of more realistic epidemiological models [27]. Here only one set of HA gene sequence data was used in association with one specific site of an avian influenza outbreak. Even though some similarity exists for all the HA gene sequence data isolated from the same species at the same place, there are still some differences among them. The conclusions drawn here regarding the evolutionary relationship of the HPAI H5N1 virus from wild birds and poultry that could infect humans lacks experimental validation of the model used to draw those conclusions. Therefore, the conclusions drawn here regarding the evolutionary relationships among the HPAI H5N1 viruses from wild birds and poultry that are capable of infecting humans are preliminary rather than comprehensive. However, if the current epidemic model is viewed at the sequence level, it can be used to predict which HA gene sequences are likely to infect humans and which are not. Interestingly, in 2006, evolutionary distance based on aminoacid substitutions shrank suddenly and, accordingly, the total number of human cases of HPAI H5N1 peaked in 2006, suggesting that it may have been the small evolutionary distance in 2006 that caused the high human infection rate.

5 Conclusion

This paper describes two central accomplishments. The first was the use of a patch-based SEIR CA epidemic model to simulate the HPAI H5N1 transmission among poultry and humans and to explore the effects of human population density distribution on the number of human cases of HPAI H5N1. This facilitated a quantitative and spatially explicit combined analysis. Because most infectious disease transmission phenomena are highly non-linear and dynamic, the patch-based SEIR CA epidemic model is a useful way of studying HPA1 H5N1 transmission. In this analysis, China, from 2004–2009, served as the study area, and results showed that human population density distribution had a small effect on the number of human cases of HPAI H5N1.

The other goal was to analyze the evolution of HPAI H5N1 virus in poultry and wild birds capable of infecting humans. Results of running the patch-based SEIR CA model showed that most of the HPAI H5N1 viruses capable of infecting humans derived from poultry. Poultry was the main source of human infection with avian influenza H5N1. The evolutionary relationships among HPAI H5N1 viruses capable of infecting humans showed geographic adjacent distribution. Evolutionary distance based on nucleotide substitutions and amino acid substitutions for the HPAI H5N1 virus strains infecting humans in China from 2004–2009 went up steadily. This indicated that the HPAI H5N1 virus strains capable of infecting humans have also been evolving steadily.

References

1. Ku, A., Chan, L.: The first case of H5N1 avian influenza infection in a human with complications of adult respiratory distress syndrome and Reye's syndrome. J. Paediatr. Child Health **35**, 207–209 (1999)
2. World Health Organization. http://www.who.int/
3. Watanabe, Y., Ibrahim, M.S., Suzuki, Y., Ikuta, K.: The changing nature of avian influenza A virus (H5N1). Trends Microbiol. **20**(1), 11–20 (2012)
4. Shu, Y., Yu, H., Li, D.: Lethal avian influenza A (H5N1) infection in a pregnant woman in Anhui Province, China. N. Engl. J. Med. **354**, 1421–1422 (2006)
5. Wang, H., Feng, Z., Shu, Y., Yu, H., Zhou, L., Zu, R., Huai, Y., Dong, J., Bao, C., Wen, L., Wang, H., Yang, P., Zhao, W., Dong, L., Zhou, M., Liao, Q., Yang, H., Wang, M., Lu, X., Shi, Z., Wang, W., Gu, L., Zhu, F., Li, Q., Yin, W., Yang, W., Li, D., Uyeki, T.M., Wang, Y.: Probable limited person-to-person transmission of highly pathogenic avian influenza A (H5N1) virus in China. Lancet **371**, 1427–1434 (2008)
6. Neumann, G., Chen, H., Gao, G.F., Shu, Y., Kawaoka, Y.: H5N1 influenza viruses: outbreaks and biological properties. Cell Res. **20**, 51–61 (2010)
7. Yang, Y., Halloran, M.E., Sugimoto, J.D., Longini Jr., I.M.: Detecting human-to-human transmission of avian influenza A (H5N1). Emerg. Infect. Dis. **13**(9), 1348–1353 (2007)
8. Iwami, S., Takeuchi, Y., Liu, X.: Avian flu pandemic: can we prevent it? J. Theor. Biol. **257**, 181–190 (2009)
9. Horimoto, T., Kawaoka, Y.: Influenza: lessons from past pandemics, warnings from current incidents. Nat. Rev. Microbiol. **3**, 591–600 (2005)

10. Liu, J., Xiao, H., Lei, F., Zhu, Q., Qin, K., Zhang, X.W., Zhang, X.L., Zhao, D., Wang, G., Feng, Y., Ma, J., Liu, W., Wang, J., Gao, G.F.: Highly pathogenic H5N1 influenza virus infection in migratory birds. Science **309**, 1206 (2005)
11. Keawcharoen, J., Oraveerakul, K., Kuiken, T., Fouchier, R.A.M., Amonsin, A., Payungporn, S., Noppornpanth, S.: Avian influenza H5N1 in tigers and leopards. Emerg. Infect. Dis. **10**, 2189–2191 (2004)
12. de Jong, M.D., Hien, T.T.: Avian influenza A (H5N1). J. Clin. Virol. **35**, 2–13 (2006)
13. Roberton, S.I., Bell, D.J., Smith, G.J.D., Nicholls, J.M., Chan, K.H., Nguyen, D.T., Tran, P.Q., Streicher, U., Poon, L.L.M., Chen, H., Horby, P., Guardo, M., Guan, Y., Peiris, J.S.M.: Avian influenza H5N1 in viverrids: implications for wildlife health and conservation. Proc. Biol. Sci. **273**, 1729–1732 (2006)
14. Abdel-Moneim, A.S., Abdel-Ghany, A.E., Shany, S.A.S.: Isolation and characterization of highly pathogenic avian influenza virus subtype H5N1 from donkey. J. Biomed. Sci. **17**, 25 (2000)
15. Watanabe, Y., Ibrahim, M.S., Ellakany, H.F., Kawashita, N., Mizuike, R., Hiramatsu, H., Sriwilaijaroen, N., Takagi, T., Suzuki, Y., Ikuta, K.: Acquisition of human-type receptor binding specificity by new H5N1 influenza virus sublineages during their emergence in birds in Egypt. PLoS Pathog. **7**, e1002068 (2011)
16. Webster, R.G., Bean, W.J., Gorman, O.T., Chambers, T.M., Kawaoka, Y.: Evolution and ecology of influenza A viruses. Microbiol. Rev. **56**, 152–179 (1992)
17. Taubenberger, J.K., Reid, A.H., Lourens, R.M., Wang, R., Jin, G., Fanning, T.G.: Characterization of the 1918 influenza virus polymerase genes. Nature **437**, 889–893 (2005)
18. Smith, D.J.: Predictability and preparedness in influenza control. Science **312**, 392–394 (2006)
19. Iwami, S., Takeuchi, Y., Liu, X.: Avian-human influenza epidemic model. Math. Biosci. **207**, 1–25 (2007)
20. Kim, K.I., Lin, Z., Zhang, L.: Avian-human influenza epidemic model with diffusion. Nonlinear Anal.: Real World Appl. **11**, 313–322 (2010)
21. Agarwal, M., Verma, V.: An avian-human influenza epidemic model with vaccination. J. Appl. Sci. **5**(6), 451–458 (2010)
22. Samanta, G.P.: Permanence and extinction for anonautonomous avian-human influenza epidemic model with distributed time delay. Math. Comput. Model. **52**, 1794–1811 (2010)
23. Lucchetti, J., Roy, M., Martcheva, M.: An avian influenza model and its fit to human avian influenza cases. In: Tchuenche, J.M., Mukandavire, Z. (eds.) Advances in Disease Epidemiology, pp. 1–30. Nova Science Publishers, New York (2009)
24. Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A.: Relaxed phylogenetics and dating with confidence. PLoS Biol. **4**(5), e88 (2006)
25. Viboud, C., Bjørnstad, O.N., Smith, D.L., Simonsen, L., Miller, M.A., Grenfell, B.T.: Synchrony, waves, and spatial hierarchies in the spread of influenza. Science **312**, 447–451 (2006)
26. Ghedin, E., Sengamalay, N.A., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., Spiro, D.J., Sitz, J., Koo, H., Bolotov, P., Dernovoy, D., Tatusova, T., Bao, Y., George, K.S., Taylor, J., Lipman, D.J., Fraser, C.M., Taubenberger, J.K., Salzberg, S.L.: Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. Nature **437**, 1162–1166 (2005)
27. Nelson, M.I., Holmes, E.C.: The evolution of epidemic influenza. Nat. Rev. **8**, 196–205 (2007)
28. Ferguson, N.M., Galvani, A.P., Bush, R.M.: Ecological and immunological determinants of influenza evolution. Nature **422**, 428–433 (2003)

29. Koelle, K., Cobey, S., Grenfell, B., Pascual, M.: Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. Science **314**, 1898–1903 (2006)
30. Koelle, K., Khatri, P., Kamradt, M., Kepler, T.B.: A two-tiered model for simulating the ecological and evolutionary dynamics of rapidly evolving viruses, with an application to influenza. J. R. Soc. Interface **7**, 1257–1274 (2010)
31. Roche, B., Drake, J.M., Rohani, P.: An agent-based model to study the epidemiological and evolutionary dynamics of influenza viruses. BMC Bioinform. **12**, 87 (2011)
32. Martcheva, M.: An evolutionary model of influenza A with drift and shift. J. Biol. Dyn. **6**(2), 299–332 (2011)
33. Ito, K., Igarashi, M., Miyazaki, Y., Murakami, T., Iida, S., Kida, H., Takada, A.: Gnarled-trunk evolutionary model of influenza A virus hemagglutinin. PLoS ONE **6**(10), e25953 (2011)
34. World Organization for Animal Health. http://www.oie.int/
35. World Health Organization. http://www.who.int/influenza/human_animal_interface/avian_influenza/en/
36. Data Sharing Infrastructure of Earth System Science. http://www.geodata.cn/
37. Landscan Global Population Project. http://www.ornl.gov/sci/landscan/
38. China Statistics Yearbook. http://www.tjcn.org/
39. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/
40. Wikipedia. http://zh.wikipedia.org/wiki
41. van den Driessche, P.: Spatial structure: patch models. In: Brauer, F., van den Driessche, P., Wu, J. (eds.) Mathematical Epidemiology, pp. 179–189. Springer, Berlin (2008). (Chapt. 7)
42. Sirakoulis, GCh., Karafyllidis, I., Thanailakis, A.: A cellular automaton model for the effects of population movement and vaccination on epidemic propagation. Ecol. Model. **133**, 209–223 (2000)
43. Zhang, P., Atkinson, P.M.: Modelling the effect of urbanization on the transmission of an infectious disease. Math. Biosci. **211**, 166–185 (2008)
44. Nei, M., Kumar, S.: Phylogenetic trees. In: Nei, M., Kumar, S. (eds.) Molecular evolution and phylogenetics, pp. 73–86. Oxford University Press, New York (2000). (Chap. 5)
45. Li, S., Pearl, D.K., Doss, H.: Phylogenetic tree construction using Markov Chain Monte Carlo. J. Am. Stat. Assoc. **95**(450), 493–508 (2000)
46. Yang, Z.: Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. **42**, 294–307 (1996)
47. Saitou, N.: Property and efficiency of the maximum likelihood method for molecular phylogeny. J. Mol. Evol. **27**, 261–273 (1988)
48. Attwood, T.K., Parry-Smith, D.J.: Multiple sequence alignment. In: Introduction to Bioinformatics. Addison Wesley Longman Limited, London (1999). (Chap. 7)
49. Nei, M., Kumar, S.: Evolutionary change of amino acid sequences. In: Nei, M., Kumar, S. (eds.) Molecular evolution and phylogenetics, pp. 17–32. Oxford University Press, New York (2000). (Chapt. 2)
50. WHO report. http://www.who.int/influenza/human_animal_interface/H5N1_cumulative_table_archives/en/
51. O'Sullivan, D., Torrens, P.M.: Cellular models of urban systems. CASA Paper 22 (2000)
52. Small, M., Walker, D.M., Tse, C.K.: Scale-free distribution of avian influenza outbreaks. Phys. Rev. Lett. **99**, 188702 (2007)

The Research of 3D Geological Modeling in the Main Mining Area and East Mining Area of BayanObo Deposit

Mingchao Zhang[1,2(✉)], Jingchao Li[1], Yike Li[3], Qunchao Zuo[1], Lei Yao[1], Hui Chen[1], and Wanjuan Liang[1]

[1] Development Research Center of China Geological Survey, Beijing 100037, China
cgszhangmc@163.com
[2] Faculty of Earth Sciences and Resources, China University of Geosciences, Beijing 100083, China
[3] Research Center for Strategy of Global Mineral Resource, Chinese Academy of Geological Sciences, Beijing 100037, China

Abstract. BayanObo deposit in Inner Mongolia, an ultra-large REE-Nb-Fe deposit with various ores and mineralized elements, is currently the world's largest rare earth deposit, and has special mineralization. The deposit has strategic significance in China. Applying the professional software of three-dimension (3D) geological modeling with advanced theories and methods, a 3D visualization model of deposit in the main mining area and east mining area, including 3D engineering model, 3D terrain model, 3D rock mass model and 3D ore bodies model, which clarifies the spatial distribution of REE-Nb-Fe in BayanObo deposit in visualization way, and provides scientific basis for proving deep formation and boundary of ore body, calculating resource reserve, and well developing and protecting rare earth resource.

Keywords: 3D geological modeling · The main mining area · The east mining area · BayanObo deposit · Inner Mongolia

1 Introduction

The rare earth consumption in China lies in the emerging high-new-tech industries at present, which are in the stage of development and have increasing demand on rare earth products with great growth potential in the next few years. Therefore, the consumption of rare earth is strong.

BayanObo deposit in Inner Mongolia, an ultra-large REE-Nb-Fe deposit with various ores and mineralized elements [1], is currently the world's largest rare earth deposit, and its proven reserve of rare earth accounts for over 70% of global rare earth reserve. The deposit shows great significance with special mineralization, and attracts extensive attention for a long period.

Considered with geotectonic view, BayanObo REE-Nb-Fe deposit is located in the north margin of North China Plate, and near Xing'an-Mongolian Orogenic Belt [2]. The outcropping mainly covers early Proterozoic basement complex, shallow metamorphic sedimentary rock in Mesoproterozoic BayanObo Gr, and carbonatite vein and

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 353–362, 2017.
DOI: 10.1007/978-981-10-3969-0_39

late Paleozoic granite intruded into basement complex and BayanObo Gr. BayanObo REE-Nb-Fe deposit is mainly developed in the syncline core, which is composed of H_8 dolomite in Mesoproterozoic BayanObo Gr, and dominated by west ore block, main ore block, and eastern ore block, and several minor ore bodies.

The previous geological exploration focuses on the iron mine, where the ore body was outlined based on the iron grade, thus the rare earth beyond boundary was not further investigated. The scholars are more concerned about exploration of Nb and rare earth ore associated with iron ore, and pay less attention to separate Nb and rare earth ore. They have little knowledge about distribution, scale, and resource quantity of separate ore body, and the rare earth resources are still not clear, which affects development and protection of rare earth.

2 Progress of 3D Modeling

With rapidly developing information technology, 3D modeling, and visualization, the underground 3D visualization modeling is increasingly applied in recent years [3]. For a long period, the field geological achievement is traditionally expressed by two-dimension plane and section, resulting in loss and distortion of spatial information, complicated charting, and difficulty to update information. In this case, 3D visualization modeling arose, and it describes the geologic body and geological environment in 3D space, with computer and visualization in scientific computing, and targeting at the defects of modeling and expressing geological information in traditional method.

3D geological modeling and visualization is one of hotspots in the global geo-scientific research. With proper data structure, 3D geological modeling is to integrate management of spatial information, geological interpretation, spatial analysis and prediction, geo-statistics, analysis of entity content, and graph visualization in 3D circumstance by modern spatial information theory and computer technology, and conduct research on geometry and geological information of geological body [4–8], i.e. internal physical and chemical properties, which are applied in geological analysis and estimation of resource reserve. It is an emerging interdisciplinary technology integrating geological exploration, mathematical geology, geophysics, mine surveying, mining geology, GIS, graphic image, and visualization in scientific computing [9]. 3D geological modeling was firstly presented by Simon W. Houlding from Canada in 1993 [10]. Compared with traditional 2D method of expressing geological data, 3D model is capable of accurately expressing geologic phenomena, which realizes rapid and visual reappearance of spatial distribution and mutual relation among geological units, evacuates the implicit geological information, and facilitates engineering decision, geological analysis, and automatic mapping [10, 11].

3 Geological Characteristics

BayanObo deposit is located in BayanObo area, Baotou city, Inner Mongolia, and distributed in a nearly EW long strip shape, which is 3.5 km long, and 1.9 km wide, and has area around 4.31 km^2.

The exposure formation is dominated by five formation-complexes nearly striking EW in Proterozoic BayanObo Gr, and covers H_2–H_9. The Cenozoic Quaternary is widely distributed in the mining area. The formation mostly strikes EW, and has dip of nearly 90°, with the core of basement rock in broad groove, forming broad-groove anticline. BayanObo deposit is located in the north flank of broad moat anticline. H8 formation in the north of broad groove is carbonate of normal sedimentation, and it is dominated by limestone, and intercalated with dolomite. The outcrop shows clear bedding structures, without obvious metamorphism (shown in Fig. 1). There are multi-periods of tectonic movement, and the main fold and fault in the mining area are consistent with direction of tectonic line, forming overlapped folds and several fault structures. The broad-groove fault and anticline form the main fault and anticline structures. In BayanObo there are developed magmatite of various periods and types, and the most outcropped Hercynian granite is widely distributed in the eastern and southern parts of deposit. The basic dyke is also widely distributed.

From east to west, BayanObo deposit is divided into eastern contact zone ore block, eastern ore block, main ore block, western ore block, and northern ore block. This paper focuses on the main ore block, and eastern ore block.

The main ore block is located in NNW of BayanObo deposit mining area, with straight distance of 2.5 km, and it is 1250 m long, and 250 m wide averagely in EW direction, and 415 m wide in maximum and 1030 m deep in maximum in SN direction. In the mining area, the main ore block is the largest, with several minor ore bodies associated with its northern and southern parts. The main ore block has nearly EW strike, south trend, and dip between 50°–60°. It is lenticular, and both ends pinch out

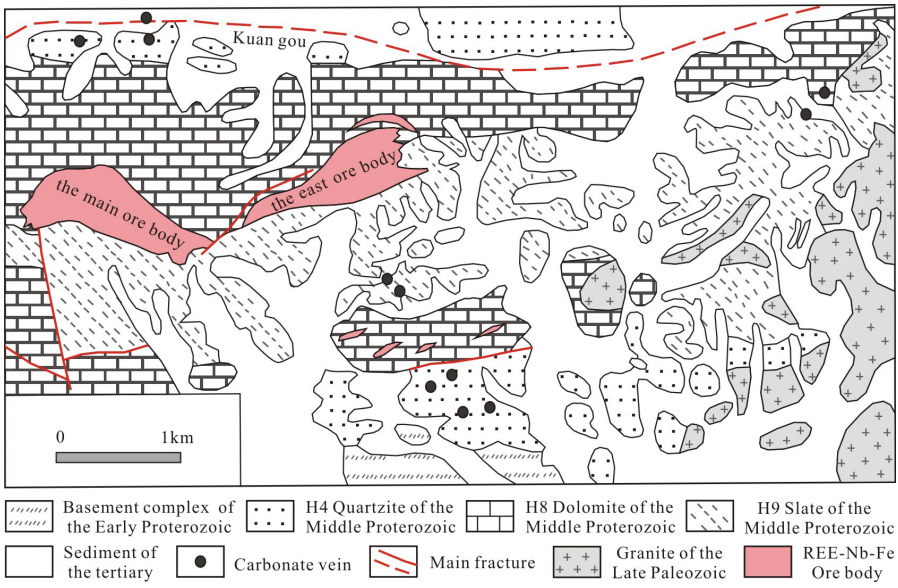


Fig. 1. The geological sketch of BayanObo mining area (modified after Fan et al. [12] and Yang et al. [13]).

into dolomite. The hanging wall is biotite slate, and the foot wall is dolomite. From north to south, there are successively distributed fluorite Nb rare earth iron ore(accounting for 73% of total ore, and rated as medium-low grade iron ore and high grade rare earth ore), massive Nb rare earth iron ore(accounting for 9.4% of total ore, and rated as high grade iron ore), aegirine Nb rare earth iron ore, riebeckite Nb rare earth iron ore, and biotite Nb rare earth iron ore, which account for 8.40%, 5.20% and 3.80% of total iron ore reserve, and are rated as medium-low grade iron ore and medium grade rare earth ore.

The eastern ore block is located in NNE direction of BayanObo deposit mining area, with straight distance of 2.5 km, and it is 1300 m long, and 180 m wide averagely in EW direction, and 350 m wide mostly and 870 m long mostly in SN direction. The eastern ore block is the secondary ore body in the mining area. The eastern ore block has NE strike, south trend, and dip between 50°–65°. It is irregularly lenticular, and wide in the west and narrow in the east. Both ends gradually turn into dolomite and slate. The hanging wall is slate and dolomite, and the surrounding rock of foot wall is dolomite. From north to south, there are successively distributed fluorite Nb rare earth iron ore, massive Nb rare earth iron ore, riebeckite Nb rare earth iron ore, and aegirine Nb rare earth iron ore, which accounts for 19.40%, 5.10%, 49.20% and 27.35% of total iron ore, and are rated as medium-low grade iron ore or medium grade rare earth ore.

BayanObo deposit covers 170 minerals in China, which is maximum in China, including 20 silver-tantalum minerals, 30 rare earth minerals, and 16 new minerals and new variety. The surrounding rock is dominated by dolomite, calcite, and quartz, the rare earth minerals dominated by monazite, bastnaesite, huanghoite, parisite, and apatite, the iron-bearing minerals dominated by magnetite and hematite, the Nb minerals dominated by niobite, aeschynite, fergusonite, pyrochlore, and ilmenorutile, the gangue minerals dominated by fluorite, aegirine-augite, riebeckite, and barite, and the sulfide dominated by pyrite and galena. According to the characteristics of paragenetic minerals, the iron ore is divided into massive iron ore, aegirine iron ore, riebeckite iron ore, fluorite iron ore, and biotite iron ore. The ores are dominated by banded, disseminated, and massive structures. The rare earth ore is dominated by dolomite rare earth ore, and occurs as disseminated, veinlet-disseminated, and massive structures, with subhedral shape, etc. The ores show various textures, and they are dominated by fine grained texture, with scaly and filamentary textures.

3.1 3D Geological Modeling of Deposit

After decades of research, there are large amount of geological data in BayanObo deposit [14–21], which provides rich theory and practical basis for 3D geological modeling and research on mathematical model in this paper.

3.2 Processing of Basic Data

The research reports, and engineering geology and other maps in BayanObo deposit over the years were collected and handled for next modeling and analysis. The basic maps and data for modeling of geological body were treated firstly.

a. Processing of borehole data

The research of 3D geological modeling covers exploration area of main mining area and east mining area. The data are from the borehole data of the past and present exploration in BayanObo deposit.

The main procedures of data processing are establishment of complete borehole data, and vectorization of profile estimating reserve in exploration line. The borehole data cover drilling No., inclinometry table, test table, and data in the drilling histogram were put into Excel to form the table, which is import into 3D mine with "Geological Database" to establish borehole database and form the map. The procedures are shown in Fig. 2.

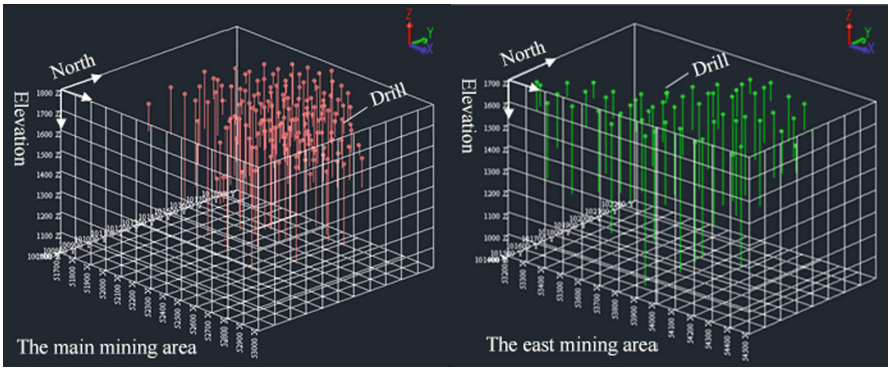


Fig. 2. Distribution of drilling engineering in the main mining area and the east area of BayanObo.

b. Processing of profile data

The contour lines of ore body, formation, and fault of same type in the profile were extracted respectively, and then connected between exploration lines to form the triangle network and establish the triangular facet TIN model through equal central angle, minimum distance, and equal division. The detection of self-intersection in each connection of triangle triangulation network makes correct calculation of entity volume. Constraint on connection shape of ore body with "Control line" and "Region" ensures good agreement with real geologic body.

The contour lines of ore body, formation boundary, vein, and fault in different ore bodies were extracted after vectorization of 13 and 14 exploration line profiles respectively in the main mining area and eastern mine area, and the corresponding lines of ore body, which reflect distribution and geological characteristics, were connected in 3D mine. The formation, vein, and fault were treated in the same way as shown in Fig. 3.

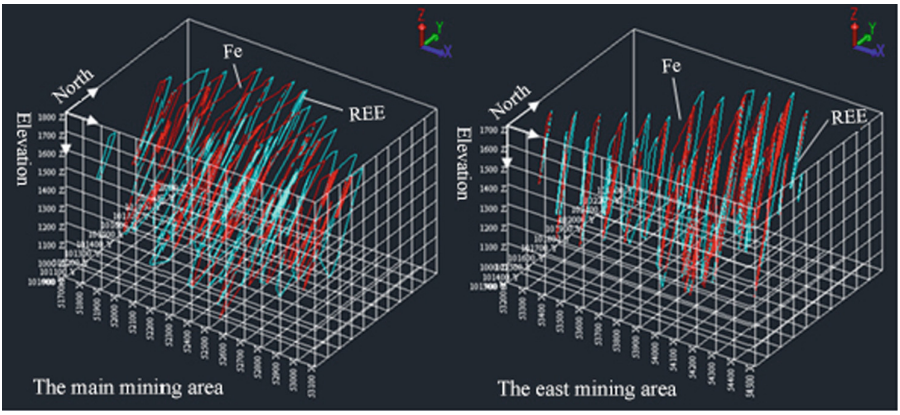


Fig. 3. The extraction of Fe and REE ore bodies in the main mining area and the east mining area of BayanObo.

3.2.1 3D Geological Model of Deposit

a. 3D Engineering Model

The exploration engineering is dominated by exploratory trench, shallow hole, and drilling. The borehole data, the firsthand information logged by geological personnel in the drilling site, play a very important role in establishing the geological profile and acquiring the geological information in deep formation. The status of line arrangement and drilling operation of drilling engineering in the main mining area and east mining area is visually displayed in 3D drilling model and terrain model after analysis and combination in 3D mine. The entity model is shown in Fig. 4.

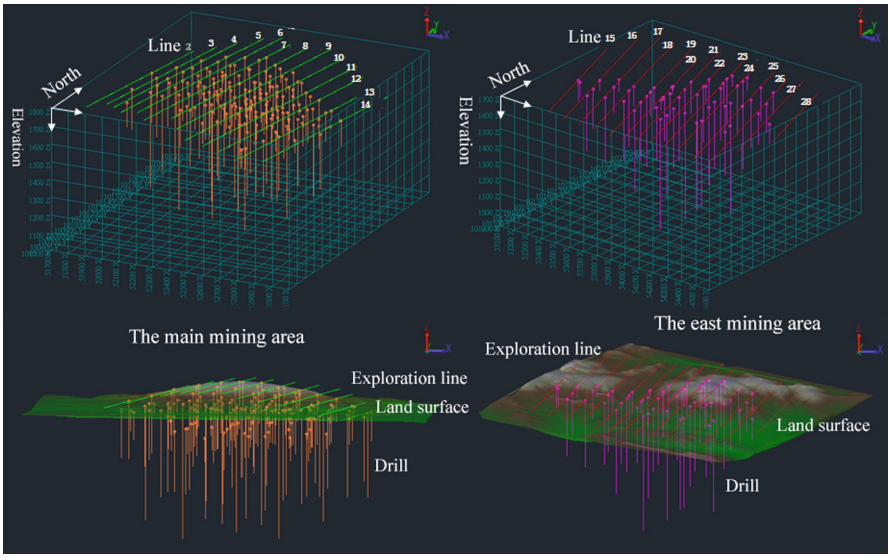


Fig. 4. Distribution of drilling engineering in the main mining area and the east mining area of BayanObo.

b. 3D Terrain Model

Vectorization of surface contour lines in the main mining area and east mining area of BayanObo deposit was conducted in Edit Module of MapGIS. After treating discontinuity and self-intersecting lines, the real altitude values were added to contour lines, which were import into 3D mine for line-generation of DTM surface model. 3D terrain model in Fig. 5, which is visual DTM diagrammatic figure, shows that the terrain is basically flat in the main mining area and relatively high in northern area. Compared with the main mining area, the east mining area is slightly rugged. On the whole, the main mining area and east mining area of BayanObo deposit are dominated by undulating topography, and have nearly EW strike. BayanObo Mountain (main mining area) has highest altitude of 1800 m, with basement exposed at medium degree, undeveloped river system, and intermittent stream in summer.

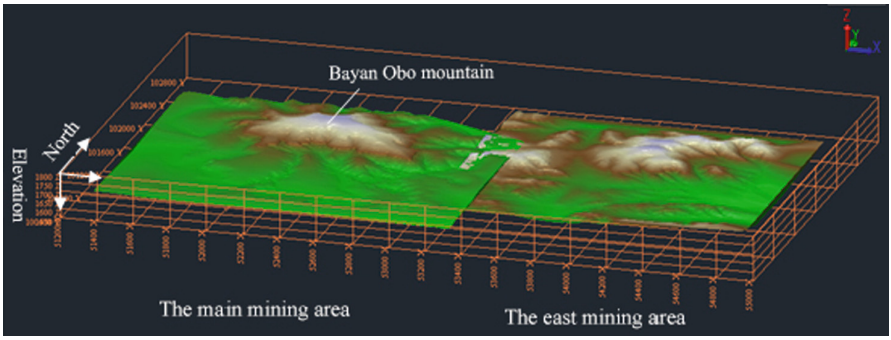


Fig. 5. The 3D Terrain model in the main mining area and the east mining area of BayanObo.

c. 3D Formation Model

The database of exploration lines was established through vectorization and geometric correction of exploration line and hole in the main mining area and east mining area. The entity lines were connected in 3D mine to establish the corresponding 3D entity model of formation. As shown in Fig. 6, the formation consists of glimmerite, dolomite, and slate, besides quartzite, metasandstone, feldspatite, and aegirinite.

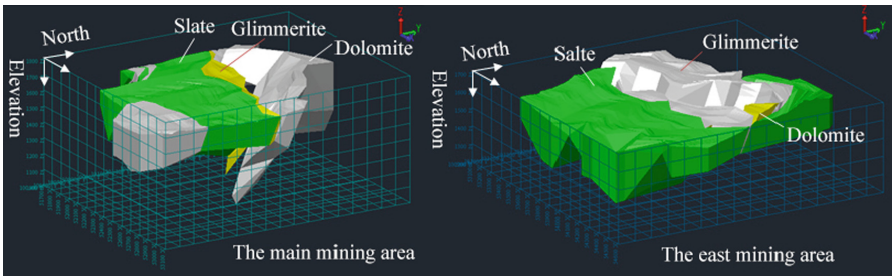


Fig. 6. The 3D formation model in the main mining area and the east mining area of BayanObo.

d. 3D Rock Mass Model

The rock mass is dominated by granite, gneissose granite, diorite, acid dyke, and intermediate and basic dyke, which are shown in Fig. 7.

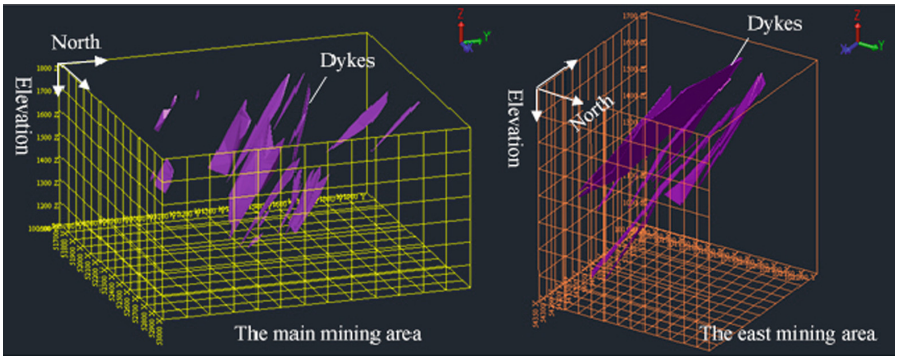


Fig. 7. The 3D rock mass model in the main mining area and the east mining area of BayanObo.

e. 3D Ore Bodies Model

The ore bodies of east mining area have greater variation in some region than the main mining area, but the whole ore bodies show some regular variation (Fig. 8).

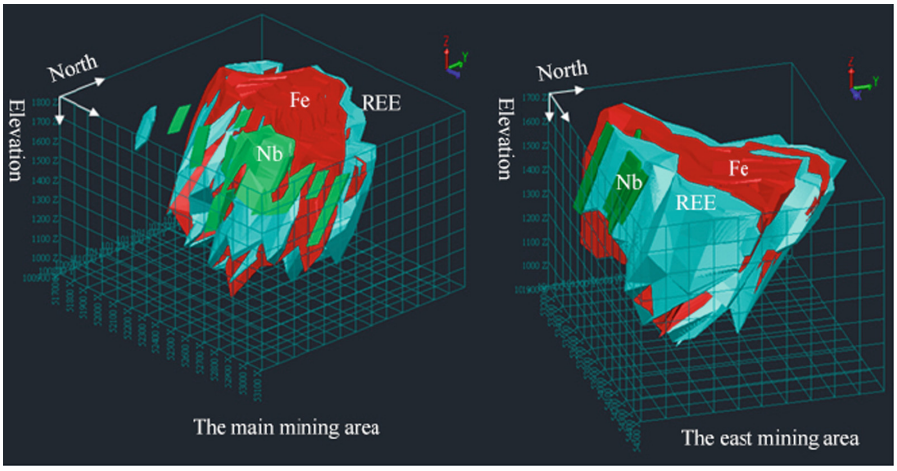


Fig. 8. The 3D ore body model in the main mining area and the east mining area of BayanObo.

In the plane, the ore bodies gradually thin, bifurcate, and pinch out toward both ends and downward, both main and east mining area vary regularly. The pinch-out occurs suddenly in some region, e.g. most of eastern block in east mining area bifurcate and pinch-out between No. 23 and No. 24 exploration lines.

There are often developed hidden lenticular minor ore bodies in the foot wall of main ore bodies, which occur both in the main and east ore bodies. ZK83-8 hole in No. 19 exploration line in the east ore bodies show greater thickness.

In the pinch-out, e.g. western end and lower ore bodies in 114-22 hole of east ore bodies, the ore bodies are enlarged suddenly, forming tuberculiform.

4 Conclusions

Applying the professional software of 3D geological modeling with advanced theories and methods, a 3D visualization model of main and east mining area in BayanObo deposit, including underground true 3D visualization model of engineering, terrain, formation, rock mass and ore bodies, which provides scientific basis for proving deep formation and boundary of ore bodies, calculating resource reserve, and well developing and protecting rare earth resource. The 3D geological modeling not only redisplays 3D information of geological bodies, which directly describes complex geological structure, but also provides supporting and qualitative and quantitative analysis method of spatial data manipulation in geological analysis and prediction.

Acknowledgments. This work was supported by the Project "Research and development of geological information product system and social service" (No. DD20160353) of China Geological Survey.

References

1. Bai, G., Yuan, Z.: Genetic Analysis of Bayan Obo Ore Deposit. China Academy of Geological Sciences Institute of Geology of Mineral Deposits, Beijing (1983)
2. Zhou, J., Zheng, Y., Yang, X., Shu, Y., Wei, C., Xie, Z.: Paleo plate tectonics and regional geology at Bayan Obo in Northern Inner Mongolia. Geol. J. China Univ. **8**(1), 46–61 (2002)
3. Liu, Y.: Geomatics and Earth Observation Science (EOS) for disaster management: an overview. J. Geomech. **14**(3), 212–220 (2008)
4. Christian, J.: 3D geoscience modeling: computer techniques for geological characterization. Earth-Sci. Rev. **40**(3–4), 299–301 (1994)
5. Mark, J.: Three-dimensional geological modelling of potential-field data. Comput. Geosci. **27**(4), 455–465 (1994)
6. Ehlen, J., Harmon, R.: GeoComp 99: GeoComputation and the Geosciences. Comput. Geosci. **27**(27), 899–900 (2001)
7. Li, Y., Guosheng, Q., Chen, J.: Realization of 3D subsurface geological modeling software in urban areas based on borehole data. Geol. Bull. China **24**(5), 470–475 (2005)
8. Zeng, Q., He, X.: Mathematical model and display method of three dimensional geological modeling. Eng. Geol. Comput. Appl. **3**, 1–8 (2006)
9. Wang, M., Bai, Y.: The status quo and development tendency of 3D geosciences modeling. Soil Eng. Found. **20**(4), 27–29 (2006)
10. Hou, E., Wu, L.: Present state and developing trend in the research on main issues of – 3D geoscience modeling. Coal Geol. Explor. **28**(6), 5–8 (2000)

11. Wang, R., Li, Y., Liu, Y., Xiang, Z.: Import and determination methods for virtual borehole in geo -3D modeling. Geol. Prospect. **43**(3), 102–107 (2007)
12. Fan, H., Xie, Y., Wang, K., Wilde, S.: Methane-rich fluid inclusions in skarn near the giant REE-Nb-Fe deposit at Bayan Obo, Northern China. Ore Geol. Rev. **25**(3), 301–309 (2004)
13. Yang, K., Fan, H., Fangfang, H., Li, X.: Skarnization age in the giant Bayan Obo REE-Nb-Fe Ore district, Inner Mongolia, China: Rb-Sr isochrone dating on single-grain phlogopite. Acta Petrol. Sin. **23**(5), 1018–1022 (2007)
14. Qingrun, M.: The genesis of the host rock-dolomite of the Bayan Obo Iron Ore deposits and the analysis of its sedimentary environment. Geol. Rev. **5**, 012 (1982)
15. Zeng, Y., Wang, F., He, Z.: Study on composition of inclusions in minerals and simulation experiment on hydrothermal metasomatic process of the Bayan Obo Iron deposit. Acta Geol. Sin.-Engl. Ed. **60**(4), 43–55 (1986)
16. Drew, L., Meng, Q., Sun, W.: The Bayan Obo Iron-rare-earth-niobium deposits, Inner Mongolia, China. Lithos **26**(1–2), 43–65 (1990)
17. Bas, M., Kellere, J., Tao, K.: Carbonatite dykes at Bayan Obo, Inner Mongolia, China. Mineral. Petrol. **46**(3), 195–228 (1992)
18. Campbell, L., Henderson, P.: Apatite paragenesis in the Bayan Obo REE-Nb-Fe Ore deposit, Inner Mongolia, China. Lithos **42**(1), 89–103 (1997)
19. Yang, X., Bas, M.: Chemical compositions of carbonate minerals from Bayan Obo, Inner Mongolia, China: implications for petrogenesis. Lithos **72**(1–2), 97–116 (2004)
20. Zhang, F., Zhao, Z., Li, Y.: Reduction kinetics of Bayan Obo coexisted iron and niobium ore by carbon-bearing pellet. Adv. Mater. Res. **418–420**, 346–352 (2011)
21. Xu, C., Taylor, R., Li, W.: Comparison of fluorite geochemistry from REE deposits in the Panxi region and Bayan Obo, China. J. Asian Earth Sci. **57**(6), 76–89 (2012)

Application of the Evidence Right in the Quantitative Evaluation of Rural Residential Area

Chao Tang[1,2(✉)] and Longyi Shao[1]

[1] China University of Mining & Technology, 11 Xueyuan Road, Haidian District, Beijing 100083, China
tangchao0312@126.com
[2] Beijing Urban Construction Exploration & Surveying Design Research Institute Company, Ltd., Beijing 100101, China

Abstract. Development and spatial ability are the main content of the quantitative distribution characteristics of rural residential areas. This quantitative study of Jincheng River Basin rural settlements distribution as an example, discusses the potential effect on spatial distribution of rural settlements distribution in the main influence factor, using weights of evidence method to quantify the influence degree of each factor, based on the rural residential space distribution characteristics of the quantitative evaluation. According to the model, select the related to the distribution and residential elevation, slope, road, water, per capita income, residential area, etc. factor as evidence layers, through the calculation of evidence layers distribution of rural settlements in the study area posterior probability, has carried on the quantitative evaluation to the whole study area, residential distribution. The study found that for the distribution of rural settlements in the study area and the factors of influencing the influencing sequence: slope > road > height > system > farmers per capita net income > per capita residential covers an area of, and the factors and the residents point was positively correlated; in the study area rural residential space layout need optimization and adjustment; the means of GIS can be effective to multi-source heterogeneous data for rapid optimization and comprehensive analysis and the forecast evaluation results in a quantitative way represented, effectively promote the multiple determinants of complex model, prediction and evaluation from qualitative analysis to quantitative development.

Keywords: Rural residential area · The model of evidence right · Quantitative research · GIS

1 Introduction

Rural residential area is an important place for production and life of rural residents. The scale sequence structure and internal function structure are forms of combination about its space layout and intra-area. At present, the quantitative research theory system of rural residential area has been established, but previous studies had focused on residential area related to land consolidation, planning, land intensive utilization, etc.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 363–371, 2017.
DOI: 10.1007/978-981-10-3969-0_40

Along with our rural social and economic development, change of natural environment, and the urbanization process deepen change of the coupling relationship about internal various influencing factors in rural residential area [1–3]. Therefore, it is necessary to make further discussion of quantitative research for the law of formation, development and evolution [4].

Many studies had shown that the location change of rural residential area is the result of residents' choices of location under the comprehensive influence of social economy, natural environment, and regional culture [5, 6]. The characteristics of the spatial layout in rural residential area are obvious for the influence of comprehensive factors, it shows the geographical differentiation of dot distribution and along the axis development of space. Evidence right model is a kind of space position relations on data and mathematics evaluation model for effective comprehensive of many favorable factors (evidence factors) combining with GIS technology. We make the quantitative evaluation for change impact factor of the spatial distribution characteristics in rural residential area combining with evidence right model, in order to further study the inherent development law of rural residential area, and provide certain theoretical basis for optimization of the rural residential areas space layout.

2 Research Data and Method

2.1 The Survey of Data in the Study Area

The study area is located in Jincheng river basin, involving 47 administrative villages in three towns of Zezhou. It includes 7 administrative villages in Xiacun, 23 administrative villages in Dadonggou and 17 administrative villages in Chuandi, the total area is 108 km^2.

This study is based on the following data: a. Extracting data of rural residential area, waters, roads and administrative boundary in river basin of Jincheng city from Zezhou county 1:1 million usage situation map of present land; building the rural residential area attribute database in river basin by inspection and processing in accordance with the thousands 1:1 scale drawing specification, on the basis of the data from 'The second national land survey results data downsizing technical indicators specification'. b. Extracting information of slope and elevation in the study area from river basin 1:50000 DEM data; c. The social and economic data of research needs and villages' population data are provided by Zezhou statistical yearbook 2010, the sixth national population census.

2.2 Evidence Theory

Evidence right model is geological statistics method of integrating mathematical statistics, image analysis and artificial intelligence. Its basic principle is first prior probability calculation to get conditional probability under the condition of corresponding geological evidence layer. Applied to the quantitative evaluation of rural residential area, the model takes each kind of influence distribution information of residential area as an evidence factor of quantitative evaluation, weight of each

evidence factor determines the value of the factor's contribution to the quantitative evaluation. In practice, every evidence factor first is converted to binary variables to represent any space position occurring or not in the study area, and then the habitable probability value of rural residential area can be calculated [7–9].

a. *Calculation of prior probability*: prior probability of evidence factor is to estimate the percentage of incident occurring or not in areas where evidence factors exist. Assuming that the total area of study is A, which can be divided into some pixel units and the number of pixel is N. Area of each pixel is u, the number of events in the study area about the main evaluation is D. Selecting randomly a pixel unit, the probability of events evaluated is:

$$P = P(D) = D/\mathrm{N} \tag{1}$$

Priori probability (O):

$$O = O(D) = \frac{P(D)}{1 - P(D)} = \frac{D}{\mathrm{N} - D} \tag{2}$$

b. *Calculation of evidence weight*: any evidence factor weight corresponding binary image as follows:

$$W^+ = \ln\left\{\frac{P(B/D)}{P(B/\bar{D})}\right\} \quad W^- = \ln\left\{\frac{P(\bar{B}/D)}{P(\bar{B}/\bar{D})}\right\} \tag{3}$$

In arithmetic expression, weight of evidence factors existent area B is W+, weight of evidence factors nonexistent area $\bar{B}$ is W−. When lacking of original data, the regional power value is 0. The degree of correlation between the evidence layers and the event is $C = W^+ - W^-$, if C > 0, evidence layer is advantageous to the events, positive correlation; if C < 0, evidence layer is disadvantageous to the events, negative correlation; if C = 0, evidence existing or not has no effect on events, irrelevance.

c. *Calculation of posterior probability*: if condition of event to evaluate about these evidence factors in study area is independent by inspection, the possibility of any pixel unit event can be expressed by logarithmic posterior probability:

$$\ln\left\{O\left(D/B_1^k B_2^k \mathrm{L} B_n^k\right)\right\} = \sum_{i=1}^{n} W_i^k + \ln O(D) \ (\mathrm{i} = 1,\ 2,\ 3\ldots,\ \mathrm{n}) \tag{4}$$

$$W_i^k = \begin{cases} W^+ & \text{Evidence factors exist} \\ W- & \text{Evidence factor does not exist} \\ 0 & \text{Data Missing} \end{cases} \tag{5}$$

Posteriori decay probability is expressed:

$$O_{\text{Posterior}} = \exp\left\{\ln\left(O_{\text{Priorprobability}}\right) + \sum_{j=1}^{n} W_j^K\right\} \tag{6}$$

Above all, Posteriori probability is:

$$P_{\text{Postprobability}} = O_{\text{Postprobability}} / \left(1 + O_{\text{Postprobability}}\right) \tag{7}$$

3 Choice of Evidence

The residential area distribution generally relates to altitude, landform, water, climate, population, transportation, economy, environment and other natural and social factors, the relationship is complex. We mainly collect and use indicators of elevation, slope, road, water system, per capita income, per capita covers an area in the study area as quantitative evaluation index (Table 1), based on the practical situation of economic development and the natural factors such as geographical environment of the study area. The characteristic value of each influence factor can be get by evaluation of coupling relationship between it and characteristics of residential area distribution, shown as Fig. 1.

Table 1. Impact factor of quantitative evaluation of residential area.

Quantitative evaluation objects	Influence factor	Index of quantitative predictors	Characteristic variables	Eigenvalue
Characteristic of residential area distribution	Social factor	Road	Road buffer	100 m buffer
			Road density analysis	(0, 160) section
		Per capita income	Per capita income distribution	163–627 yuan
		Per capita covers an area	Per capita covers an area distribution	207–261 m^2
	Natural factor	Elevation	Livable elevation range	711–800 m
				800–890 m
		Slope	Livable slope range	0°–1°
		Water system	Water system buffer	300 m buffer

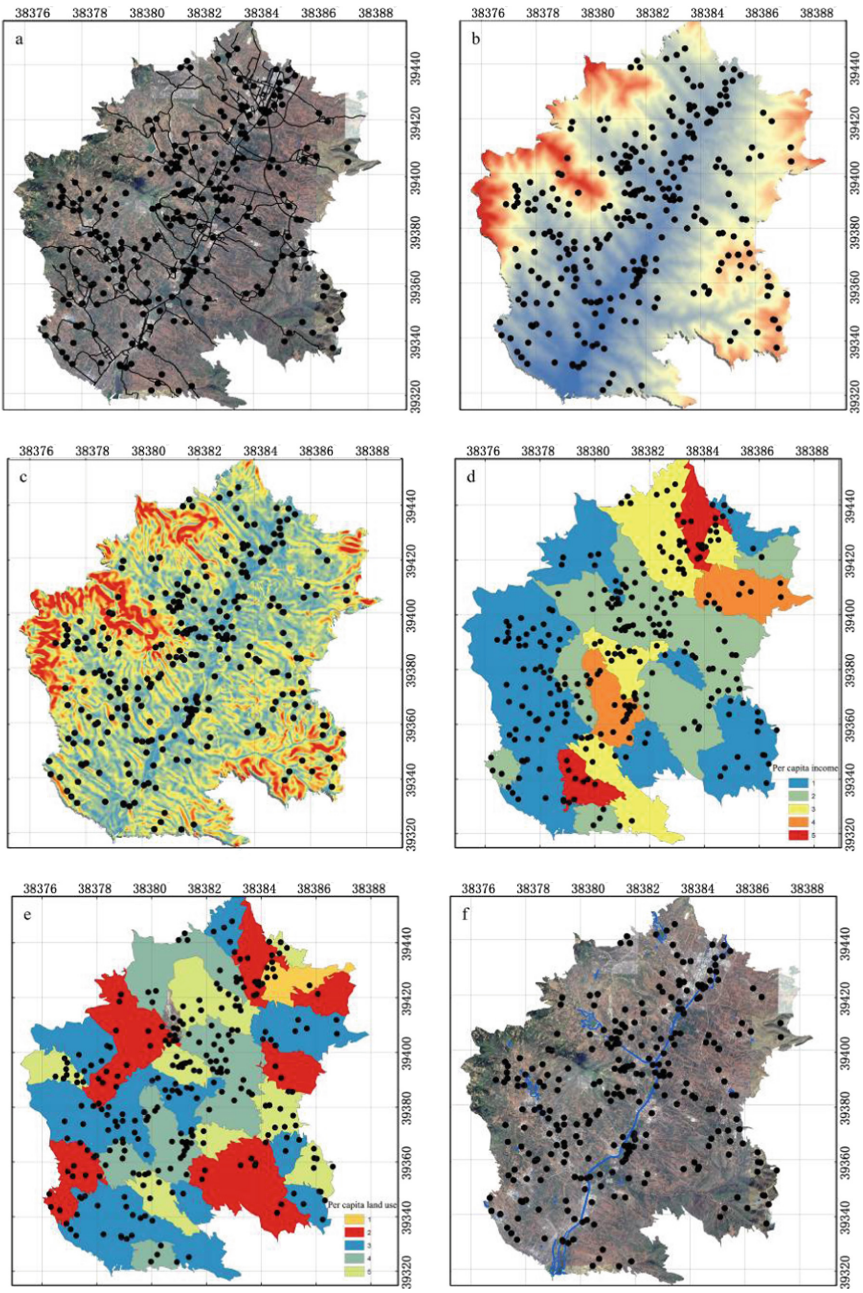


Fig. 1. Distribution of the residential area and the influencing factors. (a) Road network; (b) Elevation; (c) Slope; (d) Per capita income; (e) Per capita land use; (f) Water system.

4 The Establishment of the Model and Discussing

The results of quantitative evaluation is a residential area suitability posterior probability figure, with a value between 0–1. A variety of topography and geomorphology, social economy and the establishment of the cultural in study area provide the necessary data base for application of evidence right method; the analysis of the favorable evidence layer provides a variety of ancillary data for the application of evidence right method in study area. According to the established favorable evidence layer in the preceding thematic maps, first calculating separately the prior probability of evidence factor, then calculating the degree of relationship among the distribution of residential area and evaluation of the evidence weight, and making quantitative evaluation of the various units about residential distribution characteristics in study area (shown as Tables 2 and 3).

Table 2. A priori probability statistics of evidence factor.

Serial number	Evidence factor	PV1	PV 2	PV 3	PV 4
1	Road	0.833333	0.288363	0.166667	0.711637
2	Road equidensity	0.573643	0.164694	0.426357	0.835306
3	Elevation 1	0.461240	0.156805	0.538760	0.843195
4	Elevation 2	0.631783	0.329586	0.368217	0.670414
5	Slope	0.980620	0.481262	0.019380	0.518738
6	Per capita income	0.410853	0.259566	0.589147	0.740434
7	Per capita covers an area	0.379845	0.235503	0.620155	0.764497
8	Water	0.201550	0.086785	0.798450	0.913215

Note: PV1, when evidence factor appears, probability of residential area appear; PV2, when evidence factor appears, probability of residential area disappear; PV3, when evidence factor disappears, probability of residential area appear; PV4, when evidence factor disappears, probability of residential area disappear.

Table 3. Weight values of main evidence factors in study area.

Serial number	Name of evidence factors	W+	W−	C	Sorting
L5	Slope evidence right	0.711773	−3.287165	3.998936	1
L1	Road evidence right	1.061214	−1.451572	2.512785	2
L2	Equidensity evidence right	1.247917	−0.672522	1.926845	3
L3	Elevation evidence right 1	1.078918	−0.447929	1.526845	4
L4	Elevation evidence right 2	0.650709	−0.599223	1.249932	5
L8	Water evidence right	0.842606	−0.134299	0.976904	6
L6	Per capita income evidence right	0.459223	−0.22856	0.687783	7
L7	Per capita covers an area evidence right	0.47804	−0.209249	0.687287	8

According to the quantitative evaluation model, we calculate characteristic value of the rural resident area space distribution in every evaluation unit (after showed by the posterior probability values), to analyze the habitable degree of the rural residential areas in study area. Calculation results show that the distribution value of posterior probability distribution value is between 0.000045 and 0.972139.

The rural residential area can be divided into two classes of habitable or inhabitable as 0.6 for critical point in the study area, combining with histogram of posterior probability frequency posterior probability and the analysis of natural fracture method classification tool in ArcGIS software. According to the posterior probability relative size, it is divided into different grades with different colours, to generate colour piece figure about quantitative evaluation of posterior probability (Fig. 2). The colour piece of evaluation unit that is deeper means that the position is more conducive to the development of the rural residential area layout, habitable level is high. Specific classification result is shown as Table 4:

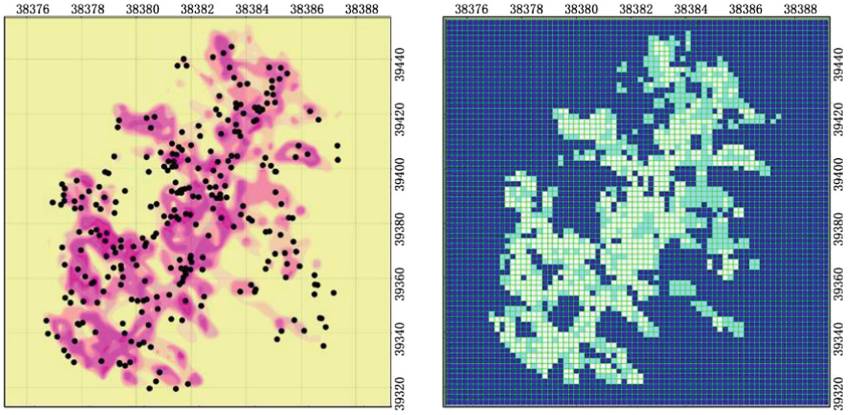


Fig. 2. Probability and color blocks graph of the quantitative prediction of the spatial distribution' characteristics of residential areas. (Color figure online)

Table 4. Evaluation and classification of rural residential area in the study area.

Posterior probability	Habitable level	The cell number	Area (hm^2)	Area of rural residential zone %
>0.6	Habitable	493	252.42	19.78
<0.6	Inhabitable	1999	1023.50	80.22

As the above figures and tables, the space distribution of residential area has obvious zoning characteristic in study area. The slope of deep color piece is gentle, main roads are packed and elevation is low, rural residential area distribution is concentrated. The evaluation units' number of evaluation unit posterior probability >0.6 is 493, the area is 252.42 hm^2, which accounts for only 19.78% of the total area of the

residential areas in the study area. It mainly distributes in Chuandi village, Hecun village, Jiaohe village, Shanglu village, Yunan village, Shuanghedi village, and Tianhu village and so on, these zones have better condition of the region, gentle topography, near the water, convenient transportation, near center town, relatively good social and economic conditions, and therefore these rural residential areas are relatively dense and suitable for rural residents. However, the area of low value about rural residential area is 1023.50 hm^2, up to 80.22% of the total area of the rural residential areas in the study area. It mainly distributes in remote mountainous area such as Gouxi, Zhongjie, Wanghushan, Heiquangou, and Dongshan and so on. These zones have poor condition of the area, low social and economic condition, far away from water, inappropriate for residents' travel and farming, these rural residential areas are small size and scattered distribution. Above all, the overall natural environment condition of study area is bad, social and economic condition is low. So the space layout of these rural residential areas is unreasonable, which needs to optimize and adjust to some extent.

5 Conclusions

According to the established evidence evaluation model, we make quantitative evaluation on the distribution of residential area in study area (taken the posterior probability of the evaluation results as final), and the distribution range of value is 0.08–0.08 in study area. The dark map-spot in figure (high posterior probability) shows good zonality, habitable residential areas are relatively dense, the opposite are relatively sparse [11, 12].

There are 273 residential areas in the whole study region, among them there are 254 residential areas where the posteriori probability value is greater than 0, accounting for more than 93% of the total. Higher posteriori probability areas are characterized by denser residential concentrations. The forecast evaluation result has a good indication [13].

The model of evidence right is based on GIS technology; it links the point type of discrete event with layer. Studies had shown that only evidence right method could closely combine with divided unit and statistical calculations in many evaluation methods [14]. We can get the following conclusions, by means of quantitative evaluation about the rural residential areas distribution in study area.

a. The advantage of using evidence right to forecast and evaluate the rural residential areas distribution is simple methods and principles; weight distribution is easy to understand.

b. It is considered that slope has a greater influence on the residential area distribution, relatively flat terrain and less slope is easy to structure and distribute building; the relationship between spatial distribution of residential area and road network is close, and they promote each other; Drainage system is extremely important to growth of animal and plant, the residential areas distribution often appear within a certain distance from the drainage system. All of these are based on the arrangement of the evidence weights and comprehensive analysis of rural residential areas distribution condition in the study area [15].

c. GIS can optimize rapidly and make effective comprehensive analysis of multi-source heterogeneous data. It can represent prediction evaluation result in the form of quantitative, and effectively promote the multiple influence factors and complex model prediction evaluation developing from qualitative to quantitative.

References

1. Agterberg, F.: A modified weights-of-evidence method for regional mineral resource estimation. Nat. Resour. Res. **20**(2), 9–101 (2011)
2. Zhao, Z., Pan, M., He, L.: Paleo plate tectonics and regional geology at Bayan Obo in Northern Inner Mongolia. Acta Scientiarum Naturalium Universitatis Pekinensis (04), 594–600 (2010)
3. Gong, J., Wang, Z., Cai, E.: Evaluation of spatial distribution of basic farmland conservation area based on fuzzy weight of evidence model. Res. Soil Water Conserv. (04), 161–167 (2015)
4. Xu, X., Wan, Q.: A quantitative study on spatial distribution of rural settlement in floodplains and discussion of its application. Geogr. Res. **16**(03), 47–54 (1997)
5. Li, C., Yang, B., Ye, S.: A study on town and village system in the mountains based on fractal theory: a case study of Nanxi in Dabie Mountain. J. West Anhui Univ. (02), 12–15 (2013)
6. Li, R., Yang, C., Jiang, X.: An analysis of optimized distribution of rural residential land in villages. GeoComp 99: GeoComput. Geosci. **28**(6), 93–98 (2011)
7. Xu, L., Hu, H., Zhang, L.: Fractal research and planning of yuan tong town residential area scale layout. Econ. Geogr. (S1), 150–153 (2001)
8. Qin, Y.: The study about optimizing spatial distribution of rural settlements and mode selection. China University of Geosciences (Beijing) (2006)
9. White, R., Engelen, G.: Cellular automata and fractal urban form: a cellular modeling approach to the evolution of urban land-use patterns. Environ. Plan. A **25**(8), 1175–1199 (1993)
10. Batty, M., Longley, P.A.: Fractal Cities: A Geometry of Form and Function. Academic Press, H Arcourt Brace & Company, Publishers, London (1994)
11. Benguigui, L.: When and where is a city fractal? Environ. Plan. B **27**, 507–519 (2000)
12. De Keersmaecker, M.-L., Frankhauser, P., Thomas, I.: Using fractal dimensions for characterizing intra-urban diversity: the example of Brussels. Geogr. Anal. **35**(4), 310–328 (2003)
13. Feng, J.: Spatial-temporal evolution of urban morphology and land use structure in Hangzhou. Acta Geogr. Sin. (03), 343–353 (2003)
14. Chen, W., Huo, M., Ma, Y.: Characteristics of the spatial distribution and evolution of rural settlements. J. Henan Agric. Univ. (03), 354–358 (2014)
15. Zhang, M., Chen, R., Tang, C., et al.: Correlation analysis of extraction mechanism of remote sensing anomaly with mineralization and ore-controlling-illustrated by the case of Qimantage Area, Qinghai Province. Commun. Comput. Inf. Sci. **482**, 471–485 (2015)

Research on Detection and Trend Forecasting Technologies of Micro-blog Hot Topic

Qi Fu[1(✉)] and Jun Tan[2]

[1] Institute of Communications and Electronics,
Jiangxi Science & Technology Normal University, Nanchang, China
jxsyfq@sina.com
[2] Physical Education Institute, Jiangxi University of Finance and Economics,
Nanchang, China
tanjun_jiangxi@163.com

Abstract. Based on the collection and study of the extensive literature, this paper concludes and classifies the detection and forecasting technologies and its current application status in the micro-blog hot topic. Furthermore, combined with research characteristics of the detection and prediction of micro-blog hot topic and including the domestic characteristics, we draw out the limitations of the current related research, and point out the direction for further improvements. Finally, it has carried on the forecast on the future prospect.

Keywords: Micro-blog · Hot topic detection · Hot trend prediction · Review

1 Introduction

Micro-blog as an important social network helps users to understand the personal and social groups a lot. Therefore, the study for the data of micro-blog has also become a research hotspot. China Internet Network Information Center (CNNIC) showed that as of December 2015, Chinese netizens reached 688 million, Internet penetration rate reached 50.3%, half of the Chinese people have access to the Internet. Meanwhile, in the field of integrated social networking, micro-blog using rate of Chinese netizens was 33.5%. The report also described, micro-blog met user demand for the main interest information. It is an important platform of user accessing to and sharing, such as hot news, interested in content, expertise, and public opinion. Over the past year, the scale of micro-blog users gradually increased, and the value of its content platform has been further improved.

In recent years, in the KDD (ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), WWW, PAKDD (Pacific-Asia Conference on Conference Knowledge Discovery and Data Mining), SIGIR (Special Interest Group on Information Retrieval Conference) and other important international and domestic publications and conferences, there has been more and more research on the micro-blog hot topic. Currently, about micro-blog hot topic research mainly contains two aspects: First, research of detection micro-blog hot topic, in order to follow the trend forecasting research to provide evidence; Second, research of prediction micro-blog hot topic. To help users find potential rule or solve real life problems hidden in the micro-blog data,

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 372–378, 2017.
DOI: 10.1007/978-981-10-3969-0_41

such as finding hot events, identifying opinion leaders, monitoring web content, and detecting negative public opinion. Based on study and combing research of micro-blog hot topic appeared in important domestic and international journals getting along with conferences in the past few years, the paper summarized research status of micro-blog hot topic, and summarized the current problems in research of micro-blog hot topic to further explore the prospects of micro-blog hot topic.

2 Detection of Micro-blog Hot Topic

2.1 The Method of Statistical Analysis

On the micro-blog platform, hot events often cause a lot of people's attention in a short time, resulting in a large number of comments and forwarding information. According to this feature, scholars believe whether monitoring occurrence frequency for the keywords of a given event can suddenly surge in a given time segment, if it is, corresponds to occurring an event; if not, do not think. Through the changes of emotional keywords to find hot events micro-blog and made emotional language distribution model, Yang achieved the hot events finding by analyzing differences between the emotional-distribution language models in the adjacent time [1]. Zheng detected a large number of emergent theme keywords online in micro-blog, clustering them to find hot news event or events [2]. Similarly, Lee defined BursT weighted formula for keyword, and introduced the sliding window, in order to achieve real-time monitoring the occurrence of top events and discrimination [3, 4].

2.2 The Method Based on Learn Model Analysis

According to topic keywords by four benchmarks chosen, Long established graph model for topic keywords to cluster, and then through the clustering results to find hot micro-blog events [5]. Moreover, putting forward graph model based on wavelet analysis in literature [6] and time and space model based on probability in literature [7], they also achieved effective detection of micro-blog events using their models. In addition, based on the traditional LDA (latent Dirichlet allocation) model, and through the micro-blog data associated with their association characteristics, many scholars improved LDA model. For example, considering contact relationship and text association of micro-blog, Zhang proposed micro-blog generational model based LDA, this is MB-LDA, to assist theme and event mining of micro-blog [8]. Based on the relationship between the micro-blog post and the expanding LDA model, Li proposed a method of hot topic discovery based on a special-topic space model [9].

2.3 The Method Based on the Improved Similarity Measure

Phuvipadawat converted text to vector space model with TF-IDF method firstly, and proposed the improved TF-IDF method based on named entities weighting [10]. The method measured the similarity of information by adjusting the feature weight to find micro-blog events more accurately. And on the basis of considering the diverse

characteristics of micro-blog data, Tong proposed an event-discovery algorithm based on semantic similarity of Chinese text feature, time similarity and social similarity [11]. Under the premise of full extraction of user characteristics, blog features and event characteristics, Gupta dig the relationship between features to achieve richer and more complete data to create an event detection model [12].

Analysis of these three methods can improve discovery or detection effect of micro-blog hot events, but that several analytical methods do not take into account that the dynamic characteristics of the event's propagation, nor take into account the specific characteristics of the micro-blog data on the impact of hot events finding. Then we need to focus on dynamic detection methods of micro-blog hot events.

3 Trend Prediction of Micro-blog Hot Topic

The research of micro-blog hot topic prediction often was based on hot micro-blog discovery. At present, there is not much achievement of trend prediction in domestic and international research. Research has been focused on the prediction of a topic micro-blog transmission range, including the topics related number to micro-blog's release and change of forwarding number with time and so on.

In literature [13], timing features and characteristics of the text combined with the micro-blog platform presented an iterative semantic analysis and prediction models of hot topic that is TopicRank, to predict trends influence on the topic in the next period of time. Attributing the user comments and social relationship network as the topic of the feature, Jamali proposed topic epidemic classification prediction algorithm [14]. Hong put forward a topic feature analysis and propagation trend prediction algorithm based on data mining and wavelet analysis [15]. By analyzing the time distribution characteristics of the users' participation in the topic discussion, Gomez proposed user behavior prediction method in short and long term [16, 17]. Through the analysis of the role of the media in the process of communication, Wu found that the topic of information from the high impact of the user as an intermediary to carry out selective transmission, while the topic of different types of users concerned is different [18]. Then Sun used the number of micro-blog fans to judge users' influence, thus predicting the potential audiences of micro-blog [19]. But Cha and Romero pointed out that the fans in the social network was a passive recipient of information, only to rely on the number of fans not only cannot accurately assess the impact of users, but would hinder the communications of topic in the user relationship network [20, 21].

Another part of the study is related to micro-blog's forwarding mechanism, for example, Yu stressed that the forwarding behavior of Sina's micro-blog in the common phenomenon, is more significant than Twitter, is the decisive factors of a hot topic [22]. Hong also analyzed the popularity of micro-blog from the forwarding behavior, and used binary-classified method to predict the hot topic [23]. Micro-blog forward behavior not only involved some of the characteristics of the topic itself, but also related to the interests of users and emotional analysis and other research areas. So the analysis of micro-blog's forwarding behavior will help to analyze and judge the trend of the spread of the topic and the possibility of becoming a hot spot. The reasons and influencing factors of the forwarding of micro-blog were analyzed, and the influence of

the users' concerns and micro-blog's content characteristics were found to be influenced to forwarding behavior in the literature [24–26]. And on the other hand, Petrovic started from the experiment to prove the feasibility of forwarding prediction [27]. The literature [23, 28] based on classification and collaborative filtering method to design a prediction algorithm of forwarding behavior, but the effect is not ideal. Yang extracted 22 features which would influence transmission, using the information-gain method for each feature is weighted for prediction of micro-blog forwarding behavior. The results showed that the user number of fans and whether referred to others in micro-blog is an important factor in the impact of micro-blog forwarding [29].

4 Research Characteristics on the Discovery and Prediction of Hot Topic in Micro-blog

In summary, hot topic detection and prediction research generally have the following characteristics:

(1) They can be realized by conventional probability statistics, data mining, and time series analysis method so on, being paid great attention to the application of statistical method and natural language processing technology, and more use of data mining and information extraction technology. Moreover, research trends are becoming more practical.
(2) In view of the specific application, we can obtain the better effect by selecting the appropriate topic characteristic and the analysis method.
(3) In China, although the research scholars have carried out much strong analysis, but the analysis is often one-sided, the accuracy of the research results should be improved. Tracing the reasons, it is mainly due to, the micro-blog information network is a complex information system, with characteristic of the community structure, which has huge number of users, the rich content of the topic and the user's behavior is not determined and so on.
(4) How to make a unified description of the micro-blog topic related features is the key issue of micro-blog hot topic detection and prediction.

5 Future Directions for Improvement

In that way, being specific to China's micro-blog hot topic research and what are the deficiencies, the next step of research should be how to further improve and deepen the study on the basis of inheritance? The specific performance is the following three aspects.

5.1 To Focus on the Dynamic Detection Methods

The following research should focus on the dynamic detection methods of micro-blog hot spots, while reducing the impact of text-data sparsity. Micro-blog topic detection

and prediction analysis is mostly affected by the influence of micro-blog text data sparsity. A Chinese micro-blog is generally not more than 140 characters, there are data sparse problem, making too much deviation through the calculated feature value of the text similarity compared with actual values. The features result in taking into account the syntax analysis, ignoring the order between words when based implicit theme approach extracts text feature information of micro-blog. And in the time of similarity matching, is unable to deal with a sequence of events orderly. Also implicit theme method is mainly aimed at the closed collections of texts. Moreover, at the step of micro-blog topic finding, we have to deal with large amounts of information released by massive users immediately, and micro-blog text information has some new characteristics, such as temporal continuity, and so on.

5.2 To Improve Analysis Research of Micro-blog Hot Topic

First, we should break through the micro-blog interface resource constraints. Information on micro-blog is rich in resources, but the available resources are limited. Such as, micro-blog's open platform to provide interfaces, restricting the content and the times of being grabbed, so that it must be combined with the crawler systems in order to get more micro-blog information. Second, improve the analysis efficiency of micro-blog hot topic detection system. By comparing the existing micro-blog hot topic detection systems in China finds the results page is slow, and the user experience is poor when the analysis results of a certain micro-blog. Third, improve the accuracy of micro-blog hot topic analysis. In China, research in the field of hot topic started late, and the accuracy of user influence analysis, sentiment analysis, event prediction and other aspects should be improved.

5.3 To Improve Application Research of Micro-blog Hot Topic

First, we should develop the application about of event prediction and management. Micro-blog event prediction is still the focus of the current study, but the results have yet to be improved. Through the research on micro-blog event prediction, we can strengthen the monitoring and control of the individual or relevant departments of the state for a particular event or series of events. Therefore, changing the state of post processing will improve the control ability and initiative of the emergency. Second, our aim is to develop commercial applications. Micro-blog has a huge number of users, where hides a lot of commercial value information, such as finding the user purchase intention to recommend goods through a hot topic.

6 Conclusion

Micro-blog hot topic research shows a trend of diversification. How to connect with international mainstream research, to realize the localization of foreign frontier theory, it is the current domestic academic should be actively thinking about the main issues. The authors think that, we should explore and grasp the future trend of micro-blog hot

topic research from two aspects: firstly, actively carry out comparative study of being on Twitter and other social media, and domestic and foreign research status. That would help to find the problems, grasp the trend, and so as to expand the research space. Secondly, to enhance communication among micro-blog researchers, especially the exchange and interaction among interdisciplinary researchers, it is true to form a blend of multidisciplinary research team which fused of computer science, communication science, management science, psychology and so on. That is one of the important factors to keep the micro-blog research activity.

In short, micro-blog hot topic research is still in the initial stage of development, but with strong development potential. From vertical and horizontal aspects to further deepen and broaden research space of micro-blog hot topic, and combining with the actual situation in China to construct the localization theory, to make the research results have practical significance, will be the main goal of the future micro-blog hot topic.

Acknowledgments. In this paper, the research was sponsored by the subject of Jiangxi "Twelfth Five-Year" plan for Social Science (Project No. 15TQ07).

References

1. Yang, L., Lin, Y., Lin, H.: Micro-blog hot events discovery based on affective distribution. Chin. J. Inf. **26**(1), 84–90 (2012). (in Chinese)
2. Zheng, F., Miao, D., Zhang, Z.: A method of Chinese micro-blog news topic detection. Comput. Sci. **39**(1), 138–141 (2012). (in Chinese)
3. Lee, C.-H., Wu, C.-H., Chien, T.-F.: *BursT*: a dynamic term weighting scheme for mining microblogging messages. In: Liu, D., Zhang, H., Polycarpou, M., Alippi, C., He, H. (eds.) ISNN 2011. LNCS, vol. 6677, pp. 548–557. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21111-9_62
4. Lee, C.: Mining spatio-temporal information on micro-blog streams using a density-based online clustering method. Expert Syst. Appl. **39**(10), 9623–9641 (2012)
5. Long, R., Wang, H., Chen, Y., Jin, O., Yu, Y.: Towards effective event detection, tracking and summarization on microblog data. In: Wang, H., Li, S., Oyama, S., Hu, X., Qian, T. (eds.) WAIM 2011. LNCS, vol. 6897, pp. 652–663. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23535-1_55
6. Weng, J., Lee, B.: Event detection in Twitter. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, pp. 401–408. AAAI, Barcelona (2011)
7. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp. 851–860. ACM, Raleigh (2010)
8. Zhang, C., Sun, J., Ding, Y.: Micro-blog theme mining based on MB-LDA model. Comput. Res. Dev. **48**(10), 1795–1802 (2012). (in Chinese)
9. Li, J., Zhang, H., Wu, H.: Chinese micro-blog hot topic mining system based on specific domain—BtopicMiner. Comput. Appl. **32**(8), 2346–2349 (2012). (in Chinese)
10. Phuvipadawat, S., Murata, T.: Breaking news detection and tracking in Twitter. In: Proceedings of the 2010 International Conference on Web Intelligence and Intelligent Agent Technology, Toronto, pp. 120–130 (2010)
11. Tong, W., Chen, W., Meng, X.: EDM: micro-blog event detection algorithm. http://www.cnki.net/kcms/detail/11.5602.TP.20121019.1017.001.html. Accessed 26 Feb 2013. (in Chinese)

12. Gupta, M., Zhao, P., Han, J.: Evaluating event credibility on Twitter. http://www.cs.uiuc.edu/hanj/pdf/sd12_mgupta.pdf. Accessed 07 Feb 2012
13. Yang, G.: Micro blog hot topic discovery strategy research. Zhejiang University master's degree thesis, Hangzhou (2011). (in Chinese)
14. Jamali, S., Rangwala, H.: Digging digg: comment mining, popularity prediction, and social network analysis. In: Proceedings of International Conference on Web Information Systems and Mining, Shanghai, pp. 32–38 (2009)
15. Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in Twitter. In: Proceedings of the 20th International Conference on World Wide Web (WWW), pp. 57–58 (2011)
16. Gomez, V., Kaltenbrunner, A., Lopez, V.: Statistical analysis of the social network and discussion threads in Slashdot. In: The Proceedings of the 17th International Conference on World Wide Web (WWW), Beijing, China, pp. 645–654 (2008)
17. Zhu, T.: Research on the role of nodes and community evolution in social networks. Doctoral dissertation, Beijing University of Posts and Telecommunications, Beijing (2011). (in Chinese)
18. Wu, S., Hofman, J.M., Mason, W.A.: Who says what to whom on Twitter. In: Proceedings of the 20th International Conference on World Wide Web (WWW), pp. 705–714 (2011)
19. Sun, S.: Hot topic detection and tracking technology in Chinese micro blog. Master degree thesis of Beijing Jiaotong University, Beijing (2011). (in Chinese)
20. Cha, M., Haddadi, H., Benevenuto, F.: Measuring user influence on Twitter: the million follower fallacy. In: Proceeding of the 4th International AAAI Conference on Weblogs and Social Media (2010)
21. Romero, D.M., Galuba, W., Asur, S., Huberman, B.A.: Influence and passivity in social media. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011. LNCS (LNAI), vol. 6913, pp. 18–33. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23808-6_2
22. Yu, L., Asur, S., Huberman, B.A.: Artificial inflation: the true story of trends in SinaWeibo, arXiv preprint arXiv:1202.0327 (2012)
23. Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in Twitter. In: Proceedings of the 20th International Conference on World Wide Web (WWW), pp. 57–58 (2011)
24. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: conversational aspects of retweeting on Twitter. In: 43rd Hawaii International Conference on System Sciences (2010)
25. Suh, B., Hong, L., Pirolli, P., Chi, H.: Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In: IEEE Second International Conference on Social Computing (SocialCom), pp. 177–184. IEEE (2010)
26. Welch, J., He, D., Schonfeld, U., Cho, J.: Topical semantics of Twitter links. In: WSDM 2011 (2011)
27. Petrovic, S., Osborne, M., Lavrenko, V.: RT to win! Predicting message propagation in Twitter. In: AAAI 2011 (2011)
28. Zaman, T.R., Herbrich, R., Gael, J.V., Stern, D.: Predicting information spreading in Twitter. In: Workshop on Computational Social Science and the Wisdom of Crowds, NIP 2010 (2010)
29. Zhang, Y., Lu, R., Yang, Q.: Prediction of the forwarding behavior of micro blog. J. Chin. Inf. **26**(4), 109–114 (2012). (in Chinese)

The Implementation of Human Tracking with Quadrotor Aircraft

Yang Yang, Dongdong Huang(✉), and Nannan Cheng

School of Electrical and Engineering, North China University of Technology, Beijing 100144, China
huang_dong_dong@yeah.net

Abstract. The small quadrotor aircraft is a kind of unmanned aerial vehicle which is realized by the combination of four rotor blades. The features of quadrotor aircraft are with a simple mechanical structure, easy to control and so on [1]. Along with the continuous development of Unmanned Aerial Vehicle in recent years, the automatic tracking function has been gradually received more and more attention. This design mainly realizes autonomous flight and the human body automatic tracking function of the quadrotor aircraft, positioning of the human body and quadrotor aircraft are obtained by GPS. Because the precision of GPS is not high, in order to overcome this shortcoming, this design proposes the human body tracking based on electromagnetic wave ranging, which greatly improves the accuracy of GPS. The quadrotor aircraft could keep within the accurate human body range. Experimental results show that, when human is walking, quickly walking, jogging and running, the quadrotor aircraft can be maintained between 2–3 m ranges from human.

Keywords: Quadrotor aircraft · Autonomous flight · Human body automatic tracking · Electromagnetic wave ranging

1 Introduction

With the development of intelligent robot technology, the research of unmanned aerial vehicle (UAV) has been valued by more and more people. In both civil and military markets, UAV have demonstrated a great value of application [2–4]. The research of unmanned aerial vehicle is mainly focused on the following aspects: the design of the mechanical structure, the research on the accuracy improvement of the sensor, and the research on the strategy of autonomous positioning, automatic tracking and autonomous navigation combined with some other external conditions. In this paper, we mainly study the human body automatically tracking of the quadrotor aircraft.

Human tracking is an important part of the human motion detection and scene understanding. It includes the identification and location of human motion, in battlefield reconnaissance, intelligent video surveillance, traffic control, human-computer interaction and other areas has broad application prospects. Common human tracking technology include:1, wireless radio frequency positioning method, this is a based RFID detection method, detected person need to carry a an electronic tag, by receiving the data within the tag to determine the location of the human body. This method is

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 379–388, 2017.
DOI: 10.1007/978-981-10-3969-0_42

easy to operate, but has low flexibility [5, 6]. 2, based on the image sequence of human body tracking technology, which is achieved through the image sequence of the moving target to detect, extract, recognize and track, so as to obtain the trajectory of the target. This method has high precision, but the demand of hardware equipment is also high, and the data processing is very complex [7, 8]. 3, human target detection and tracking methods in infrared images, when the human body passes through the infrared ray between the transmitting end and the receiving end, some infrared ray will be blocked, and the controller analyzes the position distribution of the human body through the level change. This method has high accuracy, but it is only suitable for human body tracking in short distance [9]. 4, seamless Indoor/Outdoor Positioning Handover for Location-Based Services, using the GPS and WIFI realize indoor and outdoor positioning, but in the absence of WIFI signal position can not achieve positioning [10]. Many colleges also carry on the research to the human body tracking. The human body tracking based on pyroelectric infrared sensor network in An Hui university, which tracked the motion of the human body by pyroelectric infrared sensor [11]. Vision based human tracking technology in human-computer interaction in University of Electronic Science and Technology, using a single camera to capture the image, to calculate the depth of the scene information in real time, and then using the color information for the human body tracking [12]. There are also many studies have been performed to achieve target tracking from the quadrotor aircraft. L. Carrillo designed a Quad-rotor switching control for the task of path following. The objective consists of estimating and tracking a road without a priori knowledge of such path. However, the quadrotor aircraft can only fly according to the fixed path, can not be tracked [13]. Augustin Manecy proposed a novel hyperacute gimbal eye to implement precise hovering and target tracking on a quadrotor, presenting a new minimalist bio-inspired artificial eye, able to locate accurately a target placed in its small field [14]. But the demand of hardware equipment is also high, and the data processing is very complex.

This paper focuses on the human body tracking based on electromagnetic wave ranging, firstly, positioning of the human body and quadrotor aircraft by GPS, due to the accuracy of civil GPS is relatively low, positioning is not accurate, can only measure the approximate range roughly. Install the EVB1000 ranging module on the quadrotor aircraft, be able to accurately measure to the distance between the aircraft and the human, so that the accuracy is greatly improved. The quadrotor aircraft could keep within the accurate human body range. Experimental results show that, when human is walking, quickly walking, jogging and running, the quadrotor aircraft can be maintained between 2–3 m ranges from human. Improving the accuracy, one the one hand, when the human body is in motion, the body suddenly accelerated, to make aerial effect is better. On the other hand, UAV delivery of goods requires a more accurate positioning in the future.

2 System Structure Design

The quadrotor aircraft consists of propeller, flight control, GPS, control panel, ranging module (EVB1000) etc. As shown in Fig. 1: 1. GPS, 2. PIXHAWK, 3. Control panel, 4. Ranging module (EVB1000).

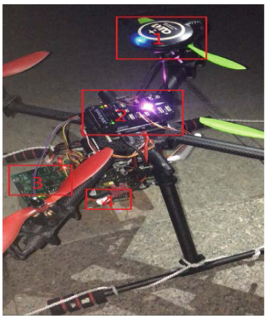


Fig. 1. The structure of quadrotor aircraft

The main purpose of this design is to study on human body tracking with quadrotor aircraft based on electromagnetic wave distance measurement. Firstly, using the control panel to generate PWM wave, sending to the flight control through I/O port, to realize the quadrotor aircraft autonomous flight, to instead of the original remote controller and the wireless receiving device, the aircraft can be separated from the remote controller to realize autonomous flight [2]. Secondly, the aircraft and the human body were installed on a GPS, to achieve the rough location. Due to the relatively low accuracy of GPS positioning, so to add EVK1000 ranging module. One EVB1000 is placed on a person, the other one is placed the flight control and connected to the flight control, which can measure the distance between the quadrotor aircraft and the human body, to achieve a more accurate positioning. Using EVB1000 as a hardware development platform, using eclipse as the software environment, the design of each module of the system is written by C language. And sending the distance between two points to the control panel of quadrotor aircraft though the serial port. When the safety distance is 2–3 m, the aircraft is in fixed mode; more than 3 m, the aircraft to follow the human body; less than 2 m, the aircraft away from the human body. To realize the automatic tracking function of the aircraft in the same direction to follow the movement of the human body. The system block diagram is shown in Fig. 2:

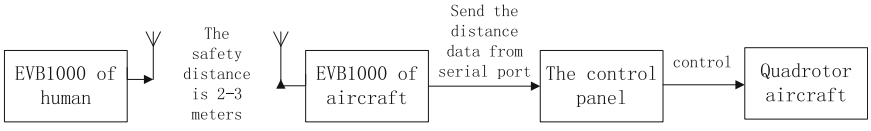


Fig. 2. The system block diagram

3 The Principle of Ranging Module and Distance Information Transmission

The EVK1000 consists of a pair of EVB1000 boards, based on the DW1000 chip of DecaWave, Complies with IEEE802.15.4-2011 UWB standard, it facilitates proximity detection to an accuracy of ±10 cm, and it is especially suitable for wireless sensor networks (WSN) applications. Taking into account its accuracy, this chip can be used as the current indoor positioning technology RFID and WiFi supplement. Taking STM32F105RC chip as processing chip, processing the collected data by DW1000, and then sent to the LCD display [15].

The ranging method uses a set of three messages to complete two-round trip measurements from which the range is calculated. As messages are sent and received the DecaRanging application retrieves the message send and receive times from the DW1000. These transmit and receive timestamps are used to work out a round trip delay and calculate the range [15]. Adding USART serial communication statement in the program, to realize the data transmission of flight vehicle control module. Figure 3 shows the arrangement and general operation of the two-way ranging [15].

DecaRanging' s Tag/Anchor two-way ranging algorithm, for this algorithm one end acts as a Tag, periodically initiate a range measurement, while the other end acts as an Anchor listening and responding to the tag and calculating the range. The two-way ranging algorithm is shown in Fig. 3:

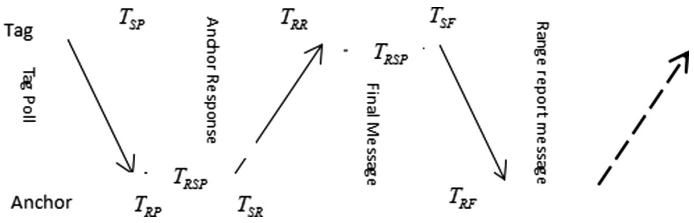


Fig. 3. Range calculation in DecaRanging

Data transmission time: $TOF = ((T_{RR} - T_{SP}) - (T_{SR} - T_{RP}) + (T_{RF} - T_{SR}) - (T_{SF} - T_{RR}))/4$, multiplying the TOF by C, the speed of light (and radio waves), gets the distance(or range) between the two devices [15]. After the EVK1000 measured distance, the distance between two points through the serial port to send to the control panel of quadrotor aircraft for communication.

4 Design and Implementation of Systems Software of Control Board

The main idea of system software design is mainly divided into the autonomous flight of the aircraft and the automatic tracking of the human body. The autonomous flight of the aircraft includes the steering engine data setting module and the control signal transmission module.

4.1 The Autonomous Flight of the Quadrotor Aircraft

The steering gear data setting part mainly completes the setting and change of each steering gear data. The quadrotor aircraft which uses Pixhawk as the flight control module, mainly adopts the four steering gears: lifting steering gear, front and rear moving steering gear, left and right moving steering gear, horizontal rotating steering gear. The different numerical values of each steering gears represent different forces in the direction of the aircraft. When the aircraft is at rest, by changing the numerical value of each steering gear to enable the aircraft to complete different actions, by adding delay to make it a complete process. The aircraft can unlock, take off, hover, drop and lock. In this design, the steering gear data is changed once every 70 ms.

The control signal encoding and sending module mainly completes the coding and sending of steering gear. The control signal of aircraft uses the S-BUS communication protocol and sent constantly, period is 14 ms, every 14 ms to collect the data of all steering gear, code and send the data once. Sending need 3 ms, and then according to the S-BUS protocol waiting for 11 ms and once again to collect the steering gear data and send.

After downloading the program to the development board, the system is initialized. The steering gear data setting module according to the set time changes the steering gear data gradually, the control signal encoding and sending module sends the signal with a period of 14 ms, after the flight control module receives the signal, it start to decode and drive motor to make the aircraft complete flight action. The flow chart is shown in Fig. 4.

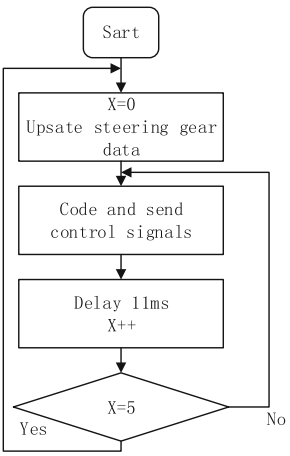


Fig. 4. Autonomous flight flow chart of quadrotor aircraft

4.2 Human Body Automatic Tracking

Positioning design of aircraft and people mainly uses the GPS module which is on the human body and flight control. After power up, turn on the GPS, and then measure the

approximate range roughly between human body and quadrotor aircraft. If it does not find the human body, then the control panel will send commands to the quadrotor aircraft land. If it is within the range of the human body, then the control panel will start EVK1000 ranging. One EVB1000 is placed on a person, the other one is placed the flight control and connected to the flight control, which can measure the distance between the quadrotor aircraft and the human body, to achieve a more accurate positioning. And sending the distance between two points to the control panel of quadrotor aircraft though the serial port. When the safety distance is 2–3 m, the aircraft is in fixed mode; more than 3 m, the aircraft to follow the human body; less than 2 m, the aircraft away from the human body. To realize the automatic tracking function of the aircraft in the same direction to follow the movement of the human body, and thereby achieve human-computer interaction. Among them, the maximum operating time is 10 min, more than 10 min, the quadrotor aircraft will automatically drop. The flow chart of the human body tracking system is shown in Fig. 5:

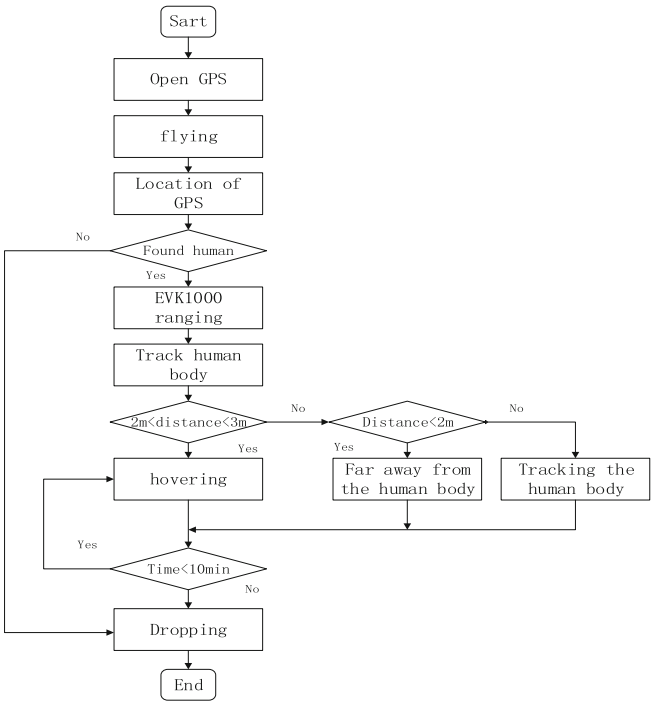


Fig. 5. Flow chart of human body autonomous tracking

5 Experimental Result

This experiment tests the effect of the human body automatic tracking function with the quadrotor aircraft in the actual environment, collecting data and then analytical processing the data. The tester is walking along the red line, as shown in Fig. 6, by A to B.

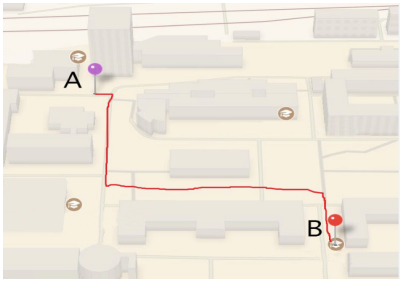


Fig. 6. Flow chart of human body autonomous tracking (Color figure online)

The tester is in the front, the quadrotor aircraft is in the rear the human body about 2.5 m. When the quadrotor aircraft takes off, the aircraft is in a fixed mode. When the tester move forward, and the distance between people and aircraft is over 3 m, the quadrotor aircraft begins to follow the body. There are four stages that include slowly walking (about 1 m/s), quickly walking (about 2 m/s), jogging (about 2.5 m/s) and running (about 3 m/s), each stage is 1 min. Each 100 ms records a data, and according to the observation data of the quadrotor aircraft more than 3 m made error statistics. As shown in Fig. 7:

Fig. 7. The quadrotor aircraft more than 3 m made error statistics

The tester is walking along the red line again, as shown in Fig. 6, by B to A. The quadrotor aircraft is in the front, the tester is in the rear the quadrotor aircraft about 2.5 m. When the quadrotor aircraft takes off, the aircraft is in a fixed mode. When the tester move forward, and the distance between people and aircraft is less than 2 m, the

quadrotor aircraft begins to away from the body. There are also four stages that include slowly walking (1 m/s), quickly walking (2 m/s), jogging (2.5 m/s) and running (3 m/s), each stage is 1 min. Each 100 ms records a data, and according to the observation data of the quadrotor aircraft less than 2 m made error statistics. As shown in Fig. 8:

Fig. 8. The quadrotor aircraft less than 2 m made error statistics

From the figures can be concluded, the error probability of slowly walking stage is very low, the error probability of quickly walking stage and jogging stage is moderate. And the error probability of running stage is relatively high. After analysis and discussion, the reasons for the error are:

- EVK1000 module itself exists ±10 cm error, so the quadrotor aircraft in 3.1 m and 1.9 m appears a lot of mistakes.
- Affected by the hardware conditions, the control of quadrotor aircraft is one of the reasons for the error.
- People suddenly accelerate, slow down, stop and turn, all will be made the error of the human body automatic tracking function with the quadrotor aircraft

6 Conclusion

This design mainly realizes autonomous flight function and the human body automatic tracking function of the quadrotor aircraft, and through the corresponding hardware and software to realize. Mainly realized:1. The realization of the ranging function, 2.

The PCB design of the control board, 3. Realization of autonomous flight function of quadrotor aircraft, 4. Realization of human body automatic tracking of quadrotor aircraft, 5. Realization the human-computer interaction. However, from the theoretical design, hardware design, software algorithm can be further improved, in order to further enhance the tracking time and accuracy of the quadrotor aircraft. As well as we will consider removing the GPS, and then adding two EVB1000 modules, using four EVB1000 modules to achieve fixed-point for quadrotor aircraft, rather than simple ranging, so that the positioning accuracy is further improved, for the UAV delivery of goods in the future.

References

1. Sun, H.: Design and implementation of the quadrotor self-positioning system based on vision. Master thesis, University of Electronic Science and Technology of China, 05 (2012)
2. Mellinger, D., Michael, N., Kumar, V.: Control of quadrotos for robust perching and landing. In: Proceedings of the International Powered Lift Conference, Philadelphia, PA, pp. 520–526, 5–7 October 2010
3. Bouabdallah, S., Noth, A., Siegwart, R.: PID vs LQ control techniques applied to an indoor micro quadrotor. In: IEEE International Conference on Intelligent Robots and Systems, pp. 2451–2456 (2004)
4. Leishman, Gordon: The Breguet-Richet quadrotor helicopter of 1907. AHS Int. Dir. **2001**, 1–4 (2001)
5. Yimin, Z., Amin, M.G.: Localization and tracking of passive RFID tags. Wirel. Sens. Process. **9**(1), 1–11 (2006)
6. Ni, L., Liu, Y., Cho Lau, Y., Patil, A.: LANDMARC: indoor location sensing using active RFID. Wirel. Netw. **10**, 701–710 (2004)
7. Barber, D.B., Redding, J.D., McLain, T.W., Beard, R., Taylor, C.N.: Vision-based target geo-location using a fixed-wing miniature air vehicle. J. Intell. Rob. Syst. **47**, 361–382 (2006)
8. Dobrokhodov, V.N., kaminer, I.I., Jones, K.D.: Vision-based tracking and motion estimation for moving targets using unmanned air vehicles. J. Guid. Control Dyn. **31**(4), 907–917 (2008)
9. Bertozzi, M., Broggi, A., Caram, C., et al.: Pedestrian detection by means of far-infrared stereo vision. Comput. Vis. Image Underst. **106**, 194–204 (2007)
10. Hansen, R., Wind, R., Jensen, C.S., Thomsen, B.: Seamless indoor/outdoor positioning handover for location-based services in streamspin. In: Mobile Data Management, pp. 267–272 (2009)
11. Zhao, W.: The research of human tracking with pyroelectric infrared sensor network. Master Anhui University, 04 (2011)
12. Xie, X.: The application of human tracking technology based on vision in human-computer interaction. Master thesis, University of Electronic Science and Technology of China, 09 (2010)
13. Carrillo, L., Flores, G., Sanahuja, G., Lozano, R.: Quad-rotor switching control: an application for the task of path following. In: American Control Conference (ACC), June 2012, pp. 4637–4642 (2012)

14. Manecy, A., Diperi, J., Boyron, M., Marchand, N., Viollet, S.: A novel hyperacute gimbal eye to implement precise hovering and target tracking on a quadrotor. In: 2016 IEEE International Conference on Robotics and Automation (ICRA) (2016)
15. Decawave Company (2014). DecaRanging_ARM_Source_Code_Guide

QvHran: A QoE-Driven Virtualization Based Architecture for Heterogeneous Radio Access Network

Luhan Wang[1,2(✉)], Zhaoming Lu[1,2], Xiangming Wen[2], Lu Ma[2], Xin Chen[2], and Wei Zheng[2]

[1] Beijing Advanced Innovation Center for Future Internet Technology, Beijing University of Technology, Beijing, China
{wluhan, lzy0372}@bupt.edu.cn

[2] Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing, China
{xiangmw, malu, chenxin2014, zhengweius}@bupt.edu.cn

Abstract. Heterogeneous cloud RANs have been proposed as a cost-effective solution to promote wireless network coverage and data rate. However, many urgent issues, such as cross-tier interference avoidance, converged network management and the fulfillment of consistent experience quality, still need to be tackled. SDN and virtualization empower the high efficiency management of networks, based on which, a QoE driven RAN architecture (QvHran) is proposed. QvHran aims at providing a new organization and management norm to perform consistent management for future H-CRANS. We design the QvHran architecture from both management and deployment aspect, and also study the relationships between them. Based on the proposed architecture, its key supporting technologies are also presented, include heterogeneous resource virtualization, network situation awareness, and elastic allocation of virtual resources. Lastly, a simulation is given to demonstrate the elastic resource allocation in QvHran. Simulation results has shown that RAN performance can be improved a lot in QvHran.

Keywords: HetNets · SDN · Virtualization · RAN architecture · QoE

1 Introduction

According to Cisco Visual Networking Index 2015 [1], global mobile data traffic will increase nearly 10-fold between 2014 and 2019 and will reach 24.3 Exabyte per month by 2019. The IMT Vision has also put forward much higher requirements in fifth generation (5G) wireless systems, such as 1000× improvement of data rate compared with 4G. To meet the overwhelming demand of radio access in different situation, various access technologies have been proposed and deployed, bringing heterogeneity into the radio networks [2].

Traditional distributed deployment of heterogeneous BSs or access points would requires a large amount of CAPEX and OPEX. Besides, the dense and random

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 389–400, 2017.
DOI: 10.1007/978-981-10-3969-0_43

deployment could cause severe inter-tier interferences resulting into spectrum efficiency and energy efficiency performance degradations. Consequently, heterogeneous cloud radio access networks (H-CRANS) were proposed by authors in [3]. H-CRANS are supposed to improve the cooperative processing and transmission, and reduce overall costs in HetNets by fully exploiting cloud computing based technologies. Another great challenge faced by RAN is the overwhelming emerging wireless services. The various types of wireless services would require a more accurate and fine-grained management of radio resources. Context-aware cognitive networks can automatically configure devices and their parameters, systems, and services based on users' contexts, in which services' quality of experience (QoE) is being deemed as a most promising criterion. QoE is a subjective measurement combines with users' perception, experience and expectations. How to precisely acquire the services' QoE and perform service oriented network management, especially how to guarantee the QoE in different RANs, are still open problems. Due to the centralized processing of baseband signals, H-CRANS has also given a great opportunity for the QoE-driven network management.

Current radio access network (RAN) architecture are usually designed for a specific RAT, which often failed to fully utilize existing heterogeneous wireless resources and provide consistent experience of quality when users move among different RANs. Future heterogeneous RANs need to realize the seamless convergence of multiple RATs. To address the above problems, we propose a new architecture for heterogeneous cloud RAN: a QoE-driven virtualization based architecture for heterogeneous radio access networks (QvHran). QvHran is an SDN and virtualization based RAN architecture, we make a separation between data plane and control plane. In the control plane, we abstract the resources in heterogeneous network into a uniformed virtual resource, which is directly related with services requirement. With the centralized control of virtual resource, QvHran could provide a consistent management mechanism for heterogeneous network resources.

In the next section, we will have a brief review of current efforts in software and virtualization based wireless networks. And in Sect. 3 we focus on describing the system overview of QvHran. In Sect. 4, several key technologies will be presented in the QvHran, such as their work flow, and some use cases. And a simple simulation of elastic resource allocation is provided to verify the performance of QvHran. Lastly, we give a brief conclusion and future work directions.

2 Related Works

Software defined networks (SDN) [4] was first introduced in wired network. The foundation of SDN is the separation of control functions and data forwarding in networks. The control functions are realized on a centralized controller, and provide open interfaces which the network administrators could use to program the behavior of traffic and network. With SDN, it can simplify network management and enable faster innovation and evolution. Due to its revolutionary design and advantages, SDN has already attracted significant attention from both academic and industry. Another emerging technology, network virtualization, is also playing a more and more

important role in networking. Network virtualization [5] separates the logical network part from the physical infrastructures, and the main advantage of virtualization is the implementation of time-sharing mechanisms, which could lead to increased efficiency in using the available physical resources. SDN and network virtualization are different from each other, but they usually work complementary to each other.

Many efforts have been done to apply SDN and virtualization to wireless networks to address the challenges posed by the complex wireless network situation and the increasing customer demands. [6–9] are some recent SDN-based research activities in wireless networks. In [9], the authors proposed a programmable wireless data plane that provides programming interfaces across the entire wireless stack and refactors the wireless protocols into processing and decision planes, which is supposed to make the radio access network more flexible to evolve and introduce intelligence into RAN. Zhou in [6] applied the resource virtualization in a novel framework for collaboration in heterogeneous networks, called CHORUS, and showed that the CHORUS could save the system energy consumption. In [7], the authors explored the SDN approach for dense deployed networks. In order to overcome the computational overhead of centralized controller, they proposed a dynamic, two-tier SDN controller hierarchy. Based on new architecture, they also identified the challenges of MAC layer reconfiguration, dynamic backhaul reconfiguration and connectivity management.

Current network virtualization researches, such as [10–12], mainly focus on the sharing of physical networks, such as infrastructure sharing, spectrum sharing and time sharing. In [11], the authors described a network virtualization substrate for effective virtualization of wireless resources in cellular networks. The authors also studied slice scheduler, flow scheduling framework, and admission control in NVS. In [6, 8], the authors mentioned the abstraction of radio resources, and proposed to virtualize the physical radio resources. Other than to share the network infrastructures or physical resources, they used virtualization to manage the heterogeneous networks. They proposed to virtualize the physical radio resources and allocate them to users. However, there still lacks a deep study of the advantages of such abstraction and how to utilize it.

3 Architecture and Features of QvHran

Benefiting from SDN, QvHran also makes a separation between the control plane and data plane, and introduce resource virtualization into RANs. Cloud computing is exploited to realize the programmable data plane and resource virtualization. The proposed architecture provides a new mechanism to organize and manage the heterogeneous radio access networks. In QvHran, we redesign the RAN architecture from both its deployment and management aspect.

3.1 Hierarchy from Network Management Aspect

As shown in Fig. 1(a), from the aspect of network management, RAN is separated into three layers, including **physical access layer**, **virtual network resource layer** and **the service layer**. The *physical access layer* consists of the necessary infrastructures in

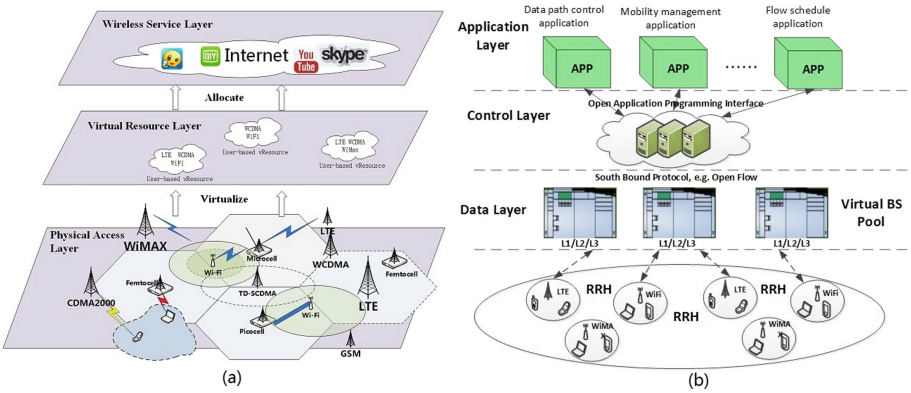


Fig. 1. (a) QvHran architecture from the network management aspect. (b) QvHran architecture from the network deployment aspect

RAN, such as base stations, relays, antenna access points, etc., and radio resources, including spectrum, time slot, space, code, etc.

The duty of the *virtual network resource layer* is to build a map between the radios resources and the available bandwidth from the network aspect. Physical radio resources from different RANs is abstracted into the format that are directly related to services' requirement, and bring the resource closer to the services. For the network management, it makes the resource manage framework more flatten. The operator could directly allocate to the services what they ultimately need. The prime goal of RAN is to satisfy the requirement of the services according to the expected quality of experience. In order to make the management more accurate and intelligent, the RAN should know what the services status are, what they still need, and what the RANs can provide. So we argue that the context-awareness would be critical important for future heterogeneous RAN. In QvHran, we not only open the basic status parameters to network administrators, but also open the synthetical results. We have verified the performance of network situation awareness in the prototype of the proposed architecture in our previous work [16]. We explored two kinds of awareness technologies:

(1) Network status situation perception technology: According to the radio interference, fading, the network load, and also service type, provide an assessment for the network status and tendency.
(2) Service QoE perception technology: quality of experience will be the optimized target in QvHran. It has the ability to percept the QoE of main types of wireless services.

3.2 Hierarchy from Network Deployment Aspect

From the aspect of RAN deployment, we introduce the generic SDN reference architecture and H-CRANS into QvHran. As Fig. 1(b) shows, there are three layers: the data layer, control layer, and the network management application layer.

Data Layer: In software defined wireless network (SDWN), data layer mainly refers to those collections of base stations, relays and other access nodes in wireless networks. In our previous work [13], we have already studied the programmable data plane in SDWN, here we introduce the dynamic programmable data layer into QvHran. By redesigning the base station, we separate the control plane from the access infrastructures. Cloud computing based processing is used to perform baseband functions in the virtual BS pool. Wireless functions are achieved by loading the software offering a variety of radio communications services. Besides, RRH entities, as mentioned in [3], consist of antenna system, power amplifier and low noise amplifier and A/D converter.

Control Layer: Logically centralized controller is the core component of SDN reference architectures [4]. In QvHran, the controller mainly performs management and awareness functions of the radio access networks. Corresponding to the resource management aspect, the network resource layer is realized on the controller. By using the unified virtual resource format, the controller hides the heterogeneity of different RANs (WiFi, 3G, LTE). Besides, network status situation awareness and service QoE assessment are also implemented in unified controller. The relationship between the controller and management hierarchy is shown in Fig. 2. Besides, The network controller communicate with the programmable data layer by south protocol, and make modular software abstractions for building complex network applications.

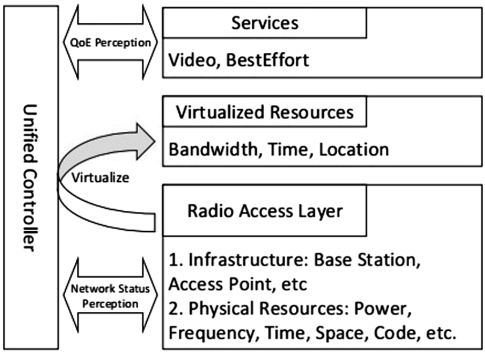


Fig. 2. Relationship between controller and management

Network Management Application Layer: Due to the programmability of SDN based architectures, network administrators could build highly scalable, flexible network services, gaining unprecedented programmability, automation and network control. In QvHran, applications running on this layer would perform network management, such as virtual resource allocation, load balance, QoE management, etc. Based on the network awareness results, more granular and accurate network management applications could be achieved.

4 Key Technologies in QvHran

4.1 RAN Resource Virtualization

Due to the introduction of programmable data plane, and the centralized controller, the consistent management of the heterogeneous network has been possible. One of the key technologies in QvHran to realize it is the virtualization of heterogeneous resources. Different RANs have different management and schedule strategies of wireless resources, such as frequency, time, code, spatial, and power. For example, in LTE networks, Radio resources are allocated into the time/frequency domain, while in 802.11 wireless networks apply a carrier-sense multiple access with collision-avoidance (CSMA/CA) scheme with exponential backoff procedure [14]. Because of the heterogeneity of different networks and their independent deployment, network resources cannot be dynamically allocated across different networks, which would cause load imbalance among different RANs. Besides, different networks have their corresponding optimization schemes, when user moves among different networks, the heterogeneous network cannot provide the consistent quality experience for users.

Heterogeneous networks have different strategies to use spectrum resources, network operators have to design different algorithm for each kind of RAN to fulfill the services' requirement. In QvHran, we build a virtual resource layer in the access network, which represents that at a certain location the number of RANs, the amount of bandwidth and time resource can be used by a UE. The virtual resources layer abstract the heterogeneous radio resources into the format that can be directly used by wireless services, which bring wireless services closer to the physical resource. Network management could allocate the virtual resources to services according to their requirement.

4.2 Network Status Awareness

QvHran has realized two main kinds of awareness, one is *network status awareness*, and the other one is *user status awareness*. The network status awareness technologies include not only current status of the RAN, but also the prediction of status trend. The status information can be opened to the network by using the open API provided by controller. Network status awareness is conducted in two layers, one is the physical layer, and the other one is network layer. In the physical layer, network status awareness refers to the channel status information perception, and the controller mainly measures the signal strength, interference, spectrum usage, etc. As to network layer, the controller will assess the network based on the network parameters, such connectivity, throughput, bandwidth, delay, jitter, packet loss rate, and also the service type running on this network.

Many researches have been done to assess the network status, but these works mainly focused on how to pick up parameters and calculate their corresponding result.

However, services have different requirements for network, which means even the same network status may have different impact on different kinds of services. For example, we pick up two main wireless services. One is best effort (BE) service, such as web page browsing, file download, etc. And another one is real-time large traffic services, such as on-line video or audio. Best effort service is usually burst and short time, while the later one usually need bandwidth reservation and delay guarantee. If one access point has already carried many BE services, the total load may not be very high, but it cannot guarantee a certain bandwidth and delay. Thus it may more suitable for BE services, not real-time services. We research the service feature based network performance assessment in the architecture of QvHran. When one service is going to access to the network, we abstract the requirements of the service and calculate the quality score for each available network. We argue that in future radio access network the service oriented network status assessment will be very important. We have studied the awareness technologies in our previous work [15].

User status awareness is mainly to measure the services' quality of experience (QoE). Wireless services QoE will be an important criterion for future network management and optimization. In QvHran, QoE perception models are built for different services and opened to network management application layer. Network administrators could use the standard QoE metric to manage the resource among services. We explored the role of human cognition and the psychophysics in QoE assessment. And built a model to map the service information, complete time, and bandwidth to get the BE services' QoE. For video services, we study a pixel-based model in the QvHran, which could obtain a more accurate result [16].

4.3 Elastic Network Resource Allocation

Due to the shortage of the spectrum and users' urgent requirement for network access, promoting the resource utilization and ubiquitous access have always been an important issue for the RANs. Providing ubiquitous access means user could connect to the network anywhere, anytime, with any object. Now many smart devices has been equipped with multiple network interfaces, and could connect to different kinds of networks. But to fully utilize the heterogeneous network resource is still a tough task for current RAN architecture. When users move into an overlapping coverage, they faced with network access selection, mobility control, load balance and other problems. Traditional RANs can't handle these problems feasibly, as there is no information sharing and traffic manage mechanism across access networks.

In QvHran, heterogeneous radio resources are abstracted as uniformed virtual resources particles. And network operators can also get the network status and user status. Collaboration between different networks, vertical handoff, and load balancing can be regarded as the process of virtual resource allocation and reallocation. We set up a simulation environment to verify the performance of elastic network resource allocation in QvHran.

As described in Sect. 4.1, the heterogeneous radio resources are virtualized into resource particles, which are represented by *bandwidth*, *time* and *location*. We define Virtual Resource Particle (VRP) as a resource unit that can be allocated to users', which represents that VRP contains β bps bandwidth resources and can be used by one user for τ ms. For a given user u_i, we use $k = \{1, 2, \ldots K\}$ to denote available BS/AP. And there will be K group of VRP for u_i maintained on the virtual resource layer. The available data rate on each BS/AP can be calculated as:

$$R_b^k = B_{avail}^k \log(1 + \frac{G_{u_i} P_{u_i}}{B_{avail}^k n_0 + \sum\limits_{u \neq u_i} G_u P_u}) \tag{1}$$

Where B_{avail}^k represents the available spectrum bandwidth on each BS/AP for u_i, G represents channel gain, and P represent the transmitting power for each user.

So the number of virtual resource particles on each BS/AP for u_i can be obtained, $n_k = \frac{R_b^k}{\beta}$. We define a comprehensive cost model to both consider service performance and spectrum cost. The first part of cost is the penalty for service degradation:

$$C_{penalty} = 1 + \log(\frac{R_{req}}{R_{provide}}), R_{req} \geq R_{provide} \tag{2}$$

Where R_{req} represents the user's requested traffic data rate, and $R_{provide}$ represents the provided data rate. It can be noticed that when $R_{provide}$ is equal to R_{req}, the penalty is 1, otherwise it will be larger than 1.

The second part of the cost is the spectrum consumption, $C_{spectrum}$. So the total cost can be denoted as:

$$C_{total} = \sum_{u_i \in U} (C_{penalty}^{u_i} + C_{spectrum}^{u_i}) \tag{3}$$

The goal of the virtual resource allocation is to allocate these VRPs to users according to their request traffic, while minimizing the total cost. We develop a Simulated Annealing (SA) based algorithm to solve this problem on our testbed platform. We apply the techniques introduced in [18] to calculate the *Temperature* used in SA. The initial temperature is calculated between initial allocation and other adjacent allocation schemes: $t_0 = (1 - \lambda_1 - \lambda_2) C_{total}^{avg} + \lambda_1 C_{total}^{\min} + \lambda_2 C_{total}^{\max}$, and later it is updated according to: $t := \frac{t}{1 + \beta t}$. In implementation, we set λ_1=0.25, $\lambda_2 = 0.25, \beta = 0.5$. The steps of the proposed algorithm is given in *Program SA-Allocation.*

```
program SA-Allocation(Output)
  Input: Users' requests set T_req , VRP set S_vrp
  Output: Allocation Scheme, Total Cost
  Generate initial allocation scheme
  Calculate initial cost
  Calculate initial temperature, t_0
  t = t_0 ,max iteration = M ,current allocation ψ
  begin
    while i< M
        ψ ← generate neighbor allocation of ψ
        Calculate C_total^curr , and annealing p(C_total^curr)
        If C_total^curr < C_total^pre then
            Accept current allocation
            If C_total^pre - C_total^curr < ε then
                Break;
            Else Continue;
         Else if rand(0,1)< p(C_total^curr) then
              Accept current allocation;
         Else if exists request can't be satisfied, then:
              Schedule the request to next allocation
         Else continue
    End while
end.
```

We consider a heterogeneous network scenario that consists of macrocells, picocells, WiFi APs and a centralized controller for RAN. Similar to the model used in [17], a macro cell is surrounded by six co-channel macrocells. Pico cells and WiFi APs are

Table 1. Simulation parameters

Parameter	Value
Radius of macrocell coverage	500 m
Pico cell density	$(\pi * 100^2)^{-1}/m^2$
WiFi AP density	$(\pi * 50^2)^{-1}/m^2$
User density in Picocell and WiFi	$30 * (\pi * 100^2)^{-1}/m^2$
User density in Macrocell	$400 * (\pi * 500^2)^{-1}/m^2$
MeNB power	46 dBm
PeNB power	37 dBm
UE power	24 dBm
Channel bandwidth in cellular	10 MHz
WiFi mode	802.11n
Noise figure	9 dB
Path loss exponent	3.5

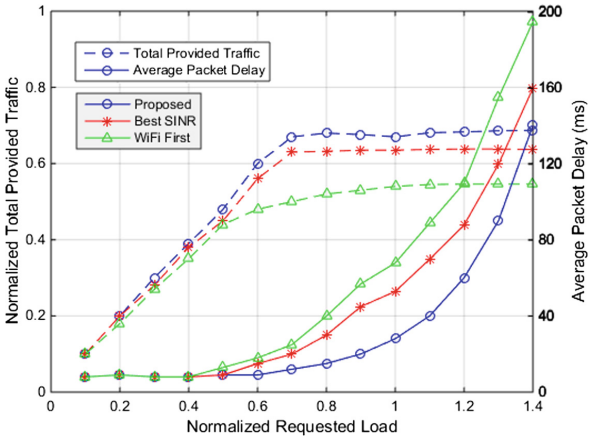


Fig. 3. Total provided traffic and average packet delay versus different requested load

deployed under the coverage of macro cell. Both of the positions of picocell and WiFi APs form an independent Poisson point process (PPP) Φ_p and Φ_w, with the density λ_p and λ_w accordingly. In practice, users are usually clustered around Pico cell and WiFi APs with a larger density. To simplify our simulation process, we suppose that the position of users in the coverage of Pico cell and WiFi APs form the same PPP Φ_u^{pw} with the intensity λ_u^{pw}, and users in the coverage of macro cell but out of the coverage form an PPP Φ_u^m with the intensity λ_u^m, where $\lambda_u^{pw} > \lambda_u^m$. Range expansion and inter-cell interference coordination are applied in our simulation. We set up parameters according to the work in [19], the main parameters used in the simulation are listed in Table 1.

We compare the elastic resource allocation in QvHran with other two approaches, one is the LTE-A with WiFi in Best SINR mode, which means that UE will always choose the access network with the best SINR, and the other one is LTE-A with WiFi

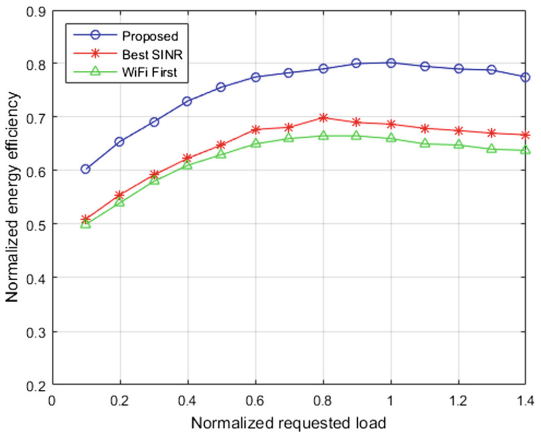


Fig. 4. Energy efficiency versus different requested load

in the WiFi-first mode. In our simulation, we compute network provided traffic average packet delay while the requested load of UEs is gradually increased, as shown in Fig. 3. Due to that resources from different RAN can be easily scheduled in QvHran, the proposed approach in QvHran can be more likely to satisfy users' requirement according to the status of different RAN. Compared with the other two approaches, QvHran can achieve a better network resource utilization, and the average packet delay is also very low. We also plot the energy efficiency with other two schemes in Fig. 4. It shows that our algorithm outperforms the other scheme about 12% in normalized energy efficiency.

5 Conclusion

In this paper, we have identified the challenges in the heterogeneous radio access networks, then proposed a QoE driven virtualization based architecture for heterogeneous radio access networks (QvHran). We have studied its hierarchy from both management and deployment aspect. The main contribution of the research is that we introduce a virtual network resource layer in the RAN management architecture. We abstract the heterogeneous radio resource into uniformed virtual resource particles. Upon this research, we have studied the key supporting technologies in QvHran and demonstrate the future RAN management, such as collaboration, load balance, mobility management, which could be equivalent to the procedure of virtual resource allocation. And finally a simulation is given to verify the performance of QvHran. In our future work, apart from studying the applications of QvHran, we also will pay more attention to quantitative analysis of QvHran. Besides, for centralized controlled architecture, the security is one of most the important factors, we will also study the security problems in our future works.

Acknowledgement. This work is supported by Beijing Advanced Innovation Center for Future Internet Technology, and Beijing Municipal Science and technology Commission research fund project No. D151100000115002.

References

1. Cisco. Cisco Visual Network Index: Global Mobile Data Traffic Forecast Update, 2014–2019 (2015)
2. Yong Sheng, S., et al.: Energy efficient heterogeneous cellular networks. Sel. Areas Commun. IEEE J. **31**(5), 840–850 (2013)
3. Peng, M., Yuan, L., Jiang, J., Li, J., Wang, C.: Heterogeneous Cloud Radio Access Networks: A New Perspective for Enhancing Spectral and Energy Efficiencies (2014)
4. ONF: Software-Defined Networking: The New Norm for Networks. ONF White Paper (2012)
5. Anjing, W., et al.: Network virtualization: technologies, perspectives, and frontiers. J. Lightwave Technol. **31**(4), 523–537 (2013)

6. Sheng, Z., et al.: CHORUS: a framework for scalable collaboration in heterogeneous networks with cognitive synergy. IEEE Wirel. Commun. **20**(4), 133–139 (2013)
7. Ali-Ahmad, H., et al.: CROWD: an SDN approach for DenseNets. In: 2013 Second European Workshop on Software Defined Networks (EWSDN) (2013)
8. Gudipati, A., et al.: SoftRAN: software defined radio access network. In: Proceedings of the Second ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking, ACM (2013)
9. Bansal, M., et al.: Openradio: a programmable wireless dataplane. In: ACM Proceedings of the First Workshop on Hot Topics in software Defined Networks (2012)
10. Wen, H., Tiwary, P.K., Le-Ngoc, T.: Current trends and perspectives in wireless virtualization. In: 2013 International Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT), IEEE (2013)
11. Kokku, R., et al.: NVS: a substrate for virtualizing wireless resources in cellular networks. IEEE/ACM Trans. Netw. **20**(5), 1333–1346 (2012)
12. Bernardos, C., et al.: An architecture for software defined wireless networking. IEEE Wirel. Commun. **21**(3), 52–61 (2014)
13. Wang, L., et al.: Open wireless network architecture in radio access network. In: 2013 IEEE 78th Vehicular Technology Conference (VTC Fall) (2013)
14. Fattah, H.: Analysis of the channel access mechanism in IEEE 802.11 wireless local area networks. In: IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, PacRim 2007 (2007)
15. Zhao, X., Lu, Z., Wang, L., Wen, X., Lei, T.: Service-oriented network performance evaluation framework based on LA-FAHP. J. China Univ. Posts Telecom **V22**(3), 74–83 (2015)
16. Xia, X., et al.: Blind video quality assessment using natural video spatio-temporal statistics. In: 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–6 (2014)
17. Guvenc, I., et al.: Capacity and fairness analysis of heterogeneous networks with range expansion and interference coordination. IEEE Commun. Lett. **15**(10), 1084–1087 (2011)
18. Misevičius, A.: A modified simulated annealing algorithm for the quadratic assignment problem. Informatica **14**(4), 497–514 (2003)
19. Banani, S.A., Eckford, A., Adve, R.: Analyzing dependent placements of small cells in a two-layer heterogeneous network with a rate coverage constraint. IEEE Trans. Veh. Technol. **65**, 9801–9816 (2016)

An ID-Based Anonymous Authentication Scheme for Distributed Mobile Cloud Computing

Tianyi Zhang and Fengtong Wen(✉)

School of Mathematical Science, University of Jinan, Shandong, China
ss_wenft@ujn.edu.cn

Abstract. Nowadays, the number of mobile users has rapidly increased. At the same time, the security problems in mobile cloud environment have a large attention from the user of mobile cloud. In this paper, a user anonymity and security authentication scheme for distributed mobile cloud computing is proposed. The proposed scheme is based on bilinear pairing cryptosystem and the theory of random number. It achieves mutual authentication, key exchange, user anonymity, and user intractability. It can resist reply attack, impersonation attack, and collusion attack with k-traitors. The proposed scheme has the function about cancel passive user for saving the storage space of server. By the security analysis, the proposed scheme is secure and efficient.

Keywords: Authentication scheme · Bilinear pairing · Mobile cloud computing services · User anonymity · User intractability

1 Introduction

As the development of network communications technology, cloud computing has aroused general concern, because of providing powerful computing capacity and huge storage space, and users can use the resource as their requirement, whenever and wherever possible [1].

Nowadays, mobile cloud computing [2–5] has become an important research field in mobile-oriented world, providing new supplements, consumption, and delivery models for IT services. In mobile cloud computing, mobile users can access computation results, resources, applications and services that are stored, implemented and deployed in cloud computing environments by using mobile devices through an insecure wireless local area network (WLAN) or 3G/4G telecommunication networks. When a mobile user who has registered wants to acquire resources by Web browser or a cloud computing application, it is necessary that the Web browser or the cloud computing application will mutually authenticate both the cloud service provider and the user. After authentication, the user can access the resources and available services from the cloud service provider. In real environment, an illegal access threats to security of system is existing. Cloud provider should support a secure authentication scheme for users using mobile devices [6–9].

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 401–409, 2017.
DOI: 10.1007/978-981-10-3969-0_44

In fact, the user's data is handled and stored by the service provider in mobile cloud environment. In other words, the data of user is lost for user's control. Therefore, the identity of user can not expose to the service provider in mobile cloud environment. It is easy to access services and information, which means that the services and information are easy to expose to everyone due to mistakes or a snag. All cloud computing have to provide a secure way, including personal identification, authorization, confidentiality and the availability of a certain level in order to cope with the security problem of private information exposed [10]. So, we proposed the authentication scheme based our understanding of mobile cloud environment security.

2 Related Works

Traditional authentication schemes are usually based on traditional public key cryptosystem. Traditional public key cryptosystems such as RSA require length key size and consume computation resources heavily. Hence, most of traditional authentication schemes are unsuitable for mobile devices, which have limited computing resources. Elliptic curve cryptosystem (ECC), which was first introduced by Koblitz [11] and Miller [12], offer the smallest key size per equivalent strength of any traditional public key cryptosystem, including RSA and Discrete Logarithm Problem (DLP). For example, a 256-bit ECC public key has the same security level as a 3072-bit RSA public key [13]. Such computational efficiency is beneficial for mobile devices.

Recently, bilinear pairing in an elliptic curve has been used in developing an ID-based cryptosystem [14–16]. Since then, several ID-based cryptosystems have been proposed. An ID-based cryptosystem is one kind of public key cryptosystems that can solve the high cost issue of public key management and authentication derived from traditional public key cryptosystems. In an ID-based cryptosystem, the identity of a user is used as the public key of this user; a user therefore does not spend extra computational cost to verify public keys of others, and no extra storage space in the user's device is required to store public keys of others and their corresponding certificates. Several studies have applied ID-based cryptosystems in cloud and grid computing environments. Lim and Robshaw [17, 18] first applied an ID-based cryptosystem to grid security in 2004, whereas in the same year, Mao [19] proposed an identity-based noninteractive authentication framework for grids. In 2009, Li et al. [20] developed a new ID-based authentication for cloud computing environment. However, the authentication protocol of Lin et al. does not provide user anonymity and intractability [21, 22]. 2015, Jia et al. [23] proposed an ID-based authentication scheme for mobile cloud computing environment. But the authentication scheme doesn't meet users anonymous, because the identity of user exposes to the cloud service provider. It is possible that a cloud service provider who isn't trusted will steal the information from users.

Since most authentication schemes based on ECC or bilinear pairing [24–30] are designed for client–server environment, they are not suitable to be directly adopted into distributed services environment in which multiple service providers compete with each other and offer various kinds of services. The most important issue is that a user needs to manage multiple private keys learned from each service provider. To resolve user key management issue, the simplest way is that all service providers share the same master

private key. However, if an adversary attacks one of the service providers successfully, he/she can learn this master private key and masquerade as any one of the service providers to cheat users. In addition, a malicious adversary, who has obtained the master private key from a service provider, can learn session keys established between another service provider and a user if the applied authentication scheme does not support perfect forward secrecy. After learning the session key, the malicious attacker can get sensitive information transmitted between the other service provider and a user. Hence, this simple approach is also unsuitable for distributed mobile cloud environments.

3 Protocol Framework

The proposed scheme assumes that the distributed mobile cloud service environment is supported by a trusted smart card generator (SCG) service. There are three roles in the proposed scheme: mobile users, distinct mobile cloud service provider and SCG. Their function as show in

1. Trusted smart card generator (SCG): generating public parameters, as well as all private keys for service providers and users.
2. Mobile users (U): requesting service, and authenticated by SP.
3. Mobile cloud service provider(SP): offering service to U, and authenticated by U.

The proposed scheme has four phases, including system setup phase, registration phase, authentication phase and cancel passive user phase.

1. System setup: This phase is implemented in SCG for generating master key and corresponding public key and public parameters.
2. Registration phase: U and SP send messages about respective identity to SCG. When SCG receives the messages, return private key to SP. If the U has not registered in SCG, SCG will store his/her information about his/her identity and return a message about registration success.
3. Authentication phase: U send a request about information of identity and identity of SP where he/she wants to obtain resource to SCG. After SCG verifies the identity of U, SCG sends a private key that is only applied to this access to U and turn the direction of access to SP. U and SP achieve mutual authentication and a session key. Next, U can obtain resource from SP.
4. Cancel passive user phase: SCG can cancel some passive users by deleting their information for verifying their identities that store in data base of SCG.

4 Proposed Scheme

The section will describe the detail of the proposed authentication scheme for distributed mobile cloud service environment. There are four aspects about the proposed authentication scheme, include system setup phase, registration phase, authentication phase and cancel passive user phase. Symbol notation used in the proposed scheme is shown in Table 1.

Table 1. Symbol notation in our proposed scheme

Symbol	Description
U_i, SP_j	A user i and service provider j
$H()$	One–way hash function
ID_i, PW_iS_i	Identity password private key of user i
Z, K_2, C_1 D_i	The authentic messages
a, b, c	Random number
e	A pairing function
$\|\|$	Concatenation operation
SCG	The smart card generator
ID_jS_j	Identity private key of SP
K_1	The session key

4.1 System Setup Phase

The phase is implemented in SCG for generating public parameters to make implementation of other phases smooth.

Let G_1 be a cyclic additive group generated by P, and let G_2 be a cyclic multiplicative group, where q is the prime order of G_1 and G_2. After choosing a pairing function $e : G_1 \times G_1 \rightarrow G_2$ and four collision-resistant hash functions: $H_1 : Z_q^* \rightarrow Z_q^*$, $H_2 : G_2 \rightarrow Z_q^*$, $H_3 : Z_q^* \rightarrow Z_q^*$, $H_4 : Z_q^* \rightarrow Z_q^*$. SCG chooses s that is the master private key of SCG and computes $P_{pub} = sP$, $e(P, P)$. Finally, SCG publishes $\{e, H_1, H_2, H_3, H_4, P, P_{pub}, e(P, P)\}$ as public parameters.

4.2 Registration Phase

Registration U_i: U_i choose ID_i and PW_i as his or her wish. U_i sends ID_i and $H_1(ID_i||PW_i)$ to SCG for registration by secure channel. Upon receiving identity ID_i and $H_1(ID_i||PW_i)$, SCG returns registration failure if ID_i has existed in data base. Otherwise SCG stores $(ID_i, H_1(ID_i||PW_i), T_i)$ to data base (DB), where T_i is time of registration success, and return a message to U_i in order to inform registration.

Registration SP_j: SP_j sends it's ID_j to SCG by secure channel. Upon receiving the ID_j from SP_j, SCG computes the private key belong to SP_j by master key s.

$$S_j = (s + H_1(ID_j))^{-1} P \tag{1}$$

Next, SCG sends S_j to SP_j. When SP_j receives S_j from SCG, SP_j stores S_j in secure memory that only the provider can access.

4.3 Authentication Phase

Step(1): U_i inputs ID_i and PW_i, computes $H_1(ID_i||PW_i)$ and ID_j of SP_j where U_i wants to obtain resource from to SCG.
Step(2): When SCG receives the message from U_i, if $H_1(ID_i||PW_i)$ exists in its date base, SCG find ID_i corresponding $H_1(ID_i||PW_i)$ and computes $N = H_1(ID_i||a)$, and $S_i = (s + H_1(ID_i||a))^{-1}P$, where a is a random number. Otherwise rejecting request from U_i. Then SCG sends S_i and N to U_i, and turns the direction of access to SP_j.
Step(3): After SP_j receives the message from SCG, SP_j computes

$$Z = e(P, P)^b \tag{2}$$

Where b is a random number. Then SP_j sends Z to U_i.
Step(4): When U_i receives Z, U_i chooses a random number c and computes

$$K_1 = H_2(Z^c) \tag{3}$$

$$K_2 = cP_{pub} + H_1(ID_j)cP \tag{4}$$

$$w = cP_{pub} + NcP \tag{5}$$

$$s_i = (c + H_3(N||Z||ID_j||w||K_1))^{-1}S_i \tag{6}$$

$$C_1 = K_1 \oplus (N||s_i||w) \tag{7}$$

U_i Sends (K_2, C_1) to SP_j
Step(5): When SP_j receives (K_2, C_1) from U_i, SP_j computes

$$K_1 = H_2(e(K_2, S_j)^b) \tag{8}$$

SP_j computes $K_1 \oplus C_1 = (N||s_i||w)$. Then SP_j computing

$$Q_i = P_{pub} + NP \tag{9}$$

Next, SP_j calculates the values

$$\begin{aligned}
&e\,(s_i, w + H_3(N||Z||ID_j||w||K_1)\,Q_i) \\
&= e\,([(c + H_3(N||Z||ID_j||w||K_1))(s+N)]^{-1}P, c(s+N)\,P + H_3(N||Z||ID_j||w||K_1)(s+N)P\,) \\
&= e\,([(c + H_3(N||Z||ID_j||w||K_1))(s+N)]^{-1}P, (c + H_3(N||Z||ID_j||w||K_1))(s+N)P\,) \\
&= e\,(P, P\,)
\end{aligned}$$

Then checks whether the values are equivalent with $e\,(P, P)$, as shown in the following equation

$$e\left(s_i, w + H_3(N||Z||ID_j||w||K_1)\, Q_i\right) = e\left(P, P\right)? \tag{10}$$

If these two values are equivalent, the validity of U_i is authenticated by SP_j. Otherwise SP_j rejects request.

Step(6): SP_j computes $D_i = H_4\left(K_1||Z||N||ID_j\right)$ and sends D_i to U_i.

Step(7): When U_i receives D_i from SP_j, U_i first computes $D_j^{'} = H_4\left(K_1||Z||N||ID_j\right)$.

And then checks whether the values of computed $D_i^{'}$ and the received D_i are same. If the values are equivalent, the validity of SP_j is authenticated. Note that K_1 is the session key shared between U_i and SP_j.

4.4 Cancel Passive User Phase

When a user sends a request about deleting account to SCG or a user doesn't use the service of SCG beyond ΔT compared T_i, SCG will delete $(ID_i), H_1(ID_i||PW_i)$ stored in data base for saving storage space. The user will be not authenticated by service provider.

5 Security Analysis

The section describes the security analysis about the proposed scheme. We assume that a attacker A can steal and tamper the message transporting in insecurity channel so that the attacker A can pretend to U or SP, and trace a user for obtaining the user's sensitive information. Next, we will describe the security requirement and attacks that are resisted in the proposed scheme.

5.1 Reply Attack

Having intercepted previous communications, an attacker can impersonate the legal user to access to the system. The attacker can replay the intercepted messages. In the proposed scheme, the attacker A can intercept Z, (K_2, C_1), D_i, and store them in a session. When the attacker A sends Z, (K_2, C_1), D_i to the corresponding entities in later session, A can't access successfully, because Z, (K_2, C_1), D_i are generated by random number, in other words, Z, (K_2, C_1), D_i are different in different session.

5.2 Impersonation Attack

An attacker attempts to modify intercepted communications to masquerade as the legal user to access the resources at a remote system. An attacker can also masquerade as the legal server to manipulate sensitive data of the legal users.

Impersonation User: If A want to pretend to U, he must know s_i and w. As U to SP authentication phase, it is based on $e\left(s_i, w + (H_3||N||Z||ID_j||w||K_1)Q_i\right) = e(P, P)$,

A has to obtain K_1 for retrieving s_i and w. But K_1 is generated by two random numbers, A can't obtain the K_1, in other words, A can't pretend to U.

Impersonation Severs: If A want to pretend to SP, he must know $D_i = H_4(K_1 ||Z||N||ID_j)$, K_1, Z, N are generated by random number which A doesn't know, so A can't retrieve D_i, in other words, A can't pretend to SP.

5.3 Collusion Attack with K-Traitors

Suppose that A obtains many S_i/S_j by registering, A can't obtain other U or SP's private key due to the K-CAA problem. Given P, sP, $\{e_1, e_2 \ldots, e_k \in Z_q^*\}$, and $\{(s+e_1)^{-1}P, (s+e_2)^{-1}P, \ldots, (s+e_k)^{-1}P\}$ for an integer $k, s \in Z_q^* P \in G_1$, it is difficult to compute $(s+e_0)^{-1}P$, where $e_0 \notin \{e_1, e_2, \ldots, e_k \in Z_q^*\}$ [23].

5.4 User Anonymity

It is very important to preserve the privacy of a user, because an adversary can eavesdrop the communication parties involved in the authentication process and can easily analyze the transaction being performed by user. In the proposed scheme, by a random number covering the identity of U, SP can't obtain the identity of U to achieve user anonymity.

5.5 Mutual Authentication

Mutual authentication should be provided between the user and remote systems. Not only can the server verify the legal users, but the users should be able to verify the legal server. In the proposed scheme, by $e(s_i, w + H_3(N||Z||ID_j||w||K_1) Q_i) = e(P, P)$, the scheme achieves the U to SP authentication. By $D_i = H_4(K_1 ||Z||N||ID_j)$, the scheme achieves the SP to U authentication.

6 Conclusion

This paper has proposed a new anonymous authentication scheme for distributed mobile cloud services environment. In the proposed, SP can't obtain the identity of user, because we make a random number covering the identity of user to guarantee the security of user's sensitive information. The proposed scheme supports mutual authentication, key exchange, user anonymity, and user intractability. Security analyses have shown that the proposed scheme can resist reply attack, impersonation attack, and collusion attack with k-traitors. By the security analysis, the proposed scheme is secure and efficient.

Acknowledgments. This work is supported by Shandong Provincial Natural Science Foundation, China (NO. ZR2013FM009).

References

1. ABI Research Report, Mobile Cloud Applications. http://www.Abiresearch.com/
2. Le, G., Xu, K., Song, M., Song, J.: A survey on research on mobile cloud computing. In: International Conference on Computer and Information Science, pp. 387–392 (2011)
3. Qiu, X.F., Liu, J.W., Zhao, P.C.: Secure cloud computing architecture on mobile internet. In: International Conference on AIMSEC, pp. 619–622 (2011)
4. Fernando, N., Loke, S.W., Rahayu, W.: Mobile cloud computing: a survey. Future Gen. Comput. Sys. **29**(1), 84–106 (2013)
5. Song, W.G., Su, X.L.: Review of mobile cloud computing. In: IEEE 3rd ICCSN, pp. 1–4 (2011)
6. Urien, P., Marie, E., Kiennert, C.: An innovative solution for cloud computing authentication: grids of EAP-TLS smart cards. In: International Conference on Digital Telecommunications, pp. 22–27 (2010)
7. Ahn, H., Chang, H., Jang, C., Choi, E.: User authentication platform using provisioning in cloud computing environment. In: Kim, T., Adeli, H., Robles, R.J., Balitanas, M. (eds.) Advanced Communication and Networking. Communications in Computer and Information Service, vol. 199, pp. 132–138. Springer, Heidelberg (2011)
8. Chang, H., Choi, E.: User authentication in cloud computing. UCMACCIS **151**, 338–342 (2011)
9. Tsai, J.L., Lo, N.W., Wu, T.C.: Secure delegation-based authentication protocol for wireless roaming service. IEEE Commun. Lett. **16**(7), 1100–1102 (2012)
10. Hyeonseung, K., Chunsik, P.: Cloud computing and personal authentication service. J. Korea Inst. Inf. Secur. Cryptol. **20**(2), 11–19 (2010)
11. Koblitz, N.: Elliptic curve cryptosystems. Math. Comput. **48**(177), 203–209 (1987)
12. Miller, V.S.: Use of elliptic curves in cryptography. In: Williams, H.C. (ed.) CRYPTO 1985. LNCS, vol. 218, pp. 417–426. Springer, Heidelberg (1986). doi:10.1007/3-540-39799-X_31
13. Recommendation for key management—Part1: General. Gaithersburg, MD, USA, August, pp. 800–57. Special Publication (2005)
14. Boneh, D., Franklin, M.: Identity-based encryption from the Weil pairing. In: Kilian, J. (ed.) CRYPTO 2001. LNCS, vol. 2139, pp. 213–229. Springer, Heidelberg (2001). doi:10.1007/3-540-44647-8_13
15. Choon, J.C., Hee Cheon, J.: An identity-based signature from gap Diffie-Hellman groups. In: Desmedt, Y.G. (ed.) PKC 2003. LNCS, vol. 2567, pp. 18–30. Springer, Heidelberg (2003). doi:10.1007/3-540-36288-6_2
16. Du, H.Z., Wen, Q.Y.: An efficient identity-based short signature Scheme from bilinear pairings. In: Proceedings of International Conference CIS, pp. 725–729 (2007)
17. Lim, H.W., Robshaw, M.J.B.: On ientity-based cryptography and grid computing. In: Bubak, M., Albada, G.D., Sloot, Peter, M.A., Dongarra, J. (eds.) ICCS 2004. LNCS, vol. 3036, pp. 474–477. Springer, Heidelberg (2004). doi:10.1007/978-3-540-24685-5_69
18. Lim, H.W., Robshaw, M.J.B.: A dynamic key infrastructure for grid. In: Sloot, P.M.A., Hoekstra, A.G., Priol, T., Reinefeld, A., Bubak, M. (eds.) EGC 2005. LNCS, vol. 3470, pp. 255–264. Springer, Heidelberg (2005). doi:10.1007/11508380_27
19. Mao, W.: An identity-based non-interactive authentication framework for computational grids. Technical report HPL-2004-96, HP Labs, Palo Alto, CA, USA (2004)

20. Li, H., Dai, Y., Tian, L., Yang, H.: Identity-based authentication for cloud computing. In: Jaatun, M.G., Zhao, G., Rong, C. (eds.) CloudCom 2009. LNCS, vol. 5931, pp. 157–166. Springer, Heidelberg (2009). doi:10.1007/978-3-642-10665-1_14
21. Hughes, V.S.: Information hiding, anonymity and privacy a modular approach. J. Comput. Security **12**(1), 3–36 (2004)
22. Tsai, J.L., Lo, N.W., Wu, T.C.: Novel anonymous authentication scheme using smart cards. IEEE Trans. Ind. Inform. **9**(4), 2004–2013 (2013)
23. Tsai, J.L., Lo, N.W.: A privacy-aware authentication scheme for distributed mobile cloud computing services. Syst. J. **9**(3), 805–815 (2015)
24. Das, M.L., Saxena, A., Gulati, V.P., Phafstak, D.B.: A novel remote user authentication scheme using bilinear pairings. Comput. Secur. **25**(3), 184–189 (2006)
25. Chou, J.S., Chen, Y., Lin, J.Y.: Improvement of Das et al.'s remote user authentication scheme. Cryptology ePrint Archive (2005)
26. Goriparthia, T., Das, M.L., Saxena, A.: An improved bilinear pairing based remote user authentication scheme. Comput. Std. Interfaces **31**(1), 181–185 (2009)
27. Khan Pathan, A.S., Hong, C.S., Hee, K.: Bilinear-pairing-based remote user authentication schemes using smart cards. In Proceedings of 3rd International Conference Ubiquitous Information Management Communication, pp. 356–361 (2009)
28. Chen, T.H., Yeh, H.L., Shih, W.K.: An advanced ECC dynamic ID-based remote mutual authentication scheme for cloud computing. In: International Conference on Multimedia Ubiquitous Engineering, pp. 155–159 (2011)
29. Wang, D., Mei, Y., Ma, C., Cui, Z.: Comments on an advanced dynamic ID-based authentication scheme for cloud computing. In: Wang, F.L., Lei, J., Gong, Z., Luo, X. (eds.) WISM 2012. LNCS, vol. 7529, pp. 246–253. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33469-6_34
30. Sun, H., Wen, Q., Zhang, H., Jin, Z.: A novel remote user authentication and key agreement scheme for mobile client–server environment. Appl. Math. Inf. Sci. **7**(4), 1365–1374 (2013)

QKDFlow: QKD Based Secure Communication Towards the OpenFlow Interface in SDN

Yan Peng[1], Chunqing Wu[1], Baokang Zhao[1,2(✉)], Wanrong Yu[1], Bo Liu[1], and Shasha Qiao[3]

[1] College of Computer, National University of Defense Technology, Changsha 410073, China
{pengyanl5, wuchunqing, bkzhao, wlyu}@nudt.edu.cn, Boliuchang@sina.com
[2] Guangxi Cooperative Innovation Center of Cloud Computing and Big Data, Guilin University of Electronic Technology, Guilin 541004, China
[3] PLA 75833 UNIT, Pudong, China
ssqiao@qq.com

Abstract. Software Defined Networks (SDN) decouples control plane and data plane, which simplifies network management. However, there are still some security threats which limit the large scale deployment of SDN. In this paper, we present a solution which integrates Quantum Key Distribution (QKD) technology with SDN in the southbound interface to fulfill secure communication between controller and switches. Rather than merely employ Transport Level Security (TLS) protocol in OpenFlow standard, the proposed scheme can prevent the Man-In-The-Middle (MITM) attack.

Keywords: SDN · QKD · TLS · Openflow · Man-In-The-Middle attack

1 Introduction

Software Defined Networking (SDN) [1], originates from the research project of Clean Slate in Stanford University in 2006. The conception of OpenFlow [2] based SDN is first proposed in 2008 by Professor Nick McKeown in Stanford University and has been listed as one of the ten breakthrough technologies which may change the world by MIT in 2009 [3].

The architecture of SDN is depicted in Fig. 1. SDN decouples control plane and data plane in the network architecture, which simplifies network management to a great extent. In the control level, the programmable controller masters the whole status of the network, which is convenient for researchers to configure network settings and deploy new protocols. In the data level, switches only provide data forwarding functions, which can process the packets fast. The communication between these two layers conform to the open uniform interface (e.g. OpenFlow). The OpenFlow channel is the interface that connects each OpenFlow Logical Switch to an OpenFlow controller. Through this interface, the controller configures and manages the switch, receives

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 410–415, 2017.
DOI: 10.1007/978-981-10-3969-0_45

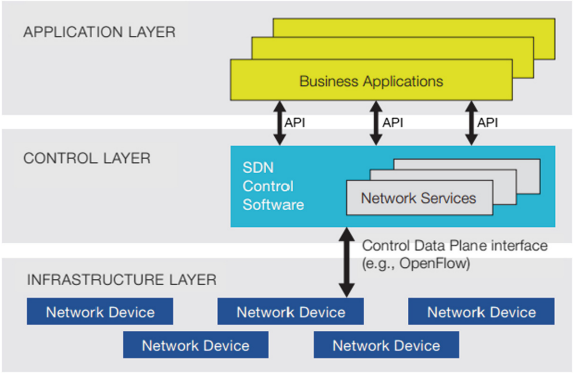


Fig. 1. Architecture of SDN

events from the switch, and sends packets out the switch [4]. Therefore, SDN could effectively reduce equipment load, decreasing operating costs, and assist operators to manage the network better.

2 TLS Protocol

In the OpenFlow standard, the Transport Layer Security [5] (TLS) protocol is employed to safeguard the single network connection between the controller and the switch. However, it is an optional choice from version 1.3.0. Furthermore, the TLS protocol itself has some loopholes, which may lead to Man-In-The-Middle (MITM) attack [6] and other attacks such as Bleichenbacher's attack [7], Ray and Dispensa's renegotiation attack [8] and stripping attacks [9–11]. Florian et al. have proposed a countermeasure which provides renegotiation security [12]. Manik et al. have proposed a solution to resist the MITM attack [6].

The TLS protocol consists of four subprotocols [5] – Handshake, ChangeCipherSpec, Record, and Alert subprotocols. The process of TLS handshake sub protocol is depicted in Fig. 2.

TLS protocol works with the following steps:

(1) Client→Server: ClientHello. Provide the version of TLS, CipherSuite and client random.
(2) Server→Client: ServerHello. Provide the version of TLS, selected CipherSuite, server random and certificate.
(3) Server→Client: ServerHelloDone. Provide some optional parameters.
(4) Client→Server: ClientKeyExchange. Provide the pre-master key which encrypted by the public key of the server.
(5) Client→Server: ChangeCipherSpec.
(6) Client→Server: Finished. Provide the encrypted Message Authentication Code of the message (1)–(5).

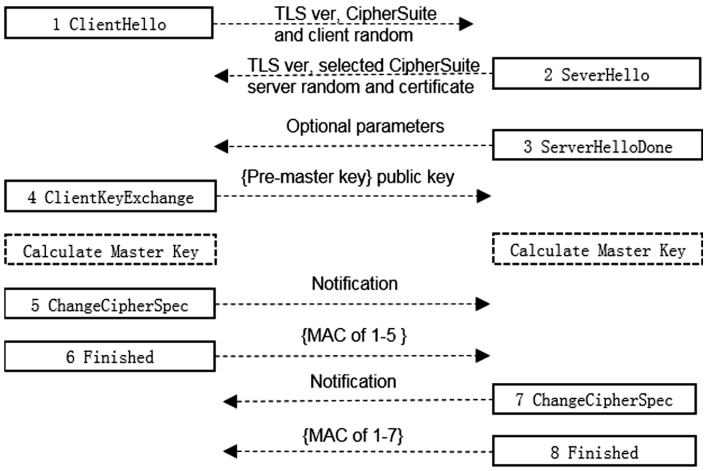


Fig. 2. Procedure of the TLS handshake protocol

(7) Server→Client: ChangeCipherSpec
(8) Server→Client: Finished. Provide the encrypted Message Authentication Code of the message (1)–(7).

There are some optional parameters the client and server can choose whether to set in the ServerHelloDone message. For example, user authentication, which can obviously improve the security of the communication. However, the security of the certificate authority is a problem. For example, a certificate authority in Holland named DigiNotar has been hacked in 2011 [13], and then, the hacker has issued some fake certificates for several popularity websites, such as Google, twitter and so on.

Moreover, the option of user authentication is not used as default. Therefore, the TLS is under more security threats.

Manik et al. have discussed MITM attacks on TLS-enabled application, and outlines existing solutions for MITM threats on TLS-enabled transactions [6]. They also proposed a soft-token based solution to mitigate the MITM threat on SSL/TLS enabled web applications. However, it does not suit the SDN environment well for the user have to participate the process.

In this paper, we proposed a security communication scheme between controller and switch in SDN, named as QKDFlow, which integrate the QKD technology with SDN.

3 The Architecture of QKDFlow

The speed of the QKD final key generation has been refreshed continuously recent years. Toshiba Corporation in 2014 has polished the speed to about 25.8 Gbits per day and sustained for 34 days [14]. The transmission distance has been refreshed to 404 km [15]. With the higher speed and the longer distance development of QKD, it will be

more practical in the real world. In this paper, we proposed a novel security communication scheme for SDN.

3.1 The Structure of QKDFlow

The structure of the QKDFlow is depicted is Fig. 3. In Fig. 3, the C-hub means the classical hub and the Q-hub means the beam-steering optical switch. The QKDFlow can time-sharing the quantum channel between the Q-hub and the controller to transfer the photons between the controller and the switches, and the classical channel can time-sharing the classical channel between the C-hub and the controller to transfer the classical data packets.

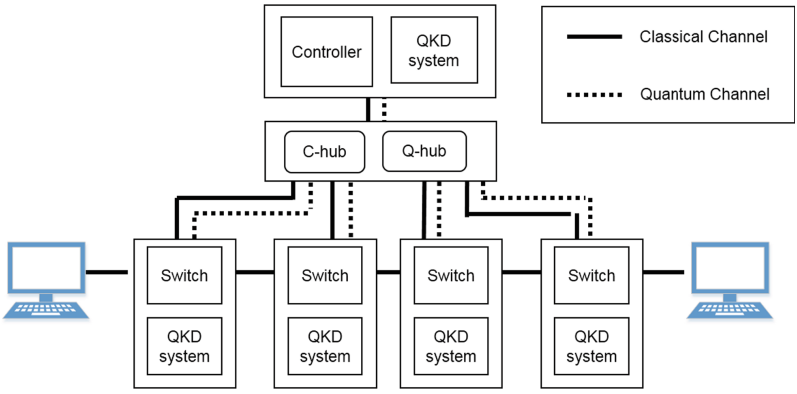


Fig. 3. Structure of QKDFlow

In the QKDFlow, mutual information of QKD post-processing is time shared with control information, and the priority of the QKD post-processing packets is higher than the control packets to ensure the key could be generated faster.

The procedure of QKDFlow is depicted as Fig. 4, it works with the following steps.

(1) The QKD systems of the switches and the controller all start, each QKD system of the switch connect to the QKD system of the controller time-shared under the beam-steering optical switch's scheduling and generate keys.
(2) The switch and the controller start, and the switch and the controller use the Handshake sub-protocol [5] to establish the connections, meanwhile, the QKD systems continue to work.
(3) The switches and the controller check the final key generated by the QKD systems. If the key between the controller and the switch is the same and do not find there are any eavesdropper and the amount of the key is enough (e.g., 100 MB), then go to step four; Else if the key is not enough, then wait; Else, give up and quit.
(4) The controller and the switches use the quantum key and One Time Pad scheme to encrypt their communication.

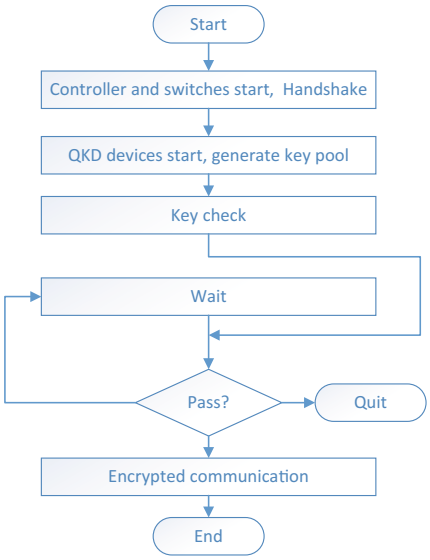


Fig. 4. Procedure of the QKDFlow

3.2 Security Analysis

We present some qualitative analysis of the QKDFlow.

First, the QKD systems is controlled by the administrator. Therefore, only the legitimate parties can get the quantum key of the switch and use it to connect to controller. This has prevent the forgery attacks. Second, the rules of quantum mechanics protect the key from been eavesdropped. If the two parties find there is an eavesdropper, the connection will be closed. This has prevent the MITM attacks. Third, the quantum key is generated synchronous in controller and switches, and is only used for one time, this has prevent the replay attacks.

4 Conclusion

In this paper, we introduced the TLS protocol employed in SDN. Towards the security threats of the OpenFlow interface of the southbound in SDN, we mainly proposed a secure communication scheme QKDFlow which integrate QKD technology with SDN and we have analyzed the security in a qualitative way. There are still some problems in QKDFlow to be solved. For example, the performance of the QKD system and the robustness of the whole systems. In the future work, we will focus on these problems to make this structure more practical and more security.

Acknowledgements. This work was supported by NSFC No. 61202488, and Guangxi Cooperative Innovation Center of cloud computing and Big Data (No. YD16505).

References

1. Open Network Foundation: Software-defined networking: the new norm for networks. ONF White Paper (2012)
2. McKeown, N., Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, L., Rexford, J., Shenker, S., Turner, J.: OpenFlow: enabling innovation in campus networks. ACM SIGCOMM Comput. Commun. Rev. **38**, 69–74 (2008)
3. MIT Technology Review: 10 breakthrough technologies, TR10: software-defined networking. http://www2.technologyreview.com/article/412194/tr10-software-defined-networking/ (2009)
4. ONF: OpenFlow Switch Specification V1.5.1
5. Dierks T., Rescorla, E.: Transport Layer Security Protocol. Network Working Group, RFC 5246 (2008)
6. Das, M.L., Samdaria, N.: On the security of SSL/TLS-enabled applications. Appl. Comput. Inform. **10**, 68–81 (2014)
7. Bleichenbacher, D.: Chosen ciphertext attacks against protocols based on the RSA encryption standard PKCS #1. In: Krawczyk, H. (ed.) CRYPTO 1998. LNCS, vol. 1462, pp. 1–12. Springer, Heidelberg (1998). doi:10.1007/BFb0055716
8. Ray, M., Dispensa, S.: Renegotiating TLS (2009). http://extendedsubset.com/Renegotiating_TLS.pdf
9. Marlinspike, M.: New tricks for defeating SSL in practice. In: BlackHat (2009)
10. Shin, D., Lopes, R: An empirical study of visual security cues to prevent the SSL stripping attack. In: Proceedings of the Computer Security Applications Conference (ACSAC 2011), pp. 287–296 (2011)
11. Zhao, S., Wang, D., Zhao, S., Yang, W., Ma, C.: Cookie-proxy: a scheme to prevent SSL strip attack. In: Chim, T.W., Yuen, T.H. (eds.) ICICS 2012. LNCS, vol. 7618, pp. 365–372. Springer, Heidelberg (2012). doi:10.1007/978-3-642-34129-8_34
12. Giesen, F., Kohlar, F., Stebila, D.: On the security of TLS renegotiation. In: Proceedings of the 20th ACM Conference on Computer and Communications Security (CCS) 2013 (2013)
13. Zetter, K.: DigiNotar files for bankruptcy in wake of devastating hack. Wired Mag. (2011)
14. Sasaki, M., Fujiwara, M., Ishizuka, H., Klaus, W., Wakui, K., Takeoka, M., Miki, S., Yamashita, T., Wang, Z., Tanaka, A.: Field test of quantum key distribution in the Tokyo QKD Network. Opt. Express **19**, 10387–10409 (2011)
15. Yin, H.-L., Chen, T.-Y., Yu, Z.-W., Liu, H., You, L.-X., Zhou, Y.-H., Chen, S.-J., Mao, Y., Huang, M.-Q., Zhang, W.-J.: Measurement device independent quantum key distribution over 404 km optical fibre. arXiv preprint arXiv:1606.06821 (2016)

Location System Design Based on Weighted RSSI for High-Speed Railway Landslide Monitoring

Bo Yang(✉), Yongqiang Zhang, Jifu Yu, Xingxia Wang, and Xinchun Jia

School of Mathematical Sciences, Shanxi University, Taiyuan 030006, China
aosmiths@126.com, zyq904914185@163.com, {1084740669, 1902758021}@qq.com, xchjia@sxu.edu.cn

Abstract. There exist such problems as high cost, difficult construction and low automation degree of existing landslide monitoring system. To treat these problems, a high-speed railway landslide monitoring system based on Zigbee wireless network technology is designed. In this system, high-speed railway landslide information is obtained by wireless sensor network location technology. Its monitoring data is transmitted by rail side room to remote monitoring center. According to the characteristics of the landslide, location coordinate system is improved. Plan of the system structure and program based on weighted Received Signal Strength Indicator (RSSI) location algorithm is introduced. The simulation experimental results show that the monitoring system can monitor effectively high-speed railway landslide.

Keywords: Landslide monitoring · Wireless sensor network (WSN) · Zigbee · Received Signal Strength Indicator (RSSI) · Location algorithm

1 Introduction

There are many railways that are in complicated geology and frequent scourge mountain areas in China, and the safety of railways is a problem that the operating administration has to face. Through the high-speed railway landslide monitoring, we can know the feature information of the change of landslide mass timely so that we can provide scientific basis to the safe operation of high-speed railway. Up to now, the technologies used to landslide monitoring are mainly include hand gaging [1], GPS positioning system, the disaster sensor detection system [3]. Although the detection technologies that mentioned above can monitor the dynamic information of landslide with efficiency in different ways, there exist such problems as high cost, difficult construction and low automation degree.

The WSN based on ZigBee has such advantages as self-organization, low power dissipation, low cost, quicker deploy ability and strong expansibility, which are very appropriate to structure network and set up infrastructure which is used to gather and transmit information in areas that are along the high-speed railway landslide environment. Meanwhile, combined with the location technology of wireless sensor network,

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 416–425, 2017.
DOI: 10.1007/978-981-10-3969-0_46

we can attain the information of the ground movement in the slippery area with efficiency, and response the movement of the surface of the landslide timely, which are very important to the detection of high-speed railway landslide and the safe operation of railway.

As an important technology in wireless sensor network, wireless sensor network positioning technology is widely used in landslide detection. According to the need to measure the distance between nodes, we divide the positioning algorithm into distance based and not. Up to now, the distance measuring technology includes RSSI [4], TOA [5], AOA [6], TDOA [7] and so on. And the positioning that don't need to measure distance only rely on the network connectivity of network and information to fix position, there are centroid algorithm [8], DV-Hop algorithm [9], APIT algorithm [10] and so on.

The article will make use of the centroid algorithm based on the RSSI algorithm which needs to measure the distance among nodes to accomplish the exact detection among unknown nodes and realize the detection of the movement of the surface of the landslide. We can use the Zigbee technology to achieve the wireless sensor positioning, which can replace the expensive geological sensor to gather the movement information of the surface of the landslide.

2 The Key Technology of the Monitoring System

2.1 Dynamic RSSI Ranging Algorithm

The fading characteristic of wireless sensor network's signal path obeys the log-normal distribution and can be influenced by background noise, multipath fading and dispersion of beacon nodes. So free space radio propagation path loss model combined with logarithmic normal distribution model, it will be more accurate to calculate path loss.

Free space radio propagation path loss model:

$$Loss = 32.4 + 10n \lg d + 10 \lg f \tag{1}$$

Loss represents for the path loss when signal transmission distance is d. The parameter d represents for the distance with beacon nodes. The parameter n represents for signal loss coefficient (often take $2 \sim 4$). The parameter f represents for the transmission signal frequency.

Logarithmic normal distribution model:

$$PL(d) = PL(d_0) + 10 \lg(d/d_0) + X_\sigma \tag{2}$$

The parameter PL(d) represents for the path loss when signal transmission distance is d. The parameter d_0 represents for reference range. The parameter n represents for signal loss coefficient (often take $2 \sim 4$), The parameter X_σ represents for average value of 0 Gaussian random variables, and it's the standard deviation between $4 \sim 10$.

Signal strength that unknown nodes received from beacon nodes:

$$RSSI(d) = P_t - PL(d) \tag{3}$$

RSSI represents for signal strength indicator. The parameters Pt represents for transmitting power, and Pl(d) represents for path loss.

The parameter d_0 = 1 m, according to the formula (1), we can know the Loss, which is $Pl(d_0)$. Combined with the formula (2) and (3), we can get the RSSI distance calculation formula which is simplified by IEEE802.15.4:

$$RSSI = \begin{cases} Pt - 40.2 - 10 \times 21g(d) d \leq 8 \\ Pt - 58.5 - 10 \times 3.31g(d) d > 8 \end{cases} \tag{4}$$

In the formula (4) signal loss coefficient is constant, but in the high-speed railway landslide environment, nodes are at different position, the outside environment condition is different and not all the signal loss coefficients are the same, so it will have a relatively big range error when using the formula (4). In order to solve this problem, firstly we can through the adjacent beacon nodes to figure out the signal loss coefficient in the current region.

Around the unknown nodes, we can choose (m + 1) data, which is received from the beacon node, and we can know that:

$$RSSI = \frac{\sum_{i=1}^{m} RSSI_i}{m} \tag{5}$$

Where, $RSSI_i$ represents the unknown node received from the beacon node.

Combined with the formula (2), (3) and (5), we can get the path loss index calculation formula:

$$n = \frac{(Pt + Pl(d_0) - X_\sigma)m - \sum_{i=1}^{m} RSSI_i}{10m \times Lg(d/d_0)} \tag{6}$$

Combined with the formula (6) and (4), we can get the RSSI distance measurement formula based on dynamic path loss coefficient:

$$d = \begin{cases} 10^{\frac{Pt-40.2-RSSI(d)}{10n}} d \leq 8 \\ 10^{\frac{Pt-58.5-RSSI(d)}{10n}} d > 8 \end{cases} \tag{7}$$

2.2 The Weighted Centroid Localization Algorithm

The weighted centroid localization algorithm is mainly aimed to add weight concerning distance to each beacon node. Using RSSI distance measurement as weighting factor can reflect the influence degree which is from each beacon nodes to the position of the centroid. We can use RSSI distance weighting and centroid algorithm to locate the nodes which need to be located.

When using RSSI distance measurement, in order to make the positioning more accuracy, we need get more exact RSSI value. But RSSI can be influenced by environmental interference which can result in data which is absent from the real condition, so we use Gaussian fitting method to filter wrong data and prove the accuracy of the distance measurement. The Gaussian fitting fiction is the following:

$$y = y_0 + \frac{A}{\omega\sqrt{\pi/2}} e^{\frac{-2(x-x_c)^2}{\omega^2}} \tag{8}$$

where, $x_c = \sum_{i=1}^{k} RSSI_i$, $\omega = \sqrt{\frac{\sum_{i=1}^{k}(RSSI_i - x_c)^2}{k-1}}$

In the formula, y_0, A are of undetermined coefficients (can be figured out by the relationship between the position of beacon nodes and RSSI), k represents for the number of the beacon nodes which had received the data. According to the literature [11], we can know that $0.5 \leq y \leq 1$, it's effective sampling data. We can save their RSSI value and average them. Then we can get the RSSI value under that condition.

It is combined with Gaussian fitting and RSSI to measure the distance. Using the weight thought, the centroid positioning algorithm which is after RSSI ranging weight weighted can be expressed as:

$$P_i(x,y) = \frac{\sum_{j=1}^{N}\left[\frac{1}{d_{ij}^n} B_j(x,y)\right]}{\sum_{j=1}^{N}\frac{1}{d_{ij}^n}} \tag{9}$$

In the formula: $P_i(x,y)$ represents for the estimated location of the pending nodes, $B_j(x,y)$ represents for the coordinate of beacon node j, n represents for weight coefficient (often taken between 1 and 4), d_{ij} represents for the distance between the beacon node j and unknown node i, N represents for the number of the beacon nodes.

2.3 WSN Positioning Monitoring Under High-Speed Railway Landslide Environment

In the high-speed railway landslide monitoring system, we use Beidou navigation system positioning module to locate the position of the beacon nodes and the position information is longitude and latitude coordinates. But the surface of the landside has a certain angle with the ground plane, merely depends on the longitude and latitude coordinates to locate will neglect the vertical displacement of the landslide. So, this system will firstly categorize the landslide slope as a plane, then make the longitude and latitude coordinate reflect to that plane, then we can get the landslide coordinate under that plane to avoid the error that caused by the neglect of the vertical displacement. In the landslide site, WSN network nodes are composed of positioning node, beacon node, routing node and gathering node. When the positioning nodes get position message, they through the routing nodes transmitting the message to the gathering nodes and then through the rail side room host of communication transmitting concentrated the gathered

landslide position message to the remote monitoring center. Through that procedure we can achieve the remote monitoring along the high-speed railway landslide surface displacement.

3 The Design of the Monitoing System

3.1 System Design

The goal of this system design is to achieve the remote monitoring of the mountain which exists landslide hazard along the high-speed railway, and is based on wireless sensor network positioning technology to get the information, and use the infrastructure that WSN and high-speed railway now have as the way of access to information, and improve the accuracy of the information gathering and strength the dealing ability of the landslide accident and achieve the landslide remote automatic detection and warning.

This system uses Zigbee wireless sensor network to monitor the landslide, and the thought of this system design is the following:

We place the wireless sensor network based on Zigbee in the high-speed railway landslide area, and nodes are composed of beacon nodes whose position are known and pending a node. Inside the pending node it operates the positioning algorithm to get its location information, every positioning node transmits the location information to gathering node through routing node, and then transmits the data to remote monitoring center through the rail side room. The remote monitoring center estimates whether there is a landslide through comparing and analyzing the displacement change of the surface of the landslide, and with that it will provide warning information to the operation of high-speed railway and achieve the remote monitoring of high-speed railway landslide.

3.2 The System Architecture

According to the specific requirements of landslide monitoring system positioning along the high-speed railway and data remote transmission and combined with architecture design method of traditional information gathering system, the design of this system is as the Fig. 1:

The data collection layer, this layer is in the bottom of this monitoring system architecture, is composed of the Zigbee wireless sensor network on the landslide scene, is mainly in charge of the gathering of the displacement information of landslide, and achieves the transmission on a regular basis of the landslide location information and quick transmission of abnormal data, finishes the gathering of the monitoring data and the function of WSN transmission.

The data transmission layer: the location information transmitted by WSN needs transmitting to the remote monitoring center, so using the communication host in the rail side room along the high-speed railway to transmit the location information of landslide to the remote monitoring center to achieve the reliable and remote transmission of monitoring data.

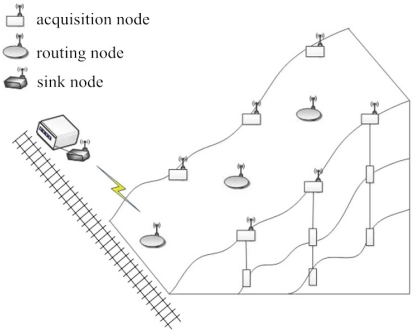


Fig. 1. The system architecture

The data stored layer: this layer is composed of the database of the remote monitoring center and is in charge of the storage and update of the data that has gathered. And it will store the relative the alarm information and history information to provide data support for the warning of landslide monitoring.

The function implementation layer: This layer uses the remote monitoring software to treat the positioning information that stored in the data stored layer comprehensively and analyzes the displacement situation of the surface of the landslide along the high-speed railway to judge whether there is a landslide or there exists the hidden trouble to a landslide and then spreads warning information to the relative units of the high-speed railway to achieve the safety operation of high-speed railway.

3.3 The System Program Design

The program design of node: Firstly, according to the actual situation of the landslide surface we can set a displacement alert threshold. Then after the positioning nodes are placed, they will operate a pre-positioning to make sure their location in order to compare if there is a abnormal situation someday. After beacon nodes and positioning nodes get and calculate their position information, they will compare it with the pre-positioning data. If the position information is within the displacement alert threshold, it will send its own position information to the remote data to count and analyze period. If the position is out of the value, the beacon node will get the position information of satellite to locate again immediately, and the positioning node will also locate again. After done 3 times, if all the position information is out of the alert threshold, it shows that there is a relative big displacement movement on the surface of the landslide and it needs to send a warning message instantly. The nodes will transmit its present position information to the remote monitoring center immediately to make a quick decision to avoid the hidden trouble that from landslide to high-speed railway. The flow chart of the node is the Fig. 2:

The program design of remote monitoring is show in Fig. 3: firstly, the remote monitoring machine will receive the positioning information on the landslide scene through the wire communication network in the rail side room along the high-speed

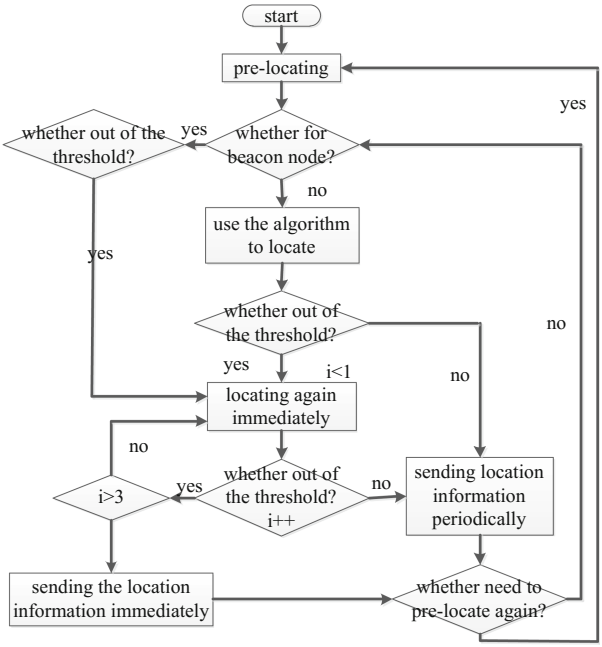


Fig. 2. The flow chart of the node

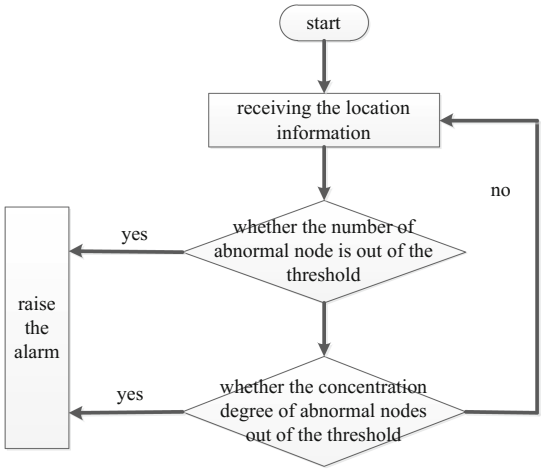


Fig. 3. The flow chart of the remote monitoring

railway and store it in the database. And according to the actual situation of the high-speed railway landslide, it will set the number of threshold and concentration threshold of abnormal nodes, and the monitoring machine will judge whether the number of the abnormal nodes are out of the standard threshold through comparing the

data that the database already has. If it is out of the threshold, the monitoring machine will alarm immediately to warm the relevant staffs to take relevant measures. If the number of the abnormal nodes are within the threshold, but the concentration is beyond the threshold, it shows that part of the landslide is abnormal and will alarm immediately to remind the relevant stuff to investigate the landslide in case for a bigger disaster. If the number of the abnormal nodes received is within the threshold, the monitoring machine will continue to receive the positioning information and update the database.

4 The Experimental Results and Analysis

In order to verify the positioning effect of the centroid algorithm based on RSSI ranging weights, in the first place, we will choose two nodes on the landslide, and we will test the RSSI value 100 times every 0.5 m in 30 m; in the second place we will use the Gaussian fitting and the mean fitting to deal with the RSSI data and use the formula (7) to calculate the distance to compare the error. Therefore, the results can be obtained by these place, as shown in following Fig. 4, compared to the mean fitting, the RSSI ranging effect of the Gaussian fitting is obviously better than the mean fitting in the whole.

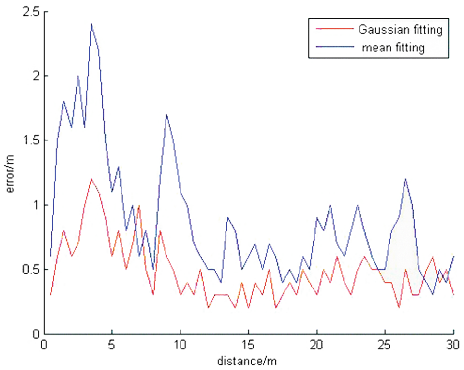


Fig. 4. The error comparison between the Gaussian fitting and the mean fitting

According to the demand of the monitoring system design, there we will build an experimental landslide model. We will set five observation points whose coordinates are known. WSN is composed of five beacon nodes and two pending nodes. The beacon nodes are fixed in the landslide model. The pending nodes will move to the next observation point every little time. In doing so, we can simulate the deformation process of the surface of the landslide and calculate the positioning error through the serial port to connect with the computer. And the positioning error is given in the following figure.

As the Fig. 5, the biggest average position error of the positioning node is 1.85 m and is in the observation point B, and the smallest average position error of the

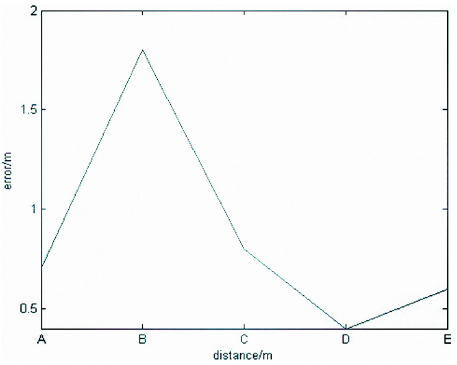


Fig. 5. The error of the actual and measured

positioning node is 0.4 and is in the observation point D. The positioning error in the whole is in a suitable range, which indicates that the system has a relative high accuracy and can monitor the displacement situation of the landslide effectively.

5 Conclusion

This article has designed a high-speed railway landslide remote monitoring system based on RSSI ranging weights positioning algorithm. Combined with the characteristic of the high-speed railway landslide to improve the positioning coordinate system and achieve the remote transmission through the infrastructure of the high-speed railway. We will achieve the remote all monitoring and disaster warning through the relevant software in the remote monitoring machine. According to the experimental research, this system can achieve the monitoring of landslide along the high-speed effectively.

Acknowledgement. This work is supported by National Nature Science Foundation under Grant 61374059; Student Scientific Training Program in Shanxi Province.

References

1. Hang, D.W., Zhang, P.Z., Wu, C.Q., et al.: The landslide monitoring research and the latest progress. Sens. World **11**(6), 10–14 (2005)
2. Wang, L., Zhang, Q., Guan, J.N.: Based on the technology of GPS landslide dynamic deformation monitoring of the experimental results and analysis. Wuhan Univ. New Pap.: Inf. Sci. Ed. **36**(4), 422–425 (2011)
3. Nichol, J.E., Shaker, A., Wong, M.-S.: Application of high-resolution stereo satellite image to detailed land a slide hazard assessment. Geomorphology **76**(1/2), 68–75 (2006)
4. Girod, L., Byehovskiy, V., Elson, J., et al.: Locating tiny sensors in time and space: a case study. In: Proceedings of the 2002 IEEE International Conference on Computer Design: VLSI in Computers and Processors, pp. 214–219. IEEE Computer Society, Freiburg (2002)

5. Harter, A., Hopper, A., Steggles, P., et al.: The anatomy of a context-aware application. In: Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking, pp. 59–68. ACM Press, Seattle (1999)
6. Nieulescu, D., Nath, B.: Ad hoc positioning system (APS) using AOA. In: Proceedings of the IEEE INFOCOM, pp. 1734–1743. IEEE Computer and Communications Societies, San Francisco (2003)
7. Girod, L., Estrin, D.: Robust range estimation using acoustic and multimodal sensing. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2001), pp. 1312–1320. IEEE Robotics and Automation Society, Maui (2001)
8. Bulusu, N., Heidemann, J., Estrin, D.: GPS-less low cost outdoor localization for very small devices. IEEE Pers. Commun. Mag. **7**(5), 28–34 (2000)
9. Nieuleseu, D., Nath, B.: DV based positioning in ad hoc networks. J. Telecommun. Syst. **22** (1–4), 267–280 (2003)
10. He, T., Huang, C., Blum, B., Stankovic, J., Abdelzaher, T.: Range free localization schemes for large scale sensor networks. In: Proceedings of the 9th Annual International Conference on Mobile Computing and Networking (MOBICOM), pp. 220–224. ACM, New York (2003)
11. Zhan, J., Liu, H.L., Liu, S.G.: Dynamic weighted localization algorithm research based on RSSI. Electron. J. **39**(1), 82–88 (2011)

Application of Computer Simulation in Interference Assessment Between Satellite Systems

Tingting Cao(✉), Dapeng Li, Aiai Ren, and Pingke Deng

Academy of OPTO-Electronics Chinese Academy of Sciences, Beijing, China
caotingting@aoe.ac.cn

Abstract. With the modernization of the Global Navigation System and the rapid development of Leo system (communication or augmentation), the frequency contention of signal is growing worse. By computer simulation and modeling, fast computation of inter-system interference has become very important. In this paper, complex systems are translated into parameters and models, then calculate the effect of the interference between different systems by computer, and analyze interference factors. The results will provide reference for the analysis of interference and the selection of frequency bands for others satellite system.

Keywords: Computer simulation · Parameters model · Interference assessment · Factors analysis

1 Introduction

WRC-12 has decided to the extension of the existing primary and secondary radio determination satellite service (space-to-Earth) allocations in the band 2483.5–2500 MHz in order to make a global primary allocation [1]. So the S band becomes a common band of navigation services and satellite communications services, S band will become the focus of competition. The model of satellite system is established in this paper and simulate the result of complex constellation interference by computer. In future, frequency selection and signal design of satellite systems may refer to the method and conclusion of this paper.

2 System Model

For the main business of S band are satellite communications and navigation, this paper selects navigation system and Leo system which share the same frequency band as an example. After simulation, the degree of the interference between the two systems is obtained. The method is suitable for the constellation system which needs to calculate the interference degree of the downlink signal, especially for the two systems or the multi-system sharing the same frequency band.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 426–432, 2017.
DOI: 10.1007/978-981-10-3969-0_47

A. Navigation system parameters model

To illustrate how the model would apply in the navigation system, a hypothetical example is presented in Table 1. Note that the values used are only for illustration.

B. Leo System Parameters Model

As in the previous section, the values used are only for illustration. For the example, there are four types of signal, but each type of signal should be independently examined.

Input constellation parameters listed in Tables 1 and 2 to model the two system using STK. The 2D track of the two systems is shown in Fig. 1 and their antenna patterns are shown in Fig. 2.

Table 1. Navigation system parameters model

Type	Parameters	Values
Constellation	Satellites no.	27
	Orbital plane no.	3
	Orbital altitude (km)	21500
	Orbital inclination angle (°)	55
Navigation signal	Center frequency (MHz)	2492.028
	Modulation	QPSK(8)
Satellite antenna	Power (dBW)	19
	Gain	13

Table 2. Leo system parameters model

Type	Parameters	Values
Constellation	Satellites no.	27
	Orbital plane no.	3
	Orbital altitude (km)	21500
	Orbital inclination angle (°)	55
Leo signal 1/signal 2	Center frequency (MHz)	2496.12
	Modulation	QPSK(4)/QPSK(2)
Leo signal 3/signal 4	Center frequency (MHz)	2492.028
	Modulation	QPSK(4)/QPSK(2)
Satellite antenna	Power (dBW)	1
	Gain	5.76

3 Calculation Methods and Results

This paper adopts the M.1831 method which is recommended by the 8D workgroup of ITU-R [2], the method bases on spectral separation coefficient (SSC) and the effective carrier-to-noise ratio (C/N_0) to evaluate the degradation of (C/N_0) due to inter-system

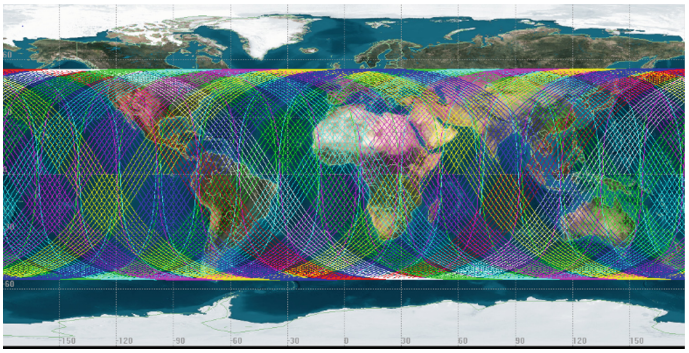


Fig. 1. The 2D track of the two systems

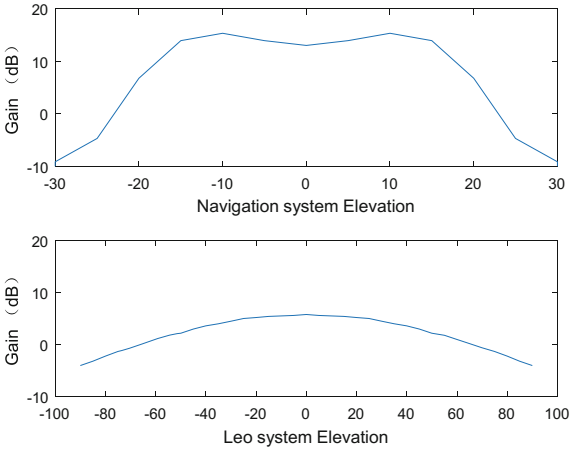


Fig. 2. Antenna patterns of the navigation system and the Leo system

interference [3], which is denoted by $\Delta(C/N_0)$. The method can ensure that the actual results will not exceed the theoretical value. It's universal and irrelevant to different receiver design.

A. The method calculation of SSC

The definition of spectral separation coefficient (SSC) is written as [4]:

$$K_{js} = \int_{B/2}^{-B/2} G_j(f + f_{dopj}) G_s(f + f_{dops}) df \tag{1}$$

Where

B: The bandwidth of the receiver;
G_j: Power spectral density (PSD) of the interference signals;

G_s: PSD of the desired signals;
f_{dopj}: The Doppler shift of the interference signals;
f_{dops}: The Doppler shift of the desired signals.

The SSC indirectly reflects the two signals' overlap and interference. PSD of the Navigation system and the Leo system signals are shown in Fig. 3.

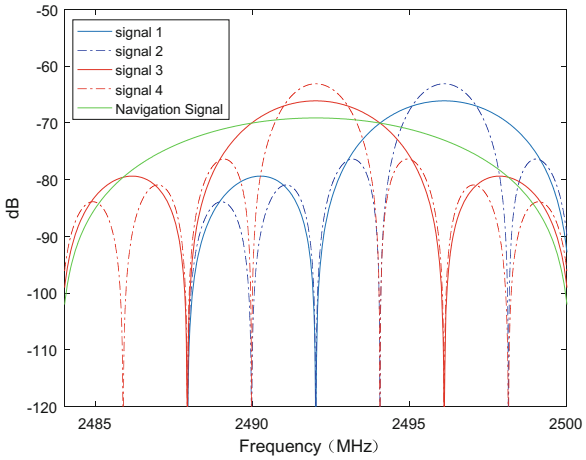


Fig. 3. PSD of the navigation system and the Leo system signals

According to the formula (1) and the different signals' PSD, SSC calculation result is given in Table 3, signal 1, signal 2, signal 3, signal 4 of Leo and the signal of the navigation is calculated respectively.

Table 3. SSC of navigation and Leo signals

Signal	Center frequency (MHz)	Modulation	SSC with navigation signal	SSC with itself
Navigation signal	2492.028	16.368	−70.9034	
Signal 1	2496.12	QPSK(4)	−73.0741	−67.8876
Signal 2	2496.12	QPSK(2)	−73.0641	−64.8711
Signal 3	2492.028	QPSK(4)	−69.9253	−67.8872
Signal 4	2492.028	QPSK(2)	−69.5092	−64.8702

B. The method calculation of $\triangle(C/N_0)$

Typically, the effective carrier-to-noise ratio (C/N_0) is used measure the impact of the interference from various sources. Now, it is used in assessing inter-system interference of different systems. C/N_0 is given by:

$$\frac{C}{N_0'} = \frac{C}{N_0 + I_{ref} + I_{\text{int}} + I_{ext}} \tag{2}$$

Where:

C: desired-signal power (W) from satellite in the reference constellation;
N_0: PSD the thermal noise;
I_{ref}: PSD of the interference signals from satellites in the same constellation as the desired signals;
I_{int}: PSD of the aggregate interference from all the in-view satellites other than those in the reference constellation;
I_{ext}: PSD of the interference signals from all radio signals other than those of the RNSS.

When considering the intra-system interference only, I_{int} and I_{ext} would be set to 0.

$$\triangle\left(\frac{C}{N_0}\right)_{\text{intra}} = \frac{C/(N_0)}{C/(N_0 + I_{ref})} = 1 + \frac{I_{ref}}{N_0} \tag{3}$$

When considering the inter-system interference, we need I_{ref} and I_{int}, I_{ext} would be set to 0.

$$\triangle\left(\frac{C}{N_0}\right)_{\text{inter}} = \frac{C/(N_0 + I_{ref})}{C/(N_0 + I_{ref} + I_{\text{int}})} = 1 + \frac{I_{\text{int}}}{N_0 + I_{ref}} \tag{4}$$

To calculate $\Delta(C/N_0)$, we should calculate the aggregate interference of the reference system (I_{ref}) and the interference system (I_{int}), and it is the maximum interference power of which is received by receivers distributed around earth.

In the paper, there is only one signal in each system, so the aggregate interference can be defined by:

$$I = \max\left(\sum_{u=1}^{U} C_u K\right) \tag{5}$$

Where:

U: the number of satellites in view;
C_u: the received power from the u-th satellite at a receiver everywhere;
K: the SSC between the desired signal and the interference signal.

Therefore, in order to get the value of I, firstly, calculate the SSC between the desired signal and the interference signal; secondly, calculate the received interference power from all satellites in view at a receiver everywhere.

C. Set up parameters and simulation [5]

The receiver is evenly distributed in 120° longitude and the simulation period is 7 days. Receiver bandwidth is set to 16 MHz, the maximum gain is 3 dB. Input the receiver parameters, constellation parameter and signal parameters to STK, to calculate the maximum aggregate power of two systems received on the earth surface. The results are shown in Table 4.

Table 4. Aggregate power of navigation and the Leo

Receiver position		Leo signal power (dBW)	Navigation signal power (dBW)
Longitude (°)	Latitude (°)		
120E	80N	−161.97	−141.753
120E	50N	−145.967	−142.227
120E	20N	−146.201	−141.644
120E	0N	−148.436	−142.151
120E	20S	−146.193	−141.646
120W	20N	−146.192	−141.704
120W	50S	−145.972	−142.229
120W	80S	−161.97	−141.754
Max		−145.967	−141.644

According to the SSC result of Table 3, the maximum the aggregate power of each system is taken as the interference power to each other. The value can ensure that the theoretical results will not be exceed by the actual results of the interference. According to the formula (4), the results of $\Delta(C/N_0)$ caused by the inter-system interference are as follows (Table 5):

Table 5. $\Delta(C/N_0)$ caused by the inter-system interference

Interference	Desired				
	Navigation signal	Signal 1	Signal 2	Signal 3	Signal 4
Navigation signal	0.3285	0.1913	0.1820	0.3862	0.4021
Signal 1	0.0704	0.2455	–	–	–
Signal 2	0.0705	–	0.4784	–	–
Signal 3	0.1440	–	–	0.2455	–
Signal 4	0.1583	–	–	–	0.4785

According to Table 4 calculation results, when Navigation system is the reference constellation, the navigation system $\Delta(C/N_0)$ caused by itself is 0.3285 and the inter-system interference is not more than 0.1583. So we learn that the worst impact on the system is the intra-system interference. The signal 3 and signal 4 has the same center frequency with navigation system, $\Delta(C/N_0)$ caused by them is 0.1440 and 0.1583. The results show clearly that the effect of the bandwidth is not obvious. The center frequency of the signal 1 and signal 2 separate 4 MHz from navigation system,

their $\Delta(C/N_0)$ is 0.0704 and 0.0705, decreased by about 50% compare with signal 3 and signal 4. The result shows the main factors affecting $\Delta(C/N_0)$ is the interval between two center frequencies. The interference of the same frequency signal is the largest, the greater frequency interval is, the smaller the $\Delta(C/N_0)$ is. The conclusions are consistent when Leo communication is the reference constellation.

4 Conclusions

By the computer simulation, the complex interference analysis is converted into the parameter model. Only need to enter the necessary parameters of each system, you can get the aggregate interference. Save a large number of repeated calculations and greatly improving the efficiency. The simulation results are true and the conclusion can be used for the interference analysis and the frequency selection for future satellite systems.

References

1. Electronic Communications Committee. Compatibility studies between RDSS and other services in the band 2483.5–2500 MHz, Amended (2012)
2. ITU-R the 8D workgroup. Recommendation ITU-R M.1831-1 (2015)
3. Wu, C., He, C.: Interference analysis among modernized GNSS. In: IEEE Proceedings (ICCP 2011), pp. 669–673 (2011)
4. Yunzhi, L.: A compatible interoperability evaluation method for global navigation satellite system signal. Appl. Electron. Tech. **40**, 95–97 (2014)
5. Xinyan, Z., Xiaodong, Z., Lili, G.: Simulation and analysis of aggregate gain factor of Compass, GPS and Galileo. Appl. Electron. Tech. **9**, 95–97 (2009)

Research and Application of Three-Dimensional Simulation Technology on Virtual Display of Skirt

Yan Wan, Zheng Tie(✉), and Zilin Shi

Computer Science and Technology Institute,
Dong Hua University, Shang Hai, China
tzzoel992@163.com

Abstract. In the paper, the mass-spring model was used as the basic cloth simulation model to simulate the skirt. After modeling, different forces were applied on the particles and generated differential equations. To solve the equations, different kinds of integration methods was used in the paper for comparisons of getting the best results. To correct the super-elastic phenomenon usually occurring in cloth simulation, the improved mass-constrained method was also adopted. The paper mainly shows the process of generating a skirt simulation model with obviously realistic wrinkles and folds fitting with an elliptic-shaping waist model.

Keywords: Skirt simulation · Mass-spring model · Virtual fitting

1 Introduction

With the development of computer simulation technology and increasing demands for virtual fitting applications, cloth simulation based on the purpose of clothing deformation has been a hot issue in the field of computer graphics. The cloth simulation technology is kind of techniques that use the physical or mathematical models complying with the similar principle of geometric, environment or property to take the place of the actual system.

It mainly includes two parts in the cloth simulation technology used mostly: building up cloth models and mathematics integral calculation. Physics-based ways [1, 2] to build up cloth models are widely used in the field comparing with geometry-based ways [3, 4]. After modeling the cloth, its movement and deformation can be described by differential equations related to particles' forces, acceleration, velocity, etc. Several methods can be used to solve the equations, and in this paper, four main representative methods were used as comparisons for choosing the best among the final results. The cloth simulation techniques are widely used in foreign countries, but their domestic application is quiet poor. As most cloth simulation techniques were applied in simulating rectangular tablecloths or curtains [5], the application of the techniques seems to be narrowly used. Also, the techniques of simulating cloth are mostly based on a square table with four corners, a handrail for handing the flag [6] or a ball whose area is much less than the area of the cloth, the wrinkles and folds seem more likely to form under which circumstances. The paper brought about a circular-shaping skirt simulation model basing on an elliptic

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 433–441, 2017.
DOI: 10.1007/978-981-10-3969-0_48

shaping waist model using cloth simulation technology to perform a three-dimensional virtual display with obvious folds and wrinkles on the skirt.

In this paper, the skirt simulation model deriving from the mass-spring model was used as the basic model. After that, different forces were applied on particles generating the differential equations which can be solved by the explicit integration methods and the Verlet integration method. Also, to prevent the model from super-elastic phenomenon, the mass-constrained method should be conducted to complete the cloth simulation. Finally, the skirt model would be used to fit an elliptic-shaping waist model with obvious wrinkles and folds [7].

2 Related Work

2.1 The Framework

To finally present a skirt simulation model with wrinkles and folds, firstly physics-based mass-spring model was brought in as the basic cloth model shaping as square. After that, force analysis should be done to generate the dynamic differential equations of each particle. Four integral methods were used in the paper to solve these equations for comparisons, among which the best method could be found out depending on the eventual presentation. With the super-elastic phenomenon appears, mass-constrained methods should be implemented on each particle after each iteration of calculation to correct the over-stretched spring length in mass-spring model. Detailed description is as follows.

2.2 Mass-Spring Model

The mass-spring model proposed by the Provot [8] is a primary method of cloth modeling based on physics. It has been widely used with its simplicity, practicality and efficiency. In the model, the cloth is seen as a series of uniformly distributed particles, and the particles are connected by springs with different types. These springs are used to simulate the interaction between the particles in different directions with no mass.

There are three types of springs connecting the particles respectively known as the structure spring simulating stretching and compressing forces, the shearing spring simulating the ductility of the cloth and the bending spring preventing from tearing up the cloth (Fig. 1).

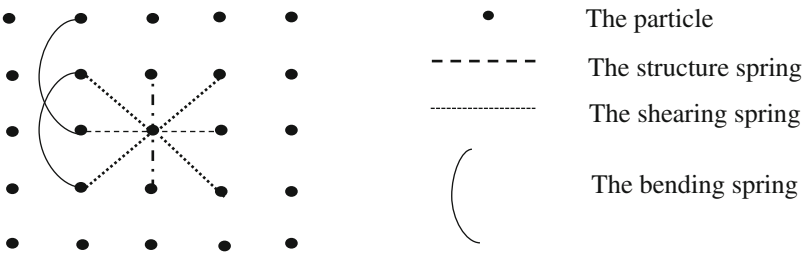


Fig. 1. The mass-spring model

2.3 Force Analysis

For a cloth model composing of discrete particles, the motion state of each particle depends on the sum of external and internal forces.

The Newton's second law of motion shows the relationship between the resultant force with mass and acceleration, wherein $\partial^2 X/\partial^2 t$ is the differential acceleration expression:

$$F_{internal}(X,t) + F_{external}(X,t) = m * \partial^2 X/\partial^2 t \tag{1}$$

2.3.1 The Internal Forces

The internal forces applied on particles include the spring force and the damping force.

The springs connecting two particles i, j are one of three basic springs. Each one has different spring force applied on the particles according to their elastic coefficient. By the Hooke's law, the spring force for a single particle is:

$$F_{elastic}(i,j) = K_s * \left(x_{ij} - l_0\right) * \left(x_{ij}/|x_{ij}|\right) \tag{2}$$

Wherein, K_s is the elastic coefficient which needs to be set manually, $x_{ij} = x_j - x_i$ refers to the difference of distance vector at time t of particles i, j.

In the mass-spring model, since forces applied on the spring may lead to a certain vibration, to reduce excessive vibration during particle's motion, the damping force should be added on the model described as:

$$F_{damping}(i) = -K_d * \left(v_i - v_j\right) \tag{3}$$

Wherein, K_d is the damping coefficient, v_i, v_j represents the velocity of particles i, j at time t.

2.3.2 The External Forces

The external forces applied on particles include gravity and air resistance.

$$F_{external} = G + F_{air} \tag{4}$$

As the cloth is composed of uniformly distributed particles, the gravity of the cloth can be seen as the sum of the gravity of each particle expressed as: $G_i = m_i * g$.

The air resistance refers to the air friction during the motion of particles which will slow the change of length of the spring and make the model more stable. The air resistance also belongs to the damping force which can be defined as follows, wherein k_{air} is the damping coefficient:

$$F_{air}(i) = -v_i * k_{air} \tag{5}$$

2.4 Numerical Integration Method

After the simulation of particles connected by different types of springs has been completed, dynamic differential equations should be listed based on kinetics to study the trajectory of particles and to be solved. Several numerical integration methods can be applied now in the cloth simulation.

The explicit integration methods [9] are methods based on the time difference, including the explicit Euler method, the explicit midpoint method and the Runge-Kutta method.

The explicit Euler method [10, 11] is a simple integration method which is easier to implement and is more effective compared with other explicit methods.

The formula of the explicit Euler method used in the mass-spring model is as follows, wherein a_i is the acceleration of particles, Fi is the resultant force, v_i is the velocity, x is the position of particle at current time step, Δt is the time step:

$$a_i(t+\Delta t) = F_i(t)/m_i \tag{6}$$

$$v_i(t+\Delta t) = v_i(t) + \Delta t * a_i(t+\Delta t) \tag{7}$$

$$x_i(t+\Delta t) = x_i(t) + \Delta t * v_i(t+\Delta t) \tag{8}$$

The explicit midpoint method [9] is an extension of the Euler method. During the method, two times of the Euler's formula will be invoked each iteration. The core of the midpoint method is to use slope Eq. (9) to get the approximate Eq. (10).

$$y'(t) \approx (y(t+h) - y(t))/h \tag{9}$$

$$y(t+h) \approx y(t) + h * f(t, y(t)) \tag{10}$$

For the explicit midpoint method, a more precise value of (t + h/2) can be used in the Eq. (10), the result can be obtained as:

$$y(t+h) \approx y(t) + h * f(t+h/2,\ y(t) + h * f(t, y(t))/2) \tag{11}$$

The most commonly used Runge-Kutta method is the fourth-order Runge-Kutta method [12], which firstly estimates slope values of four points within the interval [x_i, x_i + 1], then uses a weighted average of the slope values as the approximate value of the average slope to obtain an equation as follows:

$$y_n + 1 = y_n + h * (K_1 + 2K_2 + 2K_3 + K_4)/6 \tag{12}$$

$$K_1 = f(x_n,\ y_n) \tag{13}$$

$$K_2 = f(x_n + h/2,\ y_n + h * K_1/2) \tag{14}$$

$$K_3 = f(x_n + h/2,\ y_n + h * K_2/2) \tag{15}$$

$$K_4 = f(x_n + h,\ y_n + h * K_3) \tag{16}$$

The Verlet integration [13] uses the position x(t) and the acceleration a(t) of particles at time t as well as the position at time t−h to calculate the position at the time t + h, the expansion of the position equation with the Taylor format is as follows:

$$x(t+h) = x(t) + hx'(t) + h^2x''(t)/2! + h^3x'''(t)/3! + \ldots \tag{17}$$

Replace the h in the equation as (–h):

$$x(t-h) = x(t) - hx'(t) + h^2x''(t)/2! - h^3x'''(t)/3! + \ldots \tag{18}$$

Obtained by adding the former two equations:

$$x(t+h) = 2 * x(t) - x(t-h) + h^2x''(t)/2! + \ldots \tag{19}$$

2.5 The Mass-Constrained Method

Generally, the deformation curve influenced by the forces is non-linear, which may cause the super-elastic phenomenon. The traditional mass-constrained method [8] to prevent it from the phenomenon is to calculate the spring elongation after each iteration, when the spring elongation is larger than the critical value, constraints should be imposed to the particles. This method would still cause the super-elastic phenomenon when the forces increasing to a certain extent, so the improved mass-constrained method [13] was used in the paper described as follows:

(1) Traverse the particles currently and compare the elongation of the springs connected to certain particle. When finishing, correcting the position of the particles linked by the spring which has the largest elongation of certain particle determined above. The correction formula is as follows:

$$\Delta l = \left|x_j - x_i\right| - l_0 \tag{20}$$

$$\Delta p_i = +m_i/(m_i + m_j) * \Delta l * (x_j - x_i)/(|x_j - x_i|) \tag{21}$$

$$\Delta p_j = -m_j/(m_i + m_j) * \Delta l * (x_j - x_i)/(|x_j - x_i|) \tag{22}$$

Wherein, x_j, x_i refers to the current position of two particles, l0 refers to the original length of the spring, Δp_i, Δp_j refers to the displacement correction of the particles i, j, m_i, m_j refers to the mass of the particles.

(2) For other springs whose elongation is larger than the set critical value, velocity should be changed as setting the velocity component which will cause excessive stretch to zero [14].

3 The Initial Skirt Model and Waist Model

Application of the mass-spring model has its particularity as particles in it are connected in warp and weft directions or inclined longitudinal directions. These connections are in a square area. To simulate a skirt model in the paper, a circle was chosen as the initial shape of the cloth. Considering the structure of the mass-spring model, the circle cloth model was also derived from a square area. In a square where springs connecting readily with particles as described in mass-spring model, the center point had been selected as the circle's center too. Then, only the particles in the circle area were left to simulate the cloth to shape it as a circle. That is to say, the initial area was a square, then the distance between each particle with the center particle should be estimated whether or not smaller than the set radius. Remaining the particles and the spring connections which are corresponding to the estimation and setting no connections of others.

It can be concluded during the process of cloth simulation that it would cause serious super-elastic phenomenon if the numbers of particles of the square were set too much. On the other hand, the simulation results would be affected if the numbers were set too little. As well, for the simulation of circle cloth, the degree of the cloth approaching to the circle depended on the density of the square since it was cut along the inner square. To sum up, proper numbers of particles should be chosen after conducting many times of experiments. The numbers chosen in the paper is 45*45 after several experiments.

As to simulate the wrinkles and folds of a skirt, waist model is needed as basic model. According to several studies, the body's waist contour can be substantially seen as the elliptical configuration out of the skirt model. The waist contour was set as an oval, of which the center of the major axis was exactly the center of the circle skirt model. As we can see from the figure, the simulation result for the waist model is quiet similar to an oval. The use of the waist model is different from the square as it has no corners which may be easy to lead to the folds. Also, the elliptical model can be more authentic for visual effects (Fig. 2).

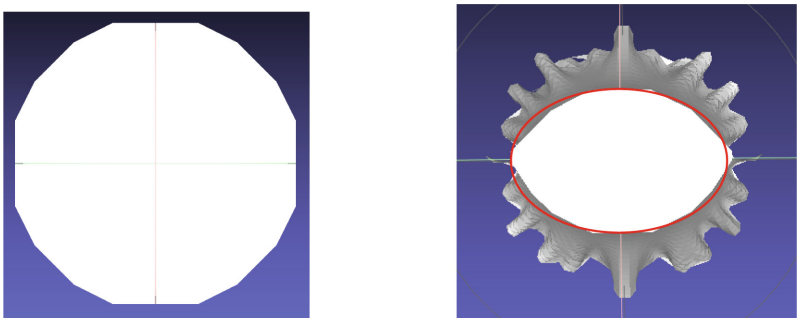


Fig. 2. The initial model of a skirt and a waist

4 Final Results and Analysis

In the paper, firstly the mass-spring model was used to construct square cloth, then inner particles within the scope of the set radius constituted a circle skirt model. Setting corresponding springs between these particles, then fitting out the elliptical area which represented the body's waist contour. The particles in the elliptical area being set without any forces were called fixed particles. For other non-fixed particles, the force analysis should be conducted at first to calculate the resultant forces by summing up internal forces and external forces within the current iteration. The position of each particle could be figured out using referred integration method. After an iteration of the integration, the improved mass-constrained method should be invoked to correct the deviation caused by the super-elastic phenomenon to generate a more realistic simulation effect.

As mentioned, four main integration methods known as the explicit Euler integration, the explicit midpoint integration, the fourth-order Runge-Kutta integration and the Verlet integration were used for finding out the most authentic result to solve the differential equations based on the same model during same times of iteration and the results show as follows (the screenshot both from the above and the front) for comparisons. It can be concluded from the results that the skirt simulation using the explicit midpoint integration still has obvious super-elastic phenomenon and stretches the cloth longer. As two times of Euler integration equation being invoked each iteration, the actual times of iteration is doubled which may cause stretching the cloth. The simulation using the fourth-order Runge-Kutta integration has slower process while falling near the fixed particles which results in obvious distortion and the effect looks like an umbrella skirt. As it showed in the figure, the presentation using the methods is the most unnatural. The simulation using the Verlet integration lacks of folds with visual effect similar to tight fishtail skirt. While experiments conducted on the four-cornered table using this methods to simulate the tablecloths presenting quite good folds of the cloth near the corner, this method may just fit to simulate cloth based on edged foundation. Since the improved mass-constrained method cannot be used in the Verlet integration, after traditional constrained method, it has obvious super-elastic phenomenon as well. The simulation using the explicit Euler integration has the most realistic effect and the wrinkles as well as the folds highly correspond to the real state. In summary, the explicit Euler integration is the most suitable integration method in the paper to study skirt simulation. As well, the comparisons between long skirt and short skirt using the explicit Euler integration show as below (Figs. 3, 4, 5, 6 and 7).

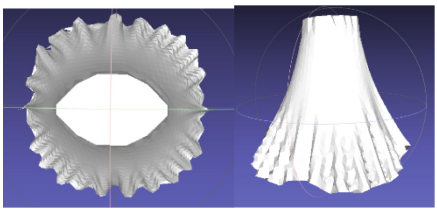


Fig. 3. The explicit Euler integration

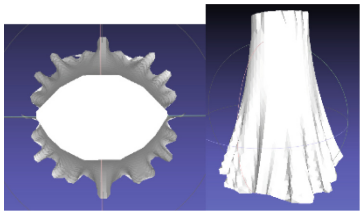


Fig. 4. The explicit midpoint integration

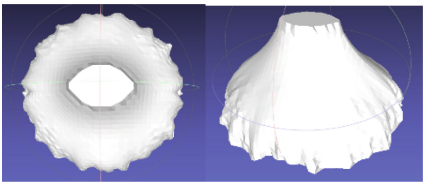


Fig. 5. The fourth-order Runge-Kutta integration

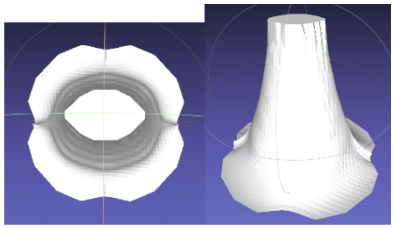


Fig. 6. The Verlet integration

Fig. 7. The long skirt and short skirt contrast using the explicit Euler integration method

5 Conclusion and Future Work

The paper studied the three-dimensional cloth simulation and its application in skirt simulation. Also, to simulate the reality of skirt on body model, an elliptic-shaping waist model was brought in innovatively. The mass-spring model was used to build up the basic cloth model in the paper and different integration methods were experimented to conduct the calculation. As it can be seen from the results that using the explicit Euler integration method can better simulate the deformation effects of the cloth and can also produce a more realistic simulation effect. The shape of the model of the skirt in the paper is similar to circle which is closer to real-life prototype of cloth cutting, also the waist contour of body model is imitated with ellipse which is more in line with the reality human body model. Experimental results show that using the mass-spring model as basic cloth model, meanwhile conducting the calculation with the explicit Euler integration method can better simulate the wrinkles and folds of skirts in a more realistic way. Improved mass constrained method was used while simulating in the paper can better make the skirt look more realistic during the progress of producing wrinkles while falling. As well, obvious super-elastic phenomenon will not occur during the circumstance. Although it increases the amount of calculation, it has a more pronounced realistic cloth simulation. The study of the cloth simulation in the paper was only conducted on the relatively simple models such as seeing waist as ellipse, so the simulation on the more complex upper body model needs further study. A relatively simple circular shape was used while building up the prototype of the skirt, but for a

formed cloth, the shape was too monotonous and simple and needed to be improved. These are the future directions to be studied and improved in the application.

References

1. Schröder, K., Zhao, S., Zinke, A.: Recent advances in physically-based appearance modeling of cloth. SIGGRAPH Asia 2012 Courses, p. 12. ACM (2012)
2. Breen, D.E., House, D.H., Getto, P.H.: A physically-based particle model of woven cloth. Vis. Comput. **8**(5–6), 264–277 (1992)
3. Weil, J.: The synthesis of cloth objects. ACM SIGGRAPH Comput. Graph. **20**(4), 49–54 (1986)
4. Agui, T., Nagao, Y., Nakajma, M.: An expression method of cylindrical cloth objects-an expression of folds of a sleeve using computer graphics. Trans. Soc. Electron. Inf. Commun. **J73-D-II**(7), 1095–1097 (1990)
5. Benameur, S., Djedi, N.E.: Multi-resolution cloth simulation based on particle position correction. Int. J. Comput. Appl. **143**, 29–36 (2016)
6. Huang, W.X., Sung, H.J.: Three-dimensional simulation of a flapping flag in a uniform flow. J. Fluid Mech. **653**, 301–336 (2010)
7. Bridson, R., Marino, S., Fedkiw, R.: Simulation of clothing with folds and wrinkles. In: Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 28–36. Eurographics Association (2003)
8. Provot, X.: Deformation constraints in a mass-spring model to describe rigid cloth behaviour. In: Graphics Interface, p. 147. Canadian Information Processing Society (1995)
9. Volino, P., Magnenat-Thalmann, N.: Comparing efficiency of integration methods for cloth simulation. In: International Proceedings on Computer Graphics 2001, p. 265. IEEE (2001)
10. Müller, M., Heidelberger, B., Hennix, M., et al.: Position based dynamics. J. Vis. Commun. Image Represent. **18**(2), 109–118 (2007)
11. Hutzenthaler, M., Jentzen, A., Kloeden, P.E.: Strong and weak divergence in finite time of Euler's method for stochastic differential equations with non-globally Lipschitz continuous coefficients. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 467(2130), pp. 1563–1576. The Royal Society (2011)
12. Pronk, S., Páll, S., Schulz, R., et al.: GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics **29**(7), 845–854 (2013). btt055
13. Vassilev, T., Spanlang, B., Chrysanthou, Y.: Fast cloth animation on walking avatars. In: Computer Graphics Forum, pp. 260–267. Blackwell Publishers Ltd. (2001)
14. Desbrun, M., Schröder, P., Barr, A.: Interactive animation of structured deformable objects. Graph. Interf. **99**(5), 10 (1999)

Database Construction and Map Compilation of Provincial Common Geographic Maps

Guizhi Wang[1(✉)] and Wen Zhou[2]

[1] Database Department, National Geomatics Center of China, Beijing 100830, China
wmj@nsdi.gov.cn
[2] Technology Research and Development Department, HASMG, Haerbin 150086, China
34010303@qq.com

Abstract. The project of new century version of the National Huge Atlas of the People's Republic of China has an important achievement that is the National General Atlas of China. Moreover, different provincial common geographic maps are main content of the National General Atlas of China. Database construction and map compilation of different common provincial geographic maps abide by the design idea of mapping after establishing database, make full use of the national 1:1000 000 fundamental geographic information database, and adopt the technology of automatic illustrated based on database. This paper mainly introduces the achievement form of different provincial common geographic maps, compiled technical route, process flow and key technology, etc.

Keywords: Design and development · New century version of the national general atlas of China · Different provincial common geographic maps · Automatic illustrated based on database · Database construction · Electronic map compilation

1 Introduction

The national general atlas can fully display the fundamental geographic information content of China. It is one of the important basic data which can make people full understanding of the present situation and the level of development in our country. The compilation of the national general atlas has received high attention in the government and academic circles in China. Since the founding of China, China has compiled and published the national general atlas of China twice in the '50–'60s and '80–'90s. However, hasn't been updated and republished the national general atlas of China in recent 20 years. This 20 years is the rapid development period in infrastructure construction, economic and social. Existing Atlas can't objectively reflect the real situation of China's national geographical conditions. People's demand and using patterns for atlas has become increasingly diverse, and also put forward higher requirements for the

Taking the new century edition of the national general atlas of China for example.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 442–450, 2017.
DOI: 10.1007/978-981-10-3969-0_49

up-to-dateness of contents in Atlas. Therefore, it is urgent need that rapidly compile the new century version of Atlas using the latest technology measures and expression ways.

At the same time, the China have made major progress in the data achievements accumulation and mapping technology innovation, successively built multi-scale fundamental geographic information database from national to local one after another, completed the first phase project of island (reef) mapping, fully carried out the geographical conditions census. 'No. 3 resource' surveying and mapping satellite can quickly obtain high resolution image data. With gradual perfecting of the surveying and mapping science and technology innovation system, the digital mapping technology and quick updating and mapping technology based on database has achieved fruitful results. Mastered a batch of the graph-database integration technology with independent intellectual property rights, such as multi-scale linkage downsizing, incremental updating based on database, fast mapping, etc. The above laid a solid foundation of data and technology for compiling the new century version of the national general atlas of China.

In 2013, the ministry of science and technology started basic work project of science and technology on research and compilation of new century version of the National Huge Atlas of the People's Republic of China. The project needs to complete the demonstration projects which include the national general atlas, the economic atlas and the administrative area atlas. Among them, the national general atlas is the foundation of other atlas, it consists of general maps, national thematic map, provincial geographic maps, provincial level city maps, place names index, etc. And provincial geographic maps are its important component. This article mainly discusses the compilation technology of provincial geographic map.

2 Achievement Forms of Provincial Geographic Maps

The achievements of provincial geographic maps include the two forms which are basic database and the electronic map sheets.

2.1 Basic Database of Provincial Geographic Maps

Provincial geographic maps, with relative balance degree, comprehensively display the basic geographic information of China in detail. They mainly include the spatial distribution and attribute information of basic elements such as hydrology, topography, residents, transportation, administrative zones, soil and vegetation. They are basically consistent with the national fundamental geographic information database on expression content and data structure. Therefore, the provincial geographic maps are in harmony with 1:1000 000 scale database of national fundamental geographic information in the database structure, coordinate system, feature content organization, data classification and coding, attribute table structure, data storage format, etc. Such the

provincial geographic maps are easy to continuously updated with the national fundamental geographic information database.

The database of provincial geographic maps includes 9 kinds of basic elements vector data. The database covers the scope of each provincial geographic map, with the 2000 national geodetic coordinate system and the 1985 national elevation datum, no projection. It is stored using the geographic coordinates of latitude and longitude, the unit is degree.

2.2 Sheets of Provincial Geographic Maps

Provincial geographic maps are cut and extracted based on the databases taking province-level administrative area as mapping units. There are 34 province-level mapping units or provincial geographic maps all over the country. Each mapping unit includes three sheets which are administrative zoning map, relief map and surface overlay map. The administrative zoning map mainly represents humanities elements, the relief map mainly represents natural elements, and the surface overlay map mainly represents all types of surface. Three types of map cooperate with each other so that roundly represent basic geographical elements of various provinces, municipalities directly under the central government and autonomous regions. They detailedly display geographical landscape and characteristics in each mapping unit.

The administrative zoning maps take regional city as basic color unit in most of the provinces, but take district or county as basic color unit in municipality directly under the central government, Hainan province and Ningxia autonomous region. They detailly represent the human elements such as residents, transportation, boundary, etc. and roughly represent the natural elements such as hydrology and soil, but do not represent vegetation and topography.

The relief maps represent landscape pattern with contour and color hill-shading. They mainly represent the natural elements such as hydrology, soil, vegetation, natural scenic reserve, and roughly represent the human elements such as residents at or above the county-level, transportation at or above province-level, boundary at or above regional city but at or above district or county in municipality directly under the central government, Hainan province and Ningxia autonomous region.

The surface overlay maps represent 10 types of surface coverage elements using qualitative bottom method such as cultivated land, forest, grassland, shrub land, wetland, water body, tundra, artificial surface, bare land, glaciers and permanent snow. They roughly represent basic geographical elements such as residents at or above county-level, boundary at or above regional city but at or above district or county in municipality directly under the central government, Hainan province and Ningxia autonomous region.

Provincial geographic maps emphasize overall effect of regional shape on vision and require correctness in direction. So they use the double standard parallel conformal conic projection. The determination of double standard parallel and central meridian is according to scope of latitude and longitude of the latest provincial boundary in various

provinces, using the calculation formula of standard parallel and central meridian under the condition of equal deformation absolute value in edge-weft and centre-weft, referring to the projection parameters in 1995 version of the national general atlas and 2004 version of the atlas of China, passing program calculation, deformation analysis, and fine-tuning rounding.

Under the premise of making full use of the atlas type area size and fully displaying graphics area within a double-paged or a single-paged, unified design each map scale of provincial geographic map, keep simple ratio relations between the map scales as far as possible, and types of scale should not be too much.

3 Technical Route and Process Flow

The general idea of provincial geographic map is that firstly establish spatial database which can meet mapping needs and then proceed intelligent distribution map, setting projection and scale, cutting according to frame and so on based on the database. Furthermore, carry out human-computer interaction editing process and finally generate the prepress data of provincial geographic map.

3.1 Technical Route of Building Spatial Database

The national 1:10 00000 fundamental geographic information database is the main data source of compiling provincial geographic map. In order to meet the needs of different province geographical map, first of all, need to supplement and extend database scope in surrounding areas of China based on national 1:10 00000 fundamental geographic information database using the global 1:10 00000 database, and establish the 1:10 00000 common geographic map database for whole China and neighboring countries which cover various provincial geographical map and keep current character of 2012 year.

Due to the global 1:10 00000 database do not agree with the national 1:10 00000 fundamental geographic information database in model, the former need to be normalized on the basis of the latter in production process, including data analysis, extraction, merger, coding and layering.

In order to improve current character, need to update the geographical map database. Therefore, comprehensively utilizing the latest national fundamental geographic information database and other forms of current information, finding the changes in data content through comparative analysis, extracting incremental change data, fast incrementally update the 1:10 00000 common geographic map database for China and its surrounding areas and set up the common geographic map database for meeting the needs of different province geographical map compilation. Its current character is provided by 2015.

Technical route and process flow of building the common geographic map database is shown in Fig. 1.

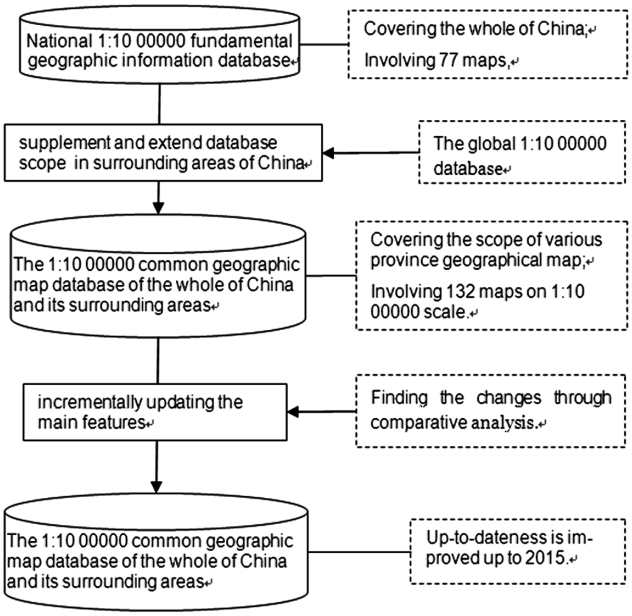


Fig. 1. Technical route and process flow of building common geographic map.

3.2 Process Flow of Compiling Map

Compiling of provincial geographical maps is mainly based on the 1:10 00000 common geographic map database of the whole of China and its surrounding areas, use the 30 m global surface coverage data (GlobeLand30), adopt the new technology of full digital desktop mapping. The specific technical route and mapping process is shown in Fig. 2.

According to design requirements of the general atlas of China, carry out projection transformation based on the 1:10 00000 common geographic map database. Further, set the scale and determine the cutting box size according to the size of graphics area and edition core, and guarantee basic drawing area in center position and determine the direction of due north. Then, separately compile the administrative zoning maps and the relief maps of each province through cutting database, extracting sheet data, converting data format, editing graphics, configuration symbols, doing cartographic generalization, adjusting layers and so on. At the same time, compile the surface overlay maps using the data from the GlobeLand30, through cartographic generalization, editing and processing.

Each of provincial common geographic maps is made up of three sheets which are the administrative zoning maps, the relief maps and the surface overlay maps. Every map need to be printed to paper map after completing map compilation. The paper maps are used to review along with their electronic maps. In order to ensure the quality of maps, it is essential not only the self-check of production departments but also the quality control of employed experts in the whole compiling process. If there are some

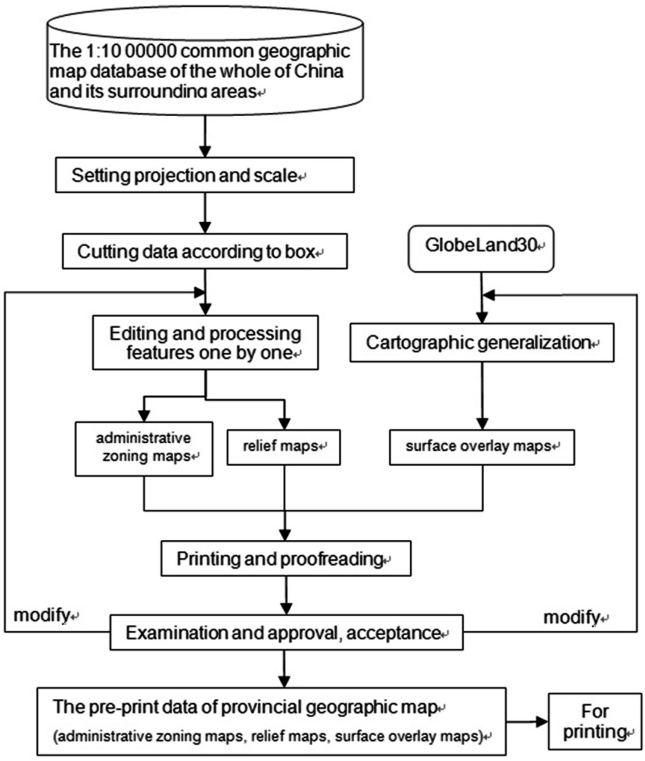


Fig. 2. Technology route and process flow of compiling map.

problems or mistakes in maps, they must be return to original production departments for modifying. The results of modified maps will be sent away to the security examination department of map for examination and approval. Need to modify the maps once again according to the opinions of examination and approval, and finish the acceptance work. Final, the pre-print data of converted and derived provincial geographic map will be used to print paper atlas.

4 The Research of Key Technology

The new century version of National General Atlas of China is a large, authority and basic national science reference atlas. It should comprehensively, systematically and intuitively reflect the basic national conditions about the status of the natural geography and social economy development of China. It also should fully embody the development level of science and technology of our country. Therefore, the provincial geographic maps serve as main body content of National General Atlas of China, many fields should be deeply researched and designed in detail such as material authority, data current character, content scientificity, technology advance, product practicability

and art expressionism. In order to ensure the authority of "atlas", should take all necessary measures. The main technical difficulties embodied as follows:

Workload Heavy and Technology Complicated. The research design and editing producing for new century version of provincial geographic map, not only must do a large amount of data processing and map editing to meet the requirements of printing-publishing and embody the national height of map products, but also need to adopt the most advanced technical method and process flow. Further, through increasing the intensity of science and technology innovation, design and optimize fast technology route and process to improve the level of automation and intellectualization.

Therefore, formulated the basic idea of "firstly building library, then compiling maps", not only guarantee the advanced of research and development achievements, but also consider the sustainability of linked updating library in future. In producing data and building database, using the technique methods of multivariate data integration, automatic downsizing, incremental updating and so on. In order to improve the automation level, promote the work efficiency, reduce the labor intensity, researched and developed the corresponding software tools such as automatic illustrating, intelligent adjustment, automatic cutting, automatic conversion, template customization and quality control based on database.

Mapping Technology Requests High. The provincial geographic maps take the national 1:10 00000 fundamental geographic information database as main data source. Various provinces adopt different scales according to their actual range size. The majority of map scales are at 1:10 00000 or so, but the scale span is bigger in the rest of map scales, the biggest scale is 1:30000 (for example, Macau) and the minimum scale is 1:40 00000 (for example, Xinjiang and Tibet).

Therefore, the technology development difficulty is larger to realize automatically illustrating based on database. Furthermore, the technical difficulty and workload of map downsizing are also larger. In order to guarantee the system coordination between each sheet, need to formulate flexible cartographic generalization index system, use the combination way of automatic illustrating and human-computer interaction mapping, neatly master the editing-processing techniques of different maps by relying on experience.

Material Acquisition and Analysis Difficult. Firstly, the difficulty of data acquisition is very large, so need to arouse all enthusiasm and widely collected relevant data and information through various channels. Further, need to give more support from related state department so that provide the authority data of public use, to ensure information authority, data currency and content scientific.

Secondly, the workload of data analysis and processing is very heavy after obtaining materials. Due to some reasons such as data type complexity, data standard denormalization, data index disunity and so on, make the data processing work very difficult and hard, including data fusion, screening extraction, indexes determination, etc.

So the project will employ experienced mapping experts to guide on-site and strictly control the quality.

Chart Data Processing Complex. In coastal provinces especially Hainan province, the scope of sheet covers and includes sea area. But nowadays the chart data that can be gleaned from search queries is incomplete and its coverage is not whole too. Therefore, need to make integrated processing and analysis integration for multiple sources information from both at home and abroad.

In addition, the achievement of 927 first-phase construction has a huge amount of data quantity. The screening available data content from them is not easy, it is necessary to deeply thoroughly analyze data content and carefully choose reasonable index. Therefore, the data processing technology is very complex and its workload is also heavy.

5 Conclusion

The research and compiling of the new century version of National Huge Atlas of the People's Republic of China, is the first time to carry out research, design, edit and production on the state-level huge atlas in our country, using full new technology method. Each entity undertaking give full play to its own advantage of technology and data. National Geomatics Center of China, as the department of study, design, organization, implementation, construction and maintenance about national fundamental geographic information database, accumulated the rich data resources and technology research achievements and formed a set of technical system on database building and update. Successively finished rapid continued updating of 1:50000, 1:250000 and 1:10 00000 database one after another and realized the dynamic management and quick service on multi-scale, multi-type, multi-version database.

The studying and compiling of provincial geographic map in the national general atlas make full use of existing achievements from the national 1:10 00000 fundamental geographic information database. The national 1:10 00000 database has rich integrity data content, standard data specification, rigorous topology relationship and precision mathematical foundation which laid a very good foundation and design standard for database establishment of different province geographic map. At the same time, use many kinds of mature technology such as multi-scale automatic downsizing, element-level incremental updating, graphics-database linkage update, automatic illustrated based on database and so on, and carry out studying and innovating in cartographic generalization, cartographic symbols, visualization expression and so on, so that make the map products more meet the needs of the era development from content to form.

The development and research of the new century version of provincial geographic map use the technical route of "firstly building library, then compiling maps" and the results will be published online, so the final results include three kinds of form such as database, pre-printing data of electronic map and online map. All three are coordination and unity each other. They can mutually show geographical conditions in our country

from multiple levels and multiple points of view. They also can provide important basis to central and provincial leaders organ and competent business department for macro decision-making such as economic policy, long-term planning, rational allocation of productive forces, etc. Moreover, in order to adapt to the new norm and implement "four comprehensive", they can provide basic geographic information security for general layout.

References

1. Yi, C.: Study and Design of Modern Atlas. Science Press, Beijing (2005)
2. Liu, J.: Construction and update of the national fundamental geographic information database. Bull. Surv. Mapp. **10**, 1–3 (2015)
3. The Atlas of China. Map Publishing House of China, Version 2, January 2011
4. Wang, D., Shang, Y., Liu, J.: Technical discussion of database-driven topographic fast mapping. Geomat. World **2**, 6–9 (2012)
5. Guizhi, W.: The design and product of national 1:1000000 cartographic data of topographic map, pp. 1–5. IEEE Conference Publications. doi:10.1109/Geoinformatics.2015.7378691
6. Liao, K., Qi, Q., Chi, T.: The national natural atlas of China — development of electronic atlas. Earth Inf. Sci. **10**(03), 284–290 (2008)
7. Han, J.: The map visualization design oriented atlas compiling — take population and environmental change atlas of the People's Republic of China for example. J. Earth Inf. Sci. **12**(6), 777–783 (2010)
8. Kramers, E.: Interaction with maps on the internet — a user centred design approach for the atlas of Canada. Cartogr. J. **45**(2), 98–107 (2008)
9. Liao, K.: The milestone of Chinese modern cartography development — characteristics and innovation of China's national atlas. Geogr. Sci. Prog. **20**(03), 200–207 (2001)

Building Geospatial Health Applications from the EASTWeb Framework

Yi Liu[1(✉)], Michael D. DeVos[1], Muhammad Abdul-Ramin[1], and Michael C. Wimberly[2]

[1] Department of Electrical Engineering and Computer Science, South Dakota State University, Brookings, SD, USA
{yi.liu, Michael.devos, muhammad.abdulrahim}@sdstate.edu
[2] Geospatial Sciences Center of Excellence, South Dakota State University, Brookings, SD, USA
michael.wimberly@sdstate.edu

Abstract. Disease maps and forecasts developed using satellite remote sensing data can inform the decisions of public health officials and improve disease control and epidemic response. However, it is time consuming to construct a geospatial health system using remote sensing data products from scratch and it is quite expensive to maintain the system if changes are needed to the data products. We have developed an open-source Epidemiological Applications of Spatial Technologies (EASTWeb) framework to facilitate constructing applications to automate the retrieval, processing, and storage of satellite remote sensing data for public health research and applications. This paper briefly introduces the EASTWeb framework and uses a case study to demonstrate how the EASTWeb framework can be easily extended to be an application. The paper also illustrates the EASTWeb V2.0 system, which is extended from the framework and has been using in forecasting the West Nile Virus and Malaria epidemic risks in our research.

Keywords: Geospatial health applications · EASTWeb framework · Earth observation data stream

1 Introduction

Satellite remote sensing data provide environmental information about patterns of vegetation moisture, temperature and other spatially and temporally variable characteristics of the Earth's surface. Such information can be used in the study of human health [1]. It is a promising field to use satellite remote sensing data to develop disease models and inform the decisions of public health officials to improve the disease prediction and response [2, 3]. Accurate disease forecasts allow public health officials to intelligently deploy resources to prevent or reduce the effects of disease outbreaks. However, researchers and public health officials face a barrier to producing or accessing geospatial data, which requires geographical information studies (GIS) expertise and time consuming retrieval, processing and storage.

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 451–464, 2017.
DOI: 10.1007/978-981-10-3969-0_50

To respond to this need, we have developed open-source and client-based application Epidemiological Applications of Spatial Technologies (EASTWeb) V1.0 [4, 5] to facilitate public health researchers to access and process the earth science datasets. EASTWeb V1.0 automatically connects to earth science data archives (MODIS land surface temperature product MOD11A2, nadir BRDF-adjusted reflectance product MCD43B4 from the Land Processes Distributed Active Archive Center [6], and TRMM products 3B42 and 3B42RT from the Goddard Earth Sciences Data and Information Services Center [7]), and acquires, processes, and summarizes selected remote sensing datasets based on the time period and geographic extent that the user provides. The environmental data summaries are stored in a relational database that can be easily queried and linked to ecological and epidemiological datasets for analysis and forecasting in software environments such as R. EASTWeb V1.0 has been used to facilitate the forecast of epidemic risks of Malaria and West Nile Virus [4].

We believe that EASTWeb application could be used more broadly to support the integration of remotely-sensed environmental data into public health research. Different earth science data archives have different data structures, access methods, and file formats and requires different data processing methodologies. It is very expensive to update EASTWeb V1.0 if changes are required to upgrade the user interfaces, implement new data products and refactor the implementation of the processing steps. Thus, we expanded EASTWeb V1.0 to a customizable plugin framework to allow developers to easily build geospatial health application by using different earth observation data products [8, 9].

This paper describes how to build geospatial health application from the EASTWen framework we constructed. Section 2 introduces the architecture, implementation and the user types of the EASTWeb framework. Section 3 uses a case study to illustrate how to use the framework to develop applications. Section 4 demonstrates the EASTWeb V2.0 application that was extended from the framework. Section 5 evaluates the current work and points to the directions for future research.

2 EASTWeb Framework

Four major steps [4, 8] illustrated in Fig. 1 are generalized to illustrate the process of earth observation data stream products.

The Downloading step accesses data stream files from online earth observation archives and stores them locally for further processing.

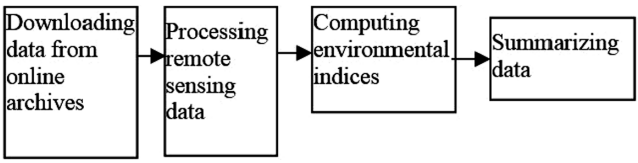


Fig. 1. The major processing steps.

The Processing remote sensing data step contains sub- steps such as mosaicking, converting, compositing, data filtering, masking, clipping and reprojecting. For tiled data mosaicking is needed for combining multiple tiles into a single geospatial data layer. Converting is applied to data whose data format cannot be read directly by the GDAL library. Compositing combines multiple data files into a single data layer. Data filtering screens each data value and converts "bad" data to a NoData value. Masking applies the mask grid and converts water pixels to NoData. Clipping subsets the spatial extent of the data to a smaller area of interest. Reprojecting puts all data into the same geographic projection as the zone dataset. The resulting data layers are stored in GeoTiff format after the processing step.

The Computing environmental indices step calculates different indices for each data product. For example, the MODIS reflectance product requires calculating standard spectral indices such as the normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), soil- adjusted vegetation index (SAVI), and various forms of the normalized difference water index (NDWI). The environmental index data are stored in GeoTIFF format.

The Summarizing data step generates spatial and temporal summaries of the environmental indices. We assume that spatial summaries are mandatory but temporal summaries are optional. A spatial summary results from the statistics computation based on the user-supplied zone shapefile. Zones typically represent counties, districts, census tracts, zip. Temporal periods are day, WHO epi-week, CDC epi-week, month, the data archive's native temporal resolution and others customized by users.

2.1 The Architecture of EASTWeb Framework

A framework is a skeleton of an application that can be customized by an application developer efficiently [10]. The difficult part of the framework design is the identification of the abstractions that can be tailored to specific applications within the framework domain to support the customization.

In the EASTWeb framework, each of the four processing steps is designed to be a sub-framework to support flexibilities. As shown in Fig. 2 [9], the four sub-frameworks Downloading, RS Data Processing, Environmental Indices, and Summarization are designed to map to the four major steps.

In the EASTWeb framework, a plugin customizes the four sub-frameworks to be concrete components by implementing the processing steps for a data stream product. For example, the TRMM 3B42RT plugin will customize the Download sub-framework for downloading TRMM 3B42RT dataset from Goddard center, specialize the RS Data Processing sub-framework to process the data (including converting, clipping, masking and reprojecting), implement the calculation methods in the Environmental Indices sub-framework and generate summaries in the Summarization sub-framework. In addition to the code of customizing the four sub-frameworks for processing a specific data stream product, each plugin contains an xml storing associate metadata information, such as download method (e.g., FTP, HTTP, etc.) and associated URI, the

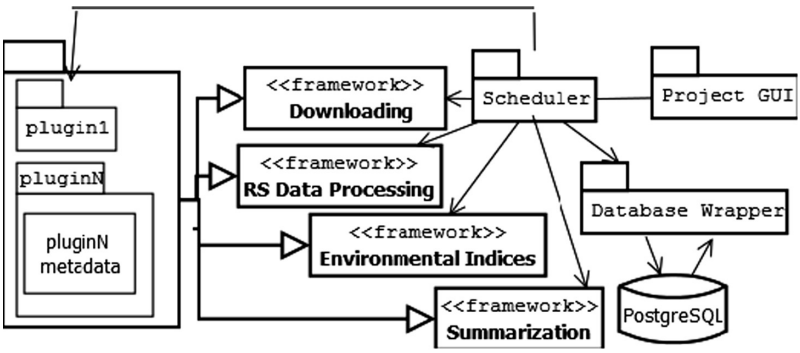


Fig. 2. The architecture of EASTWeb framework.

process task dependencies, the calculations that need to be performed to compute the environmental indices, and other information such as quality control levels defined for the data stream product.

2.2 The Implementation of the EASTWeb Framework

The EASTWeb framework is coded using JAVA for overall system control and using Java Swing for the Graphical User Interfaces. The Geospatial Data Abstraction Library (GDAL) [11] is used to carry out the spatial analyses and PostgreSQL is used to store and manipulate the resulting data summaries. Each sub-framework is held in a java package. Components Scheduler, Project GUI and Database Wrapper are in java packages as well. There are several other packages holding the additional information, for example, a PluginMetaData package for gathering the plugin metadata parser and the plugin metadata file.

The EASTWeb framework is open source. We have released the source code of the framework at https://github.com/eastweb/EastWeb.V2.

Constructing an application from customizing the EASTWeb framework involves building a plugin for each data stream product that the application supports, and assembling the plugins in the application. For example, if an application supports TRMM 3B42RT product and MODIS MOD11A2 product, plugins for these two products should be constructed. EASTWeb framework provides flexibilities to the requirement changes in applications. If an application needs to support a new data stream product, it will simply have a plugin for the product implemented and added in. If an existing data stream product needs to be removed from an application, the associated plugin can be removed.

2.3 The Types of the Users

The users of the EASTWeb framework are categorized into three types: the plugin developers, the application level developers and the end users.

A plugin provider develops a plugin for a data stream product by extending the four sub-frameworks in the EASTWeb framework. This type of the user should have experience in both geospatial sciences and computer programming using the JAVA language. For example, the plugin provider A builds a MODIS Land Surface Temperature (LST) plugin, a MODIS BRDF-corrected reflectance and a TRMM 3B42RT plugin.

An application developer uses the available plugins to construct a concrete application. This type of the user should have the knowledge of Java packages and the application configurations. For example, the application developer B constructs an EASTWeb 1.3 application by adding the MODIS LST plugin, and the TRMM 3B42RT plugin in to the framework.

An end user uses a concrete application to create and run the projects. This type of the user is a professional scientists or technicians who have experience with geospatial datasets. He should be familiar with the key concepts such as projections and coordinate systems. For example, the end user C uses the EASTWeb 1.3 application to create and run a project for summarizing the environmental indices in the northern Great Plains to facilitate the forecasts of West Nile Virus risk.

3 Case Study

This case study demonstrates how to build an application with a plugin of TRMM 3B42RT [12] product.

3.1 Building a Plugin

A plugin is built by the plugin provider. The steps for building a plugin of a data stream product are (1) defining the plugin information in the plugin metadata file, and (2) customizing sub-frameworks.

As described in Sect. 2.1, a plugin metadata file describes the key information on the processing steps that a plugin has. XML elements <Download>, <Process>, <Indices> and <Summary> hold the information for the four sub-frameworks. The download method (e.g., FTP, HTTP, and etc.), associated URI (hostname and root directory), and credential information (username and password) are needed for the Download sub-framework and they are described in the elements <mode>, <hostname>, <rootDir>, <username> and <password> nested in element <download>. If a sub-framework is customized, the name of the class that holds the implementation needs to be specified in the element <className> inside the sub-framework tag. For example, TRMM3B42RTDownloader is the Java class that extends the Downloader sub-framework, so it is specified in the <className> inside the <Download> as shown in Fig. 3.

TRMM 3B42RT needs to have sub-steps converting, reprojecting, clipping and masking (in this order) in the Processing remote sensing data step. As shown in Fig. 3, in the <Process> element, the name of the classes implementing those sub-steps while

```
<?xml version="1.0" ?>
<PluginMetadata>
    <Title>TRMM3B42RT</Title>
    <Download>
        <DownloadFactoryClassName>
            TRMM3B42RTFactory
        </DownloadFactoryClassName>
        <Mode>ftp</Mode>
        <FTP>
            <HostName>disc2.nascom.nasa.gov</HostName>
            <RootDir>
              /data/TRMM/Gridded/Derived_Products/3B42RT/Daily/
            </RootDir>
            <UserName>anonymous</UserName>
            <PassWord>anonymous</PassWord>
        </FTP>
        <TimeZone>CST6CDT</TimeZone>
        <FilesPerDay>1</FilesPerDay>
        <DatePattern>\d{4}</DatePattern>
        <FileNamePattern>
            3B42RT dailv\.(\d{4})\.(\d{2})\.(\d{2})\.bin
        </FileNamePattern>
    </Download>
    <Processor>
        <ProcessStep>TRMM3B42RTConvert</ProcessStep>
        <ProcessStep>TRMM3B42RTReproject</ProcessStep>
        <ProcessStep>TRMM3B42RTClip</ProcessStep>
        <ProcessStep>TRMM3B42RTMask</ProcessStep>
        <NumberOfOutput>1</NumberOfOutput>
    </Processor>
    <Indices>
        <ClassName>TRMM3B42RTIndex</ClassName>
    </Indices>
    <Summary>
        <Temporal>
                <MergeStrategyClass>
                    SummationGdalRasterFileMerge
                </MergeStrategyClass>
            </Temporal>
    </Summary>
    <QualityControl/>
    <ExtraInfo>
        <Tiles>false</Tiles>
    </ExtraInfo>
</PluginMetadata>
```

Fig. 3. The plugin metadata file for TRMM3B42 plugin.

customizing the RS Processing sub-framework are specified in the <className> elements. The order of the class names matches the order of the sub-steps.

TRMM3B42RT plugin needs to customize Downloading sub-framework, RS data processing sub-framework and Environmental indices sub-framework. It uses the implemented temporal periods in the Summarization sub-framework and will not add new ones, so it will not customize the Summarization sub-framework. The customized

class(es) for each sub-framework should be organized into a package with the plugin's name and further be placed in each sub-framework package.

Figure 4 shows the code of class TRMM3B42RTFactory that customizes the Downloading sub-framework. Figure 5 shows the partial code of class TRMM3B42RT Convert which extends the Convert in the RS data processing sub-framework. EASTWeb framework has implemented Clipping, Masking and Reprojecting steps in RS Data Processing sub-framework. Since TRMM3B42RT plugin does not need extra customizations on these steps, it just inherits the implementations from the framework.

As for example, Fig. 6 illustrates the code for the Clipping step in TRMM3B42RT. The implementations of the RS processing steps of TRMM3B42RT are placed in a package named "TRMM3B42RT" inside the "Processor" package (holding classes for

```
public class TRMM3B42RTFactory
        extends DownloadFactory {

    public TRMM3B42RTFactory(Config configInstance,
       ProjectInfoFile projectInfoFile,
       ProjectInfoPlugin pluginInfo,
       DownloadMetaData downloadMetaData,
       PluginMetaData pluginMetaData, Scheduler scheduler,
       DatabaseCache outputCache, LocalDate startDate) {
        super(configInstance, projectInfoFile, pluginInfo,
         downloadMetaData, pluginMetaData, scheduler,
         outputCache, startDate);
    }

    @Override
    public DownloaderFactory
        CreateDownloaderFactory(ListDatesFiles listDatesFiles) {
        return new LocalStorageDownloadFactory(configInstance,
           "TRMM3B42RTDownloader", projectInfoFile, pluginInfo,
           downloadMetaData, pluginMetaData, scheduler,
           outputCache, listDatesFiles, startDate);
    }

    @Override
    public ListDatesFiles CreateListDatesFiles()
        throws IOException {
        return new TRMM3B42RTListDatesFiles(new DataDate(startDate),
        downloadMetaData, projectInfoFile);
    }
}
```

Fig. 4. Code of TRMM3B42RTFactory.java.

RS data processing sub-framework). Figure 7 demonstrates how the Environmental indices sub-framework is customized for TRMM 3B42RT product.

```
public class TRMM3B42RTConvert extends Convert {
    private Integer noDataValue;

    public TRMM3B42RTConvert(ProcessData data,
            Boolean deleteInputDirectory) {
        super(data, deleteInputDirectory);
        noDataValue = data.getNoDataValue();
    }
    @Override
    protected void convertFiles() throws Exception{
        GdalUtils.register();
        synchronized (GdalUtils.lockObject) {
            //set xSize and ySize
            ...
            for (File f:inputFiles)
            {
                DataInputStream dis = new DataInputStream(new FileInputStream(f));
                // set filename
                ...
                Dataset outputDS = gdal.GetDriverByName("GTiff").Create(
                        mOutput.getPath(),
                        xSize, ySize,
                        1,
                        gdalconstConstants.GDT_Float32
                        );

                double[] array = new double[xSize];
                int row = 0, col = 0;
                for (row=0; row<ySize; row++)
                {
                    for (col=0; col<xSize; col++)
                    {
                        array[col] = dis.readFloat();
                    }
                    outputDS.GetRasterBand(1).WriteRaster(0, row, xSize, 1, array);
                }
                dis.close();
                outputDS.GetRasterBand(1).SetNoDataValue(noDataValue);
                // set GeoTransform, set projection
                ...
                outputDS.delete();
            }
        }
    }
}
```

Fig. 5. Partial code of TRMM3B42RTConvert.java.

```
public class TRMM3B42RTClip extends Clip{
    public TRMM3B42RTClip(ProcessData data,
            Boolean deleteInputDirectory) {
        super(data, deleteInputDirectory);
    }
}
```

Fig. 6. Code of TRMM3B42RTClip.java.

```
public class TRMM3B42RTIndex extends IndicesFramework
{
    private final int INPUT = 0;
    public TRMM3B42RTIndex(List<File> inputFiles, File outputFile,
            Integer noDataValue) {
        super(inputFiles, outputFile, noDataValue);
    }

    @Override
    protected double calculatePixelValue(double[] values)
            throws Exception {
        if (values[INPUT] == noDataValue) {
            return noDataValue;
        } else {
            return values[INPUT];
        }
    }

    @Override
    protected String className() {
        return getClass().getName();
    }
}
```

Fig. 7. Code of TRMM3B42RTIndex.java.

3.2 Building an Application from Plugins

The EASTWeb framework has built the graphical user interfaces (GUI) for the applications and controller to cooperate all the components to work together.

An application developer can construct an application from implemented plugins easily. The actions needed are: (1) to place the code of the plugins into the right packages in the framework and (2) to bundle all the codes into an executable jar file for delivery. The plugin metadata files should be placed in the PluginMetaData package provided by the framework. The customized code for each sub-framework should be placed in the package with the plugin's name under each sub-framework's package. An executable jar file for the codes can be easily made through a Java IDE (e.g., Eclipse [13], which is used in the EASTWeb framework development) or through the jar command in a console.

4 EASTWeb V2.0 System

As an enhancement to the EASTWeb V1.0 system, we have developed EASTWeb V2.0 system containing 6 plugins (a MODIS Land Surface Temperature (LST) plugin, a North American Land Data Assimilation System (NLDAS) [14] NOAH model plugin, a NLDAS forcings plugin, a MODIS BRDF-corrected reflectance, a TRMM 3B42 plugin, and a TRMM 3B42RT plugin) is under test. The 6 plugins were developed by 4 different individuals with different programming background. 2 developers are new to Java programming, one developer has moderate experience with Java and another is an experienced Java programmer. Each plugin requires relatively small amount of coding because the framework provides considerable amount of reusable code. From extending the EASTWeb framework, only 3,600 extra lines of code were used to program the 6 plugins and thus complete the implementation of the

EASTWeb V2.0 system. The code of the EASTWeb V2.0 system is released along with the source code of the EASTWeb framework at https://github.com/eastweb/EastWeb.V2.

The end user can select the plugins from the plugin pool of the application while creating a new project as shown in Fig. 8.

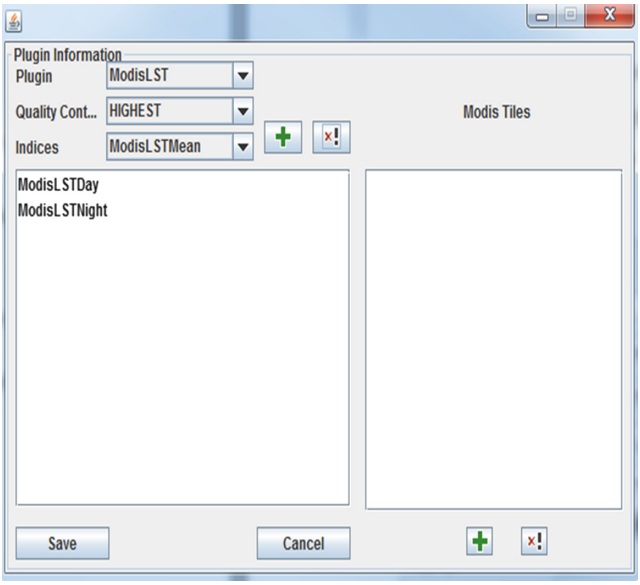


Fig. 8. Create a new project – choosing plugins.

As for each plugin, the end user can choose to calculate the environmental indices from the indices list that the plugin provides. As shown in Fig. 9, the end user needs to provide the basic project information such as the start date, project name, and working directory and such. He also can choose to apply a mask grid to exclude some area from summarization and to clip the data to focus on the area of interest. The user also needs to set the projection information and choose the spatial summary units and temporal summary period. In the main window as shown in Fig. 10, the end user can select to run existing project(s) and choose whether to save the intermediate files for each project. Each project can be stopped or deleted. The user can use the query tool as shown in Fig. 11 to retrieve the summary results from the database.

We have been using EASTWeb system to integrate with the R environment to carry out modeling and mapping and have been extensively tested through applications to support mosquito-borne disease forecasting for West Nile virus in the United States and epidemic malaria in the highlands of Ethiopia. EASTWeb V2.0 is now working as a core component in the Epidemic Prognosis Incorporating Disease and Environmental Monitoring for Integrated Assessment (EPIDEMIA) system [15].

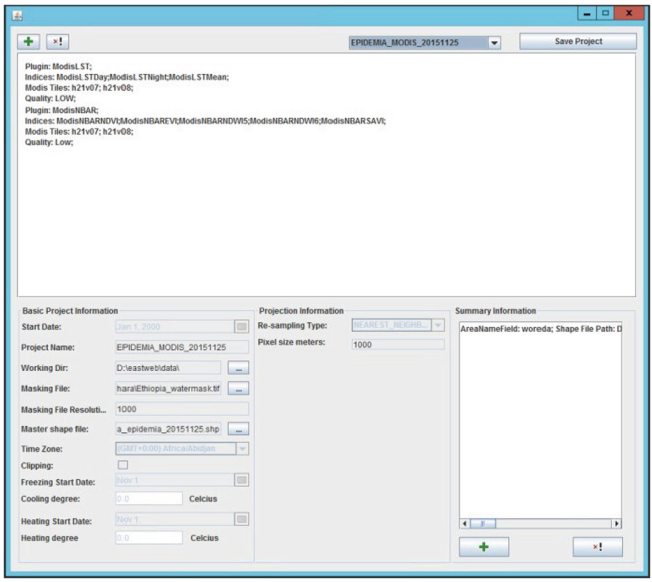


Fig. 9. Create a new project – adding project information.

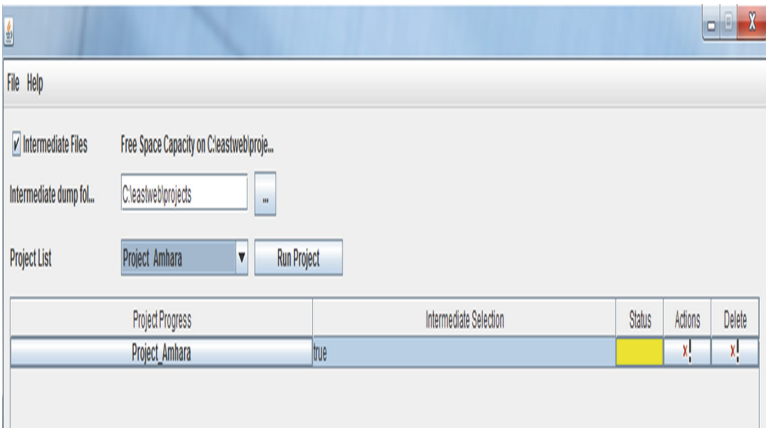


Fig. 10. Project main window.

As an example, Fig. 12 illustrates the results of using MODIS LST and TRMM plugins of the EASTWeb V2.0 system in forecasting the malaria cases in Dera Woreda in the Amhara Region of Ethiopia based on the data retrieved from May 9th to May 16th, 2016. The data indicated that Dera Woreda had exceeded the outbreak detection threshold in 2 weeks of the past 6 weeks. The malaria cases were forecast to exceed the outbreak threshold in 4 weeks of the next 4 weeks.

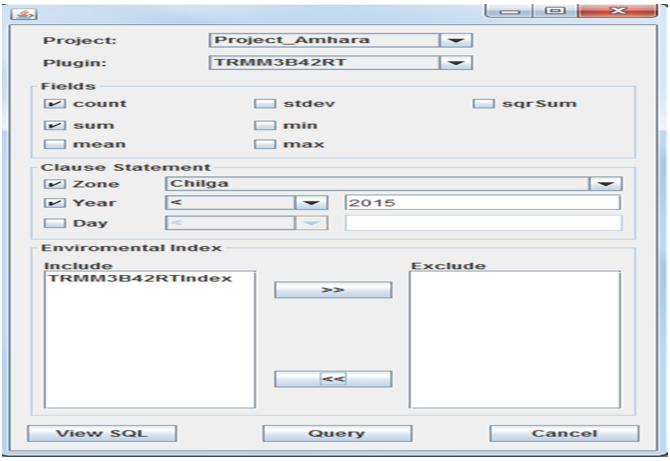


Fig. 11. Summary query window.

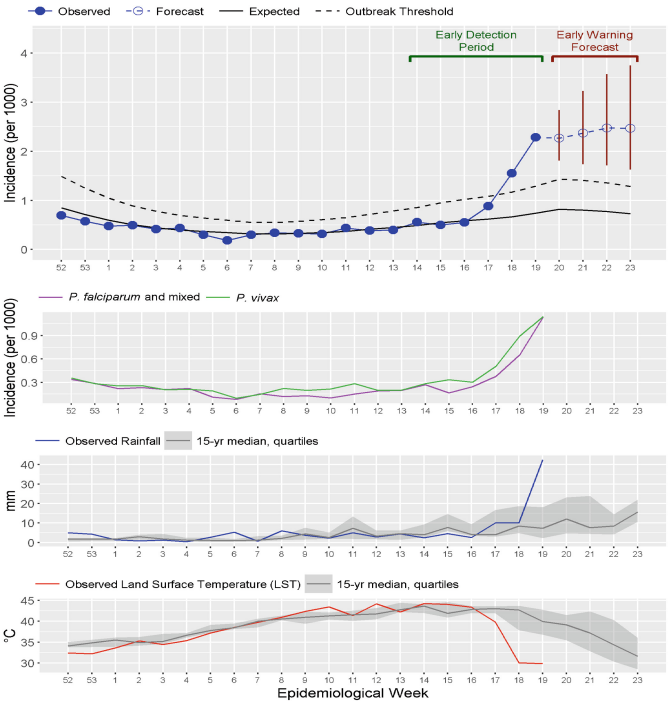


Fig. 12. Malaria early detection and early warning for Dera Woreda.

5 Discussion

It is time consuming to construct geospatial health applications from scratch. The EASTWeb framework is designed and constructed to allow geospatial health application developers to easily build applications by using different earth observation data products. The framework provides reusability and flexibility in constructing concrete applications. The development of a plugin needs only a small amount of coding since the framework provides considerable amount of reusable code. Building a concrete application is straightforward, and requires assembling the available plugins using a simple strategy. Introducing plugins in the framework also make an EASTWeb application flexible in changing requirements. For example, TRMM is going to be replaced by the Global Precipitation Measurement (GPM) [16]. Replacing TRMM by GPM in an application can be simply done by removing the TRMM plugin and adding the GPM plugin in.

There are many opportunities for improving and enhancing EASTWeb framework. There are several areas which will be focused on in the near future. First, the Application Programming Interface (API) of the EASTWeb framework, sample code, and tutorial of using the framework will be released along with the code. Second, the current EASTWeb framework is hosted on a single computer. As a direction of future work, EASTWeb framework will be migrated into a cloud environment to improve the performance of the execution.

Acknowledgments. This work is supported by NASA ACCESS grant NNX14AI37A "Expanding Earth Science Data Access for Public Health Research and Applications" and NIH grant R01AI079411 "An Integrated System for the Epidemiological Application of Earth Observation Technologies".

References

1. Ford, T.E., Colwell, R.R., Rose, J.B., Morse, S.S., Rogers, D.J., Yates, T.L.: Using satellite images of environmental changes to predict infectious disease outbreaks. Emerg. Infect. Dis. **15**, 1341–1346 (2009)
2. Chuang, T., Wimberly, M.C.: Remote sensing of climatic anomalies and west nile virus incidence in the northern great plains of the United States. PLoS ONE **7**, e46882 (2012)
3. Midekisa, A., Senay, G., Henebry, G.M., Semuniguse, P., Wimberly, M.C.: Remote sensing-based time series models for malaria early warning in the highlands of Ethiopia. Malaria J. **11**, 165 (2012)
4. Liu, Y., Hu, J., Snell-Feikema, I., VanBemmel, M.S., Lamsal, A., Wimberly, M.C.: Software to facilitate remote sensing data access for disease early warning systems. Environ. Model. Softw. **74**, 247–257 (2015)
5. EASTWeb site. https://epidemia.sdstate.edu/eastweb/
6. Land Processes Distributed Active Archive Center (LP DAAC). https://lpdaac.usgs.gov
7. Goddard Earth Sciences Data and Information Services Center (GES DISC). http://disc.sci.gsfc.nasa.gov/

8. Liu, Y., Wimberly, M.C., Hu, J.: On the construction of EASTWeb framework - a plug-in framework for processing earth observation data streams. In: Proceedings of the IEEE International Conference on Electro/Information Technology (IEEE EIT 2014), Milwaukee, WI, USA, June 2014
9. Liu, Y., DeVos, M., Hu, J., Abdul-Ramin, M., Wimberly, M.C.: EASTWeb framework– a plug-in framework for constructing geospatial health applications. In: Proceedings of the IEEE International Conference on Electro/Information Technology (IEEE EIT 2016), Grand Forks, ND, USA, May 2016
10. Schmid, H.A.: Framework design by systematic generalization. In: Fayad, M.E., Schmidt, D.C., Johnson, R.E. (eds.) Building Application Frameworks, pp. 353–378. Wiley, Hoboken (1999)
11. Geospatial Data Abstraction Library (GDAL). http://www.gdal.org
12. Tropical Rainfall Measuring Mission (TRMM). http://trmm.gsfc.nasa.gov/
13. Eclipse. https://eclipse.org/
14. North American Land Data Assimilation System (NLDAS). http://ldas.gsfc.nasa.gov/
15. Wimberly, M.C., Henebry, G.M., Liu, Y., Senay, G.B.: EPIDEMIA - an EcoHealth informatics system for integrated forecasting of malaria epidemics. In: Proceedings of the 7th International Congress on Environmental Modelling and Software. International Environmental Modelling and Software Society (iEMSs 2014), San Diego, CA, USA (2014)
16. Global Precipitation Measurement. http://www.nasa.gov/mission_pages/GPM/main/

Ship Navigation and Warning System Based on GPS/BDS Equivalent Satellite Clock Error Method

Dongjian Cai[1,2(✉)], Zhanyong Fan[1,2], Zongkun Zhen[1], and Wanghui Zhou[1]

[1] Engineering Technology Center, Suzhou Industrial Park Surveying, Mapping and Geoinformation Co., Ltd., Suzhou, Jiangsu, China
{caidj, fanzy, zhenzk, zhouwh}@dpark.com.cn
[2] College of Surveying and Geo-informatics, Tongji University, Shanghai, China

Abstract. The Automatic Identification System (AIS) uses a single GPS system and the accuracy of positioning is limited by the broadcast ephemeris, which is around 10 m. Thus, AIS is not competent for precise applications like ships in the crowded waterways. This contribution develops an improved navigation and warning system by integrating GPS and BeiDou satellite (BDS) together and calculating equivalent satellite clock errors of different constellation. Users can receive satellite clock corrections w.r.t broadcast ephemeris via Frequency Modulation (FM) or Very high Frequency (VHF) technology. The final result shows that accuracy of real-time positioning is within 1 m.

Keywords: Multi-GNSS · Navigation and warning · AIS · Equivalent satellite clock error

1 Introduction

Over the past decade, Global Navigation Satellite System (GNSS) has experienced a dramatic development. Starting from two full operation constellations (GPS and GLONASS), a set of five global or regional navigation satellite systems BeiDou, Galileo, Quasi-Zenith Satellite System (QZSS) and the Indian Regional Navigation Satellite System (IRNSS) are offering, or at least preparing positioning, navigation and timing (PNT) service [1–3]. The booming of GNSS technology produces itself a more and more important character in transportation industry. The most popular navigation and warning system AIS is an automatic tracking system used on ships for identifying and locating vessels by electronically exchanging data with other nearby ships, AIS base stations, and satellites [4]. Currently, AIS uses only GPS satellite, and as the developing of our own GNSS BeiDou satellite (BDS), most of the ships are required to integrate BDS with GPS. Multi-GNSS can indeed offer numerous advantages over stand-alone GPS navigation. However, the precision of positioning is still limited by the accuracy of broadcast products if no corrections are added. Table 1 shows precision

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 465–472, 2017.
DOI: 10.1007/978-981-10-3969-0_51

Table 1. Root-mean-square orbit errors of broadcast orbits relative to precise ephemerides (unit: m).

System	Radial	Along	Cross	T	R-T
GPS IIA	0.26	1.21	0.37	1.10	1.07
GPS IIR	0.14	1.04	0.42	0.52	0.51
GPS IIF	0.14	0.75	0.32	0.28	0.32
GLONASS	0.35	2.41	1.33	1.90	1.93
Galileo	0.63	2.65	2.29	1.62	1.58
BDS(MEO + IGSO)	0.50	2.42	1.31	0.87	0.99

of broadcast orbits and clocks of different constellation by analyzing the whole year data in 2014 [5].

It illustrates that GPS new generation IIF satellite owns a better orbit than the former two generation IIA and IIR, which the 3D RMS is within 1 m. The following GLONASS satellite has an orbit better than 2 meters. The new emerging BeiDou (MEO + IGSO) orbit is a little better than Galileo, which has a 3D accuracy of 2.8 m and 3.6 m respectively.

Obviously, broadcast products cannot satisfy precise applications, especially those based on real-time kinematic point positioning, like AIS. In order to improve the capability of AIS, this contribution calculates equivalent satellite clock errors of GPS and BDS, disseminates satellite clock corrections to users to achieve a better precision.

2 Principle of Equivalent Satellite Clock Errors

In a statistical sense, the point positioning accuracy of GNSS may be described by the product of the user equivalent range error (UERE), which combines all measurement and modeling errors of an individual pseudorange, and the dilution of precision (DOP), which maps such errors to the position uncertainty for a given number and geometric distribution of observed GNSS satellites [6]. UERE is mostly affected by the radial direction of orbits and satellite clocks. Suppose that there is a network with known station positions, we fix the broadcast satellite orbits and precise position of stations, and then calculate satellite clock corrections w.r.t broadcast satellite clocks. The corrected satellite clocks assimilate both satellite orbit errors that contributed to radial direction and satellite clock errors. So, users inside the network can obtain a better accuracy of positioning by adding satellite clock corrections.

Pseudorange observation equation of GPS and BDS is

$$P^{G,C} = \rho^{G,C} + c \cdot dt_r - c \cdot dt^{G,C} + I^{G,C} + T + ISB + Multi^{G,C} \tag{1}$$

Where the upper index G and C denote GPS and BDS, $P^{G,C}$ means pseudorange observation, $\rho^{G,C}$ indicates geometry distance, c is the speed of light, dt_r is the receiver offset, $dt^{G,C}$ is the satellite clock offset, $I^{G,C}$ is the ionosphere delay, T is the troposphere delay, ISB is inter-system bais and $Multi^{G,C}$ is the multipath affection.

As explained in Sect. 2, $\rho^{G,C}$ is calculated by fixing the known station positions and broadcast orbits. dt_r and $dt^{G,C}$ are related, and need to set a reference in solution. $I^{G,C}$ is usually eliminated by LC combination if dual-frequency measurements are accessible. A priori value of T is calculated by model, and the left part is estimated. ISB is estimated as a constant parameter, which also needs a reference [7, 8].

Error equation based on Eq. (1) is:

$$v_k = A_k \delta x_k - l_k \tag{2}$$

Where v_k is a vector of observation errors, A_k is the coefficient matrix, δx_k is the vector of variables that are to be estimated, l_k is the observation. Recall the least square performance functional:

$$J(x) = \|A_i \delta x - l_i\|_2 = \min \tag{3}$$

and let H, H be an orthogonal matrix. Because of property of orthogonal matrix [9], we can write:

$$J(x) = \|HA_i \delta x - Hl_i\|_2 = \min \tag{4}$$

In fact, $J(x)$ is independent of H and this can be exploited. For an arbitrary matrix $A \in R^{m \times n} (m \geq n)$, there exists an orthogonal transformation $H \in R^{m \times m}$ such that

$$HA = \begin{bmatrix} s & \vdots & \\ & \vdots & \tilde{A} \\ 0 & \vdots & \end{bmatrix} \tag{5}$$

Where s and $\tilde{A}$ are computed directly from A, and the matrix H is only implicit, computer mechanization requires no additional computer storage other than that used for A. So formula (4) can write as [10]

$$J(x) \left\| \begin{bmatrix} R \\ 0 \end{bmatrix} \delta x - \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \right\|_2 = \min \tag{6}$$

Where $R \in R^{n \times n}$, is an upper triangular matrix, $Hl_i = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$, $z_1 \in R^n$, $z_2 \in R^{m-n}$. Then, one can see by inspection that the minimizing δx must satisfy

$$R \delta x - z_1 = 0 \tag{7}$$

These results are more elegant than is the brute force construction via the normal equation. More importantly, the solution using orthogonal transformation is less susceptible to errors due to computer roundoff.

3 Experiment and Analysis

3.1 Data Streams

To support the real-time service the IGS Real-time Pilot Project now extends its capability to 118 tracking stations from their real-time GNSS networks, which includes 118 GPS, 111 GLONASS, 54 Galileo, and 29 BeiDou data streams (www.igs.org). After joining RTCM in 2008, the IGS real-time project adopted the RTCM-3 format for GPS and GLONASS observation messages. In addition to a format a transport protocol had to be defined, the so-called Ntrip streaming protocol (BNC), developed at BKG together with TU Dortmund is adopted for achieving real-time data streams.

Figure 1 shows the IGS real-time monitoring stations that include both GPS and BDS observations, which has a total number of 19 sites. We choose the time range from 2014 doy28 to doy34.

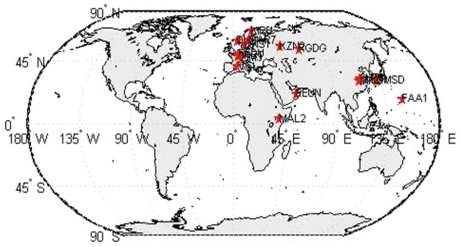


Fig. 1. Distribution of stations.

3.2 Precision of Satellite Clock After Correction

We first test the accuracy of estimated GPS and BDS satellite clocks after implementing equivalent satellite clock errors.

Figure 2 shows the accuracy of different GPS and BDS satellite clock, there is a little bit difference between satellites even in the same constellation. That might due to the number of observations and also satellites that experience eclipse. Figure 3 shows the mean precision of GPS and BDS satellite clocks over one week, the accuracy is 1.3 ns and 2.5 ns respectively, which relates tightly to the quality of pseudorange observations.

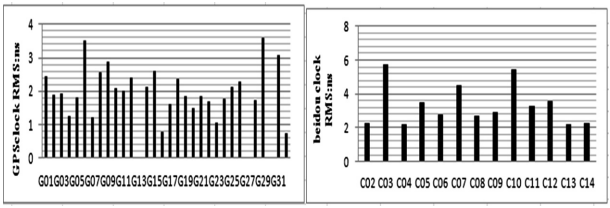


Fig. 2. Precision of GPS and BDS satellite clocks in doy 028.

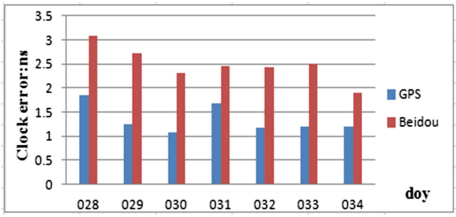


Fig. 3. Precision of all the GPS and BDS satellite clocks in one week.

3.3 Accuracy of Kinematic Positioning

This part we will test the stand-alone GPS and BDS that with and without adding satellite clock correction and also GPS/BDS integrated case. We choose a Multi-GNSS monitoring sites CUT0, and first test only using one system in doy 028.

The left side of Fig. 4 shows GPS kinematic positioning using broadcast ephemeris, precision in NEU is 1.659 m, 1.373 m and 3.906 m. The right side shows GPS kinematic positioning adding satellite clock corrections, precision in NEU is 1.372 m, 1.142 m and 2.825 m.

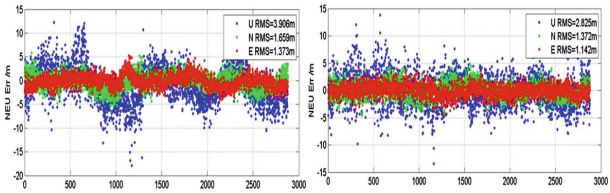


Fig. 4. GPS kinematic positioning using broadcast ephemeris and corrected clocks.

The left side of Fig. 5 shows BDS kinematic positioning using broadcast ephemeris, precision in NEU is 3.593 m, 7.492 m and 6.488 m. The right side shows BDS kinematic positioning adding satellite clock corrections, precision in NEU is 1.705 m, 0.992 m and 3.521 m.

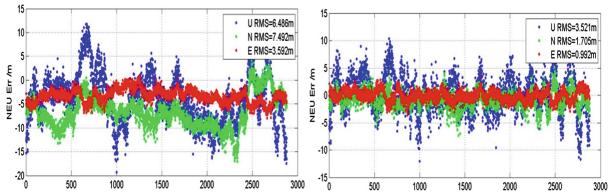


Fig. 5. BDS kinematic positioning using broadcast ephemeris and corrected clocks.

The first conclusion we can do is that satellite clock corrections can improve the kinematic positioning precision for both GPS and BDS, and especially for BDS.

The left side of Fig. 6 shows GPS/BDS kinematic positioning using precise products, precision in NEU is 0.967 m, 0.738 m and 2.265 m. The right side shows BDS kinematic positioning adding satellite clock corrections, precision in NEU is 0.946 m, 0.718 m and 2.246 m.

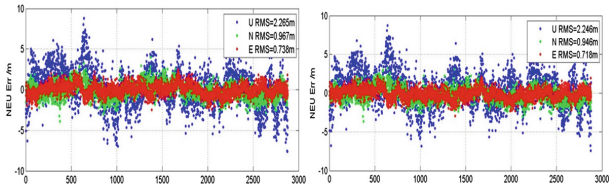


Fig. 6. GPS/BDS kinematic positioning using precise products and corrected clocks.

The second conclusion we can do is that Multi-GNSS can indeed improve the accuracy for kinematic positioning, and satellite clock correction can almost get the same accuracy as precise products. Table 2 shows the horizon precision of one week.

Table 2. Statistic of horizon precision.

Doy	GPS broadcast ephemeris	GPS satellite clock correction	Beidou broadcast ephemeris	Beidou satellite clock correction	GPS + Beidou precise ephmeris	GPS + Beidou satellite clock correction
028	2.153	1.785	8.308	1.973	1.232	1.188
029	2.343	1.883	8.024	1.856	1.100	1.166
030	2.483	1.798	4.556	1.793	1.033	1.112
031	2.303	1.727	3.143	1.549	0.877	0.945
032	2.550	1.824	4.332	1.679	0.991	1.033
033	3.001	1.982	5.565	1.409	1.221	1.173
034	2.167	1.691	4.354	1.605	0.910	0.998
Average	2.429	1.813	5.469	1.695	1.052	1.088

4 Navigation and Warning Software

Based on the mathematical method and experiment result, we developed a ship navigation and warning system using Visual C# and Arc Engine. The main function contains Mapping&Inquiring, information collection and broadcasting, ship navigation and warning, etc. The user table is shown as Fig. 7.

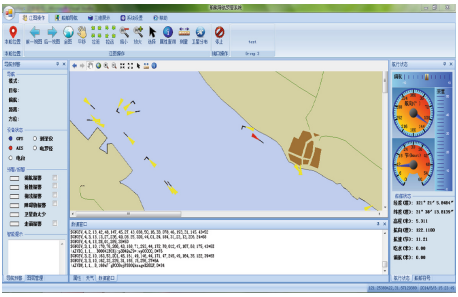


Fig. 7. Ship Navigation and warning software.

In order to test the effectiveness of our software, we did a test near the ChongMing port in ShangHai. We put AIS antenna and our software antenna together on the ship, and separated by a distance of 5 m. The observation rate was set to 1 s, for a short period, the moving the ship could be treated as uniform.

Figure 8 is the trajectory of AIS result (red color) and our software result (black color). We can find that our software result is more close to a line, which is true for the moving of ship in a short period.

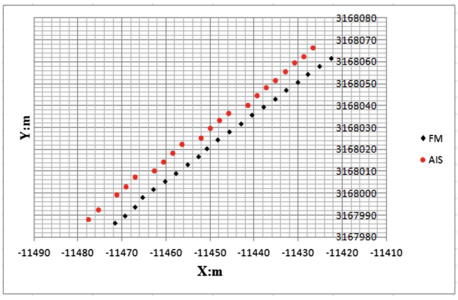


Fig. 8. Comparison of positioning accuracy. (Color figure online)

5 Conclusion

This paper aims at the poor kinematic positioning accuracy in AIS, develops an improved system that based on calculating equivalent satellite clock errors and integrating GPS and BDS. We did an experiment and proved that the improvement in 3D direction of our method is 55.2% by comparing to AIS result. The final accuracy of positioning in our software is at the level of 1 m.

References

1. Dach, R.: Bernese GNSS Software: New Features in Version 5.2. Astronomical Institute, University of Bern, Bern, Switzerland (2013)
2. Kaplan, E.D., Hegarty, C. (eds.): Understanding GPS: Principles and Applications. Artech House, Norwood (2005)
3. Galileo ICD. Galileo open service, signal in space interface control document (OS SIS ICD) (2008)
4. Filjar, R., Desci, S., Pokrajac, D., et al.: Internet AIS. J. Navig. **58**, 197–206 (2005)
5. Montenbruck, O., Steigenberger, P., Hauschild, A.: Broadcast versus precise ephemerides: a multi-GNSS perspective. GPS Solutions **19**(2), 321–333 (2015)
6. Misra, P., Enge, P.: Global Positioning System: Signals, Measurements and Performance Second Edition. Ganga-Jamuna Press, Lincoln (2006)
7. Duan, B., Wang, J., Wang, C.: Algebraic solution of GPS equations based on household transformation. J. Tongji Univ. (Nat. Sci.) **42**(7), 1123–1126 (2014)
8. Duan, B., Chen, J., Wang, J., Zhang, Y., Wang, J., Mao, L.: GNSS satellite clock real-time estimation and analysis for its positioning. In: Sun, J., Jiao, W., Wu, H., Lu, M. (eds.) CSNC 2014. LNEE, vol. 305, pp. 703–710. Springer, Heidelberg (2014). doi:10.1007/978-3-642-54740-9_62
9. Bierman, G.J.: A comparison of discrete linear filtering algorithms. IEEE Trans. Aerosp. Electron. Syst. (1), 28–37 (1973)
10. Bierman, G.J.: Factorization Methods for Discrete Sequential Estimation. pp. 33–36. Academic Press, New York (1976)

Research on Cloud Storage of Vector Data Based on HBase

Ruoxin Zhu[1(✉)], Jianqiao Cheng[2], Jianyong Fan[1], and Ke Chen[1]

[1] Zhengzhou Institute of Surveying and Mapping, Zhengzhou, China
zhuruoxin416@163.com, chxyfjy@163.com, ckl702@163.com
[2] Information and Electrics Department, Beijing Institute of Technology, Beijing, China
525301250@qq.com

Abstract. Nowadays, we enter the big data era. The amount of vector data is growing explosively. There is an urgent need for efficient storage method of vector big data. A cloud storage strategy of vector data based on HBase is proposed in this paper. Firstly, quadtree decomposition method is applied to build multi-level grid index and Hilbert space filling curve is applied to partition vector data. Secondly, vector element unique identifier is designed based on multi-level grid code and Hilbert sequence code. It is treated as RowKey of vector element in HBase. Thirdly, the storage rule of vector data is designed in detail. Finally, two contrast experiments are used to verify good feasibility and high efficiency of this proposed method.

Keywords: Component · HBase · Cloud storage · Vector data · Hilbert space filling curve · RowKey

1 Introduction

With the extending advances in technology of space data acquisition and geographic information applications, the data amount of vector data is booming. Faced with massive and complex vector data, how to achieve a lower cost and efficient storage is an urgent problem that needs to be settled [1]. This problem is something that cannot be solved easily under the concurrent environment of big data with a traditional relational database. Hadoop as an open source cloud platform with high reliability, fault tolerance, scalability features, unlimited expansion of its storage capacity and computing power offers a possibility in storing massive vector data effectively [2]. Document [3] vector data storage model adapted to the platform Hadoop Distributed File System (HDFS) is designed based on OGC simple coding model and it is combined with the MapReduce to explore the general procedures of parallel processing of vector data; [4] presents a Z filling curve and MapReduce parallel construct R-tree index methods to improve the speed of index building; [5] combines the small files into large files to improve the efficiency of reading and writing spatial data in lot of small files. However, the design based on HDFS doesn't give a good performance in supporting real-time modification of data. The investigation of building parallel structure index in R-tree and its derivative trees does not consider the impact on index reconfiguration brought by

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 473–482, 2017.
DOI: 10.1007/978-981-10-3969-0_52

the index data update [6]. Based on Hadoop HBase database system with reference to OGC simple features specification, vector data storage model is established, however, it does not consider that HBase's cluster row key uniquely identifies the design elements of vector and other issues.

Taking into account that the non-relational database HBase is suitable for large-scale data storage and support real-time modification in single or batch data without the limitation of location. This article proposes a strategy in the response of massive vector data, based on spatial indexing, data partitioning, clustering primary key design and rules of vector data storing.

2 HBase Database Storage Mechanism

HBase data sheet is organized according to Row and Column. The unique identifier of Row is RowKey which is similar to the primary key of a relational database table; the concept of Column Family is used in columns which identifies a set of columns. HBase data table use [RowKey, Column Family, Column Qualifier] to locate a data unit, the first coordinate is RowKey, the second is Column Family, the third is limited character of columns, referred to as Column. HBase data is stored in the unit as a value, each unit has to identify different versions of the same data by Time Stamp.

Table 1 is a logical data model of a HBase data table, each data record has a timestamp identity. HBase is a column-store sparse row/column matrix. The physical storage mode of row r1 is shown in Table 2 from which we can see, null values will not be stored in HBase table and in this model of storing data, new data can be added at any time to any column by using column name and timestamp distinctions. HBase does not

Table 1. HBase logical data model

RowKey	Time Stamp	Column Family: f1		Column Family: f2	
		Column	Value	Column	Value
r1	t6	f1:2	value 6		
	t5	f1:1	value 5		
	t4			f2:2	value 4
	t3			f2:1	value 3
r2	t2	f1:1	value 2		
	t1			f2:1	value 1

Table 2. HBase physical model.

RowKey	Time Stamp	Column Family	
		Column	Value
r1	t6	f1:2	value 6
	t5	f1:1	value 5
	t4	f2:2	value 4
	t3	f2:1	value 3

have rich data types. It only supports strings. Other types of data are supposed to be handle by users [7].

3 Vector Data Indexing and Data Partitioning

3.1 Based on Quadtree Split Multi-level Grid Index

In order to achieve effective storage and retrieval of massive multi-scale vector data, this article uses multi-level approach to building grid spatial index, partition way is shown in Fig. 1. According to quadtree split-level rule, dividing the entire area into a multi-level space grid area, and establish different correspondences according to different plotting scale of vector data, enabling the construction of the index of vector dataset [8].

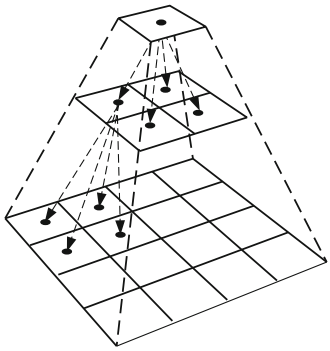


Fig. 1. Quadtree hierarchical way split

For fast access to vector data, we need to put adjacent vector data together as much as possible in the same scale in order to reduce the number of mobile disk operations and improve the efficiency of data reading and retrieving. HBase provides RowKey–the primary mean to achieve this goal. Through the design of RowKey, adjacent data stored together in index structure can be achieved. In the index table, if the identification of the grid cells is used as the memory effect that vector data cannot cluster when RowKey is storing data to multi-level grid index. In order to achieve the clustered storage of adjacent vector data in same level, according to the characteristics of RowKey arrangement, this article designs a RowKey form: "hierarchy _ grid unit identification". For example, the third level of grid level in Fig. 2, the pointer only need to move two times when searching in the specific area (sequentially moves to "03_0100" and "03_0130"), to complete reading and retrieving of the data.

This indexing mechanism has the following characteristics: First, strong dynamics, simple structure. Index doesn't have to be reconstructed when an update operation is done to a space object; secondly, convenience in inquiring. When processing data retrieval, related spatial grids can be quickly located according to the range of the

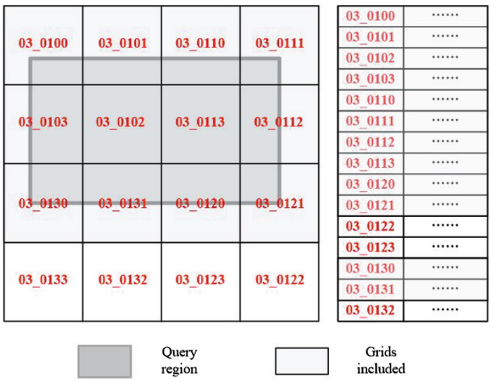


Fig. 2. A schematic arrangement of spatial data query area

search area and accuracy of the data requirements, without the need to pre-read the spatial index; In addition, the way of indexing also has a characteristic of clustering vector data. Vector data in the same hierarchy are stored into adjacent regions.

3.2 Vector Data Partitioning Based on Hilbert Filling Curve

The method of partitioning vector data is an important impact on vector data's storage efficiency, which is an important problem need to be solved when dealing with vector data distributed storage. Considering that Hilbert filling curve can retain the object's local adjacency. It is better to use Hilbert curve to partition vector data. The essence of Hilbert curve can be understood as the gradual decomposition of space, the arrangement order and the stepwise refinement process are shown in Fig. 3. The order of the grids which are passed by the curve and contain the destination point is what called the Hilbert permutation code of the destination point [9]. Hilbert permutation code generation algorithm is in reference [10].

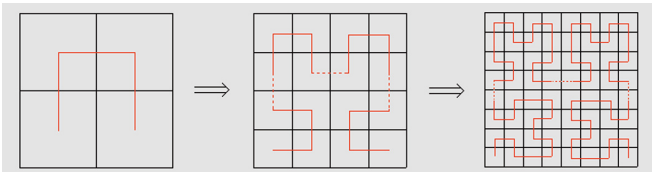


Fig. 3. Hilbert curve

In a stored procedure, after determining the corresponding the hierarchy of the multi-level spatial information and its trellis coding on the basis of the scale of vector data and spatial coordinates., we need to calculate the partition order of the Hilbert curve according to the data amount of vector data and locate it on the sub grids

partitioned by filling Hilbert curve according to the geometric position. Specific procedures are as follows:

Step 1: Determine the hierarchy of the corresponding multi-level grid according to the scale of the vector elements, and calculate the corresponding location of the grid based on the particle of vector elements.
Step 2: Determine Hilbert filling order. Calculate data amount of the vector data contained by each grid unit. Use formula (1) to calculate the partition order n of Hilbert, if n < 2, then the value is 2.

$$n = \lceil (m/G)/4 \rceil \tag{1}$$

"⌈┤⌉" represents "round up"; G represents the maximum amount of vector elements contained by a grid unit. It is set to 16 in this article.
Step 3: The grid is divided into H × H (where $H = 2^n$) sub-grid. With the number of its rows and columns we can calculate the corresponding Hilbert permutation code which is used as sub-grid Hilbert coding.
Step 4: determine the Hilbert filling sub-grid where the vector element is according to the location of the centroid. Calculate the number of vector elements contained by each sub-grid (Gk). If Gk > G, then n = n + 1 and turn to step 3. However, if Gk ≤ G, then partition is over. If n equals to the maximum partition order then stop step 4 and partition.

3.3 HBase KeyValue Design

Combined with multi-level grid index and Hilbert permutation code, this article designs identifications of vector elements which are also used as RowKey of vector elements in HBase database in order to achieve the clustered storage of vector data. Design ideas are as follows: identifications of vector elements consist of the hierarchy of the grid, coding of the grid unit, the order of Hilbert curve, the permutation code of Hilbert curve and sequence code. Macro position vector elements can be attained from the hierarchy and coding of the grid unit. All the vector elements contained in one sub-grid can be distinguished by sequence code. For vector elements such as lines, surfaces, etc., geometric center of mass is used to calculate position.

Constitution of the entire coding is shown in Fig. 4, FID = "grid level" + "break character" + "multilevel grid code" + "break character" + "sequence code" + "break character" + "Hilbert arrangement code" + "break character" + "order code". This

Grid level	Break character	Multilevel grid code	Break character	Order code	Break character	Hilbert arrangement code	Break character	Sequence code
2	1	20 in maximum	1	1	1	32 in maximum	1	3
decimal	_	"0123" code	_	hexadecimal	_	Binary	_	decimal

Fig. 4. Encoding of vector elements identifications

way of coding adapt to the rules of RowKey's storage. Thus this kind of encoding can be used as RowKey directly to accomplish data storing. This way of encoding has strong characteristics of clustering and spatial multi-scale.

3.4 Vector Data Storage Rules

This section follows rules of HBase storage, designing corresponding table structure of database and storage rules for vector elements and attribute data's dictionary respectively.

Vector Data Storage. According to the characteristics of vector data, vector data's table structure based on HBase is shown in the Table 3. Three columns in the table stores geometric data, attribute data and metadata ID sequentially. Data is stored as strings and translated into corresponding data type.

Table 3. Vector data's table structure.

RowKey	Time Stamp	Column Family: coordinate		Column Family: attribute		Column Family: metaId	
		Info	Value	Info	Value	Info	Value
Fea_ID2	t8			attribute:2	value 2		
	t7	MBR	box 2				
	t6	geocoor	WKB				
	t5					Meta_ID	metaID2
Fea_ID1	t4	MBR	box 1				
	t3	geocoor	WKB				
	t2			attribute:1	value 1		
	t1					Meta_ID	metaID1

Different data tables can be established for different data layers in order to store corresponding vector data. In the table, Fea ID, the identification of vector elements which is designed in 2.3 is used as RowKey of every record. Geometric data follows WKB standard of OGC. WKB (Well - Known Binary) uses a sequence of bytes to describe geometric objects, giving higher efficiency in reading writing and storing; Attribute data can have multiple columns. Every column represents an attribute of the vector element. It is easy to read data's attributes and translate them into specific types according to the data type and the layer's attributive character in the dictionary of attribute data.

Storage of Attributive Data's Dictionary. Attribute information of different data layers are different, so it is necessary to design appropriate dictionary of attribute data information for each data layer, facilitating reading and converting attribute data information in data table. Organization of attribute data dictionary table is shown in

Table 4. Attribute data dictionary table structure

RowKey	Time Stamp	Column Family: attribute	
		Info	Value
Control point layer			
⋮			
Transport layer	tn	"width"	"double"
		⋮	
	t4	"number of lanes"	"int"
	t3	"paving material"	"string"
	t2	"length"	"double"
	t1	"road type"	"String"

Table 4. RowKey value is the name of the data layer. Attrite can have multiple columns. The name of the attributive character is the name of the column. And the value of the column is the corresponding characteristic value.

4 Experimental Results and Analysis

4.1 Experimental Environment

There are thirteen virtual machines included in this experiment which constitute a Hadoop cluster. One virtual machine among them is used as primary node. Another eight machines of them are secondary nodes. Three machines of them are installed Zookeeper to provide coordination service. The last one of them is the device for clients. All hardware and software configurations are same for every virtual machine. Hardware configuration: 8 GB internal storage, 80 GB hard disk, 2.4 GHz CPU. Software configuration: Ubuntu 14.04 operating system, version 1.1.2 Hadoop cloud platform, distributed database HBase version 0.94.9, coordination service Zookeeper version 3.4.5.

All the vector data in this article is in Zhejiang 1:50,000 Shp format. There are three layers which give total 3785 M data. Just as shown in Table 5.

Table 5. Vector data in the experiment.

Layer	Entity amount/individual	Data amount/Mb
Contour line	434 536	2 785
Transportation	122 698	393
River system	236 673	607

4.2 Uniquely Identified Clustering Effect for Vector Elements

To test uniquely identified clustering effect for vector elements mentioned in Sect. 4.2, the experiment uses general store model of vector elements' identifications as the

reference. Build grid index respectively and retrieval space vector elements in specific regions from HBase data table. And we can verify the result by comparing the retrieval time of these two ways of storing. Retrieval time is the combination of "data read", "data transmission" and "write to the client".

Scheme 1 uses "sheet_ sequence code" as random identifications of vector elements which is also the RowKey value that will be stored in the experiment. Scheme 2 uses the unique identification for vector elements designed in Sect. 4.2 as RowKey of vector elements to store data. Retrieve vector data in range 1 (119.9°, 30.5°; 120.1°, 30.6°) and ranges 2 (118.9°, 29.5°; 120.8°, 30.6°) at the same time. Retrieval time is shown in Tables 6 and 7.

Table 6. Retrieval time of vector data for range 1 in using two different schemes.

Layer	Total number of records/piece	Number of queries/piece	Query amount/Mb	Scheme 1/s	Scheme 2/s
Contour line	434 536	850	7.2	5.447	2.893
Transportation	122 698	436	0.77	4.356	1.505
River system	236 673	2022	4.8	4.979	2.727

Table 7. Retrieval time of vector data for range 2 in using two different schemes.

Layer	Total number of records/piece	Number of queries/piece	Query amount/Mb	Scheme 1/s	Scheme 2/s
Contour line	434 536	91617	672.5	237.864	132.187
Transportation	122 698	19197	38.7	27. 476	10.658
River system	236 673	25568	63.5	63. 806	23.093

As shown in Tables 6 and 7, retrieval time in scheme 2 is much shorter than scheme 1's, which means that the method of unique identifications for vector elements has clustered effect. Because vector elements use unique identifications of vector elements as RowKey to store in HBase database. Thus the disk's pointer will move fewer times during the process of data reading because of clustered data. So as a result, retrieval time will be contracted.

4.3 Efficiency of Range Query

In order to verify the query efficiency of the vector data storage based on Hbase, the experiment use sArcGIS as reference when doing contrast experiments of range query. Scheme 1: configure an virtual machine with CPU 2.67 GHZ, internal storage 8 GB. Install Oracle11 g and use ArcCatalog to import experimental data into the database of Oracle. Use API offered by ArcSDE to program the test code for retrieving vector data. Scheme 2: implement query test based on the way of vector data storing designed in

this article. Inquire the vector data from range 1 (119.9°, 30.5°; 120.1°, 30.6°) and range 2 (118.9°, 29.5°; 120.8°, 30.6°) the query time is the combination of "reading data", "data transmission" and "writing to clients". Query time of scheme 1 is shown in Table 8. Query time of scheme 2 has already been gotten in the experiment in Sect. 3.2, as can be seen from Tables 6 and 7. The comparisons of query time gotten from scheme 1 and scheme 2 of range 1 and range 2 are shown in Figs. 5 and 6 respectively.

Table 8. Retrieval time of vector data for scheme 1.

Layer	Total number of records/piece	Range 1/s	Range 2/s
Contour line	434536	1.356	177.302
Transportation	122 698	0.913	12.776
River system	236 673	2.427	60.897

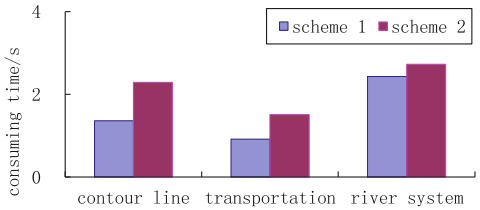


Fig. 5. Comparison of query time of vector data in range 1.

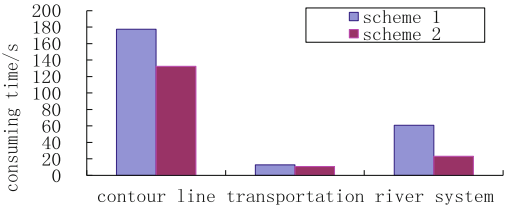


Fig. 6. Comparison of query time of vector data in range 2

From Figs. 5 and 6, we can know than scheme 1 gives higher efficiency when inquiring data in a small scale. However, when doing data query in a large scale, scheme 2 shows higher efficiency. In conclusion, the cloud storage scheme designed in this article has worse performance than ArcSDE when doing data query in a small scale but gives better performance when inquiring data in a large scale. Because cloud storage of vector data mainly focuses on storing vector data in large amount, thus the scheme of storing vector data designed in this article based on HBase has great practical value.

5 Summary

Sharp increase in the amount of vector data makes it difficult to store. Based on distributed cloud storage platform, this article designs a scheme of vector data cloud storage from several aspects: index, vector data partitioning, KeyValue, storage rules. Two experiments of clustered effect and query efficiency can verify the effectiveness of this method, which offers a solution for the storage of massive vector data. In subsequent studies, the emphasis will be investigating deeply into vector data's query algorithm of parallel space according to application requirements with the combination of MapReduce.

Acknowledgments. This work was supported by Natural Science Foundation of China (Project No. 41401462) and Scientific and technological Project of Zhengzhou (No. 112PPTGY225). The Authors would like to thank the anonymous reviewers for their valuable comments, which greatly helped us to clarify and improve the contents of paper.

References

1. Progress SURVEYING AND MAPPING navigation and geographic information science and technology - to celebrate the "Science of Surveying and Mapping Technology" founded 30 years. Surv. Map. Sci. Technol. **2014**(5): 441–449 (2014)
2. Wang, Y.-J., Sun, W., Zhou, S.: Key technologies of distributed storage for cloud computing. Software **23**(4), 962–986 (2012)
3. Wang, Y.: Several key technologies of geographic information service based on Hadoop cloud computing platform. Ph.D. thesis, Graduate School of Chinese Academy of Sciences (2011)
4. Cary, A., Sun, Z., Hristidis, V., Rishe, N.: Experiences on processing spatial data with MapReduce. In: Winslett, M. (ed.) SSDBM 2009. LNCS, vol. 5566, pp. 302–319. Springer, Heidelberg (2009). doi:10.1007/978-3-642-02279-1_24
5. Jerome, D., Cyrille, B., Flanvien, M.: Simple method for the estimation of the short-term of GNSS on-board clocks. In: Proceedings of 42nd Annual Precise Time and Time Interval (PT-TI) Meeting, pp. 215–223. The Institute of Navigation, Virginia (2010)
6. Lars, G.: HBase: The Definitive Guide, pp. 319–323. O'Reilly Media, Newton (2011)
7. Zheng, K., Fu, Y.: Vector's spatial data storage model based on HBase and GeoTools. Comput. Appl. Softw. **2015**(3), 23–26 (2015)
8. Han, H., Cheng, C.Q., Wang, Y., et al.: Rapid collection method of multi-source data based on global subdivision grid. Geomat. World **2014**(6), 6–11 (2014)
9. Lu, F., Zhou, C.: A GIS spatial indexing approach based on Hilbert ordering code. Comput.-Aided Des. Comput. Graph. **13**(5), 424–429 (2001)
10. Wang, Y., Meng, K.: Spatial partitioning of massive data based on Hilbert spatial ordering code. Geomat. Inf. Sci. Wunan Univ. **32**(7), 650–653 (2007)

Research on Visualization Methods for Academic Papers Analysis of Chinese Surveying and Mapping Journals

Jing Li, Haiyan Liu(✉), Wenyue Guo, and Ruijie Yang

Institute of Geospatial Information,
Zhengzhou Institute of Surveying and Mapping, Zhengzhou, China
{351800040, 563745754, jieruiyang1992}@qq.com,
liu2000@vip.sina.com

Abstract. Taking six influential Chinese surveying and mapping journals as the data source, and using the method of bibliometrics, research themes and authors' geographical distribution were analyzed in the field of surveying and mapping. In this paper, we propose some reasonable and effective visualization methods in accordance with the characteristics such as data types and the contents in bibliometric research through the way of analyzing Chinese surveying and mapping journals.

Keywords: Visualization · Bibliometrics · Tag cloud · Textual data · Word frequency

1 Introduction

Academic paper is an important way to transmit the concept of scientists and describes the results of academic research. It is not only a means to solve the problem, but also a tool to describe the academic exchange. In 1922, E.W. Hulme firstly proposed the concept of bibliometrics to analysis the academic from the qualitative to quantitative.

Bibliometrics is a quantitative analysis method, with various external feature as the research object, using mathematical and statistical methods to describe, and it is quantity analysis method using in analyzing the external features of the literature. It focuses on the analysis of the literature form features of "quantity", analyzing of the law of literature from a quantitative point, and indirect reflecting the relationship. Recently, with the growing number of academic papers, effective and reasonable analysis are needed to realize the development of a subject from the qualitative to quantitative.

The original bibliometrics is to analyze the data through the statistical methods without combining any graphics. With the increasing of analysis methods, the expression of complex data is limited, some significant information will be hided under the large numbers of statistical data without visualization. In 1970s, Dinsmore drew citation data in the social sciences of various fields to subject structure diagram. It was the first time that bring the concept of visualization into bibliometrics. At present,

H. Yuan et al. (Eds.): GRMSE 2016, Part II, CCIS 699, pp. 483–491, 2017.
DOI: 10.1007/978-981-10-3969-0_53

visualization technology is used more and more widely in the field of bibliometric research.

According to the characteristics such as data types and the contents in bibliometric research, we propose some reasonable and effective visualization methods in this paper.

2 Characteristics in Bibliometric Research

The academic paper is the main object in bibliometric research. The academic paper is different from the general text data, it has the certain structure, which belongs to the semi structure data. The academic paper consists of title, author, abstract, keywords, and main body, etc. According to the constitution of academic paper, three contents in bibliometric research are identified: keywords analysis, author analysis and the amount of published paper analysis.

On the basis of the three contents above, we choose three data types to visualize: textual data, attribute data of time, and attribute data of location. The specific contents of this paper are as follows:

- Visualization of word based on word frequency
- Visualization of textual data based on the time attribute
- Visualization of location attribute

3 Visualization Method

3.1 Visualization Method of Word Based on Word Frequency—Tag Cloud

Word frequency analysis is an important research method in the research of bibliometrics. Through the statistics of the word frequency to reveal the rules. For example, the high-frequency keywords could reflect the hot topic in the research field in some degrees. Involves two ingredients of vocabulary level text and word frequency, Tag cloud is the most simple and commonly used visualization method.

Tag cloud arrange the words on the screen orderly with some orders, rules or constraints. Highlighting the importance of word in the way of setting different font sizes of words.

In the paper, we count the keywords in Chinese surveying and mapping journals from 2006 to 2015. We display the top 20 keywords in the order of word frequency from high to low in Table 1.

Figure 1 will be the content of Table 1, with the frequency of the tag cloud as the constraint conditions. It can be seen that the larger font size and the central location of the keywords show the higher importance.

Table 1. Top 20 keywords in Chinese surveying and mapping journals from 2006 to 2015 ordered by word-frequency.

Ranking	Name	Word-frequency
1	Cartographic generalization	116
2	Visualization	111
3	Spatial analysis	96
4	Topological relationship	80
5	Spatial data	77
6	Geographic information	74
7	Electronic map	74
8	Spatial database	71
9	DEM	61
10	3D visualization	58
11	Database	53
12	Geographic information service	51
13	Spatio-temporal data model	48
14	Remote sensing	45
15	Digital city	44
16	3D GIS	42
17	Map	40
18	spatial relationship	38
19	Mobile GIS	37
20	Ontology	36

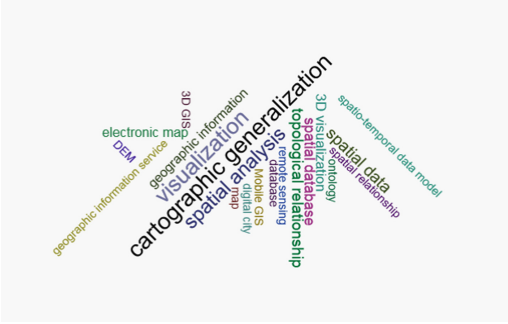


Fig. 1. Tag cloud of top 20 keywords in Chinese, surveying and mapping journals from 2006 to 2015.

3.2 Visualization Method of Textual Data Based on the Time Attribute

In the research of bibliometrics, it is an important content for researchers to observe the changes of hotspots in terms of time. Publication date of academic papers provides the opportunity for researchers to find out the rules of discipline development over time.

There are two factors should be considered: time, and high frequency keywords (which are usually used to present the hotspots).

In the time series analysis, a time bar is a straightforward way to look at data over time. In this paper, we display the top 10 keywords ordered by word frequency from high to low each year from 2006 to 2015 in Fig. 2.

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
geographic information service	spatio-temporal data model	data model	spatial analysis	spatial analysis	visualization	visualization	spatial data	visualization	visualization
spatial data	electronic map	database	database	visualization	topological relationship	cartographic generalization	cartographic generalization	geographic information	geographic information
spatial database	spatial data	spatial database	visualization	3DGIS	spatial database	spatial database	topological relationship	geographic information service	spatial analysis
visualization	spatial analysis	spatial data mining	spatial database	electronic map	spatial analysis	electronic map	visualization	cartographic generalization	database
geographic information	topological relationship	3Dvisualization	spatial data	DEM	cartographic generalization	spatial analysis	point-cloud	topological relationship	cartographic generalization
object-oriented	DEM	spatial data	DEM	remote sensing	electronic map	topological relationship	digital city	spatial analysis	big data
map	spatial database	geographic information	electronic map	spatial data	3DGIS	spatial data	geographic information	big data	cloud computing
spatial analysis	database	spatial analysis	map making	geographic information service	3D modelling	map	electronic map	digital city	smart city
spatial relationship	spatial relationship	DEM	map	ontology	spatial data mining	geographic information	mobileGIS	remote sensing	Tile Map
electronic map	map making	virtual reality	remote sensing image	spatio-temporal data model	3DGIS	geographic information service	spatial analysis	spatial data	geographic information service

Fig. 2. The top 10 keywords from 2006 to 2015.

It's difficult to find out information from Fig. 2. Then considering about the continuity of keywords on the time dimension, we draw the continuous appeared keywords with shining colors (the same key words with the same color) in Fig. 3.

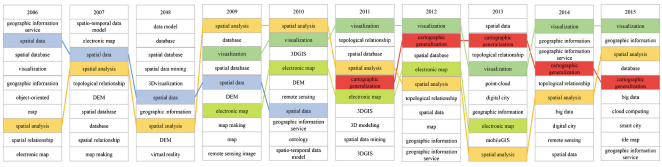


Fig. 3. The continuous appeared keywords from 2006 to 2015. (Color figure online)

From Fig. 3, we can easily find out that five keywords, "spatial data", "spatial analysis", "visualization", "electronic map", "cartographic generalization", are continuous appeared in Chinese surveying and mapping journals during this period.

But it is still too complex to find out useful information in Fig. 3. So we simplify Fig. 3 to Fig. 4 by deleting some useless keywords to highlight what we exactly focus on.

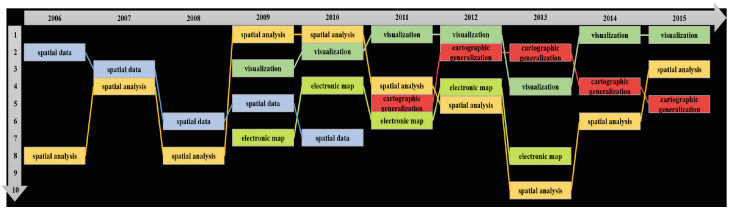
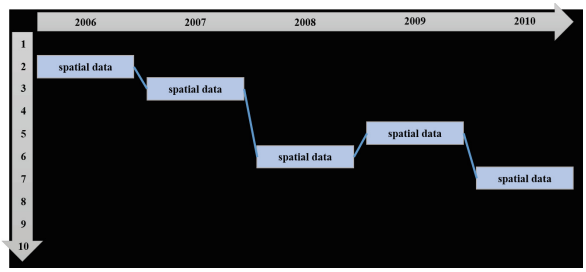
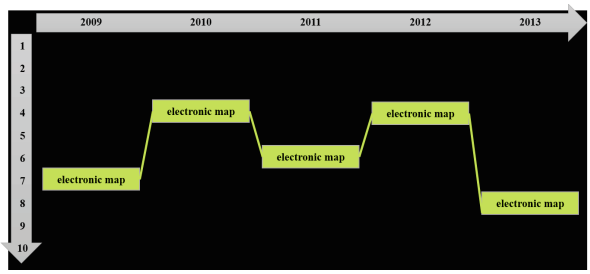


Fig. 4. The simplified graph of continuous appeared keywords from 2006 to 2015.

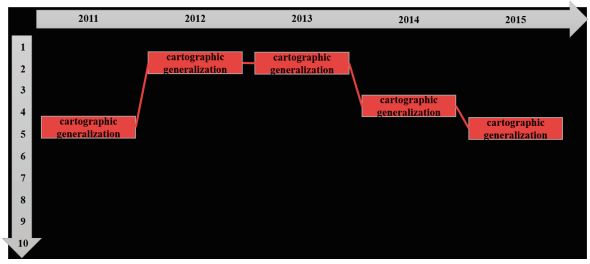
Figure 5 is the decomposition of Fig. 4. From Fig. 5(1), we can find out that spatial data once was a hotspot in the field of surveying and mapping in the early years during 2006 to 2015. From Fig. 5(2), we find that in the middle stage, the study of electric map drew more attentions of surveying and mapping researchers. And in the later years, research of cartographic generalization has been increased.



(1)



(2)



(3)

Fig. 5. Three period of Chinese surveying and mapping research (the decomposition of Fig. 4).

3.3 Visualization Method of Location Attribute

In bibliometrics, distribution of academic paper publishing is another important content to consider of. Through the analysis of geographic distribution, researchers could identify the core research area.

There are two factors to visualizing: location information and the number of paper publishing. In our study, we choose cartogram to visualizing those two factors. We use circles which size present the data to instead of physical area using shapes.

Figures 6 and 7 are drawn the geographical distribution with the data of Chinese surveying and mapping journals from 2003 to 2013 by using Tableau visualization software. Based on 5 article number for the node, we consider the place paper number above 5 as the high post area, less than 5 and greater than zero as the low post area.

From Fig. 6, the high post area mainly distribute in central and eastern regions, and only Urumchi is in the west of China. High post areas are mainly the provincial capital, municipality directly under the central government and other large cities.

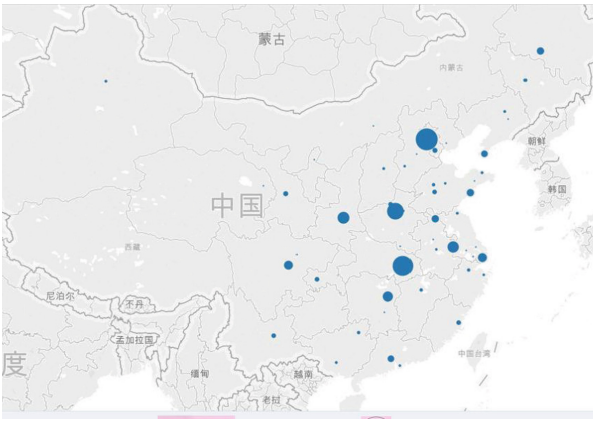


Fig. 6. Geographical distribution (the number of paper publishing > 5).

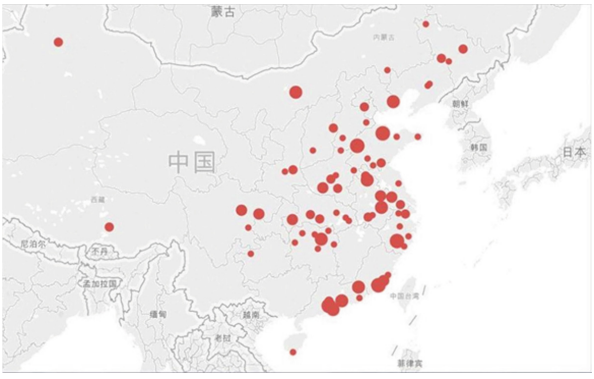


Fig. 7. Geographical distribution (the number of paper publishing $\leq$ 5).

From Fig. 7, the low post areas are also in central and eastern regions, and southeastern coastal cities. The scale is relatively smaller. Overall, regardless of the high post or the low post areas, they mainly distribute in central and eastern regions. The scale of the city and the economic development directly affect the level of the scientific research.

4 Conclusion

Visualization can be a great tool to explore data, and the key to getting the most out of the data—to understand what it represent and what it means. In this paper, according to the data types and research contents, we propose some visualization methods in bibliometrics. Through the way of visualizing the paper data of surveying and mapping, we can draw the following conclusion:

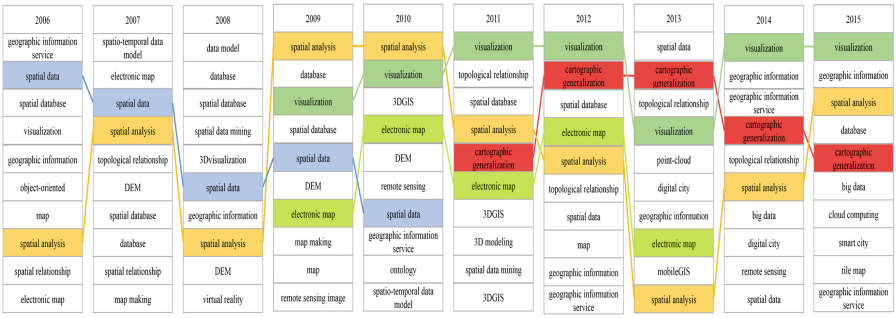
A. In bibliometrics, words analysis is a significant part. Tag cloud is a simple and useful method of visualizing words.
B. In the time series analysis, a time bar is a straightforward way to look at data over time. And when the picture is too complex, simplifying it will help.
C. Cartogram is a method of visualizing the combination of statistics and spatial data. It ignores the physical area and makes entire regions sized by data.

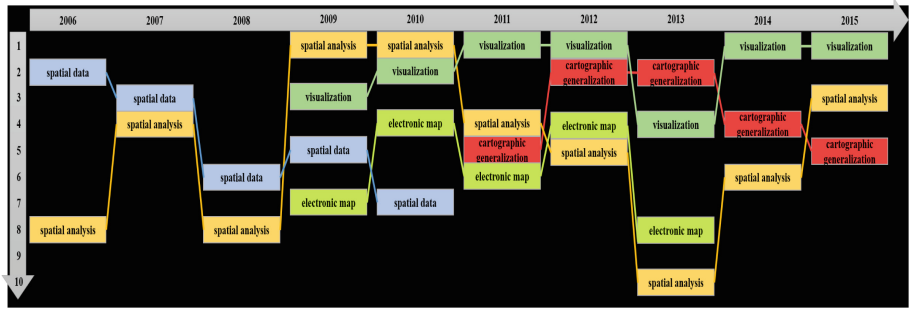
Acknowledgment. This work is funded by State Key Laboratory of Geo-information Engineering (NO. SKLGIE2015-M-4-3) and the National Natural Science Foundation of China (Nos. 41471387, 41501446).

APPENDIX (OPTIONAL)

Larger Visions:

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
geographic information service	spatio-temporal data model	data model	spatial analysis	spatial analysis	visualization	visualization	spatial data	visualization	visualization
spatial data	electronic map	database	database	visualization	topological relationship	cartographic generalization	cartographic generalization	geographic information	geographic information
spatial database	spatial data	spatial database	visualization	3DGIS	spatial database	spatial database	topological relationship	geographic information service	spatial analysis
visualization	spatial analysis	spatial data mining	spatial database	electronic map	spatial analysis	electronic map	visualization	cartographic generalization	database
geographic information	topological relationship	3Dvisualization	spatial data	DEM	cartographic generalization	spatial analysis	point-cloud	topological relationship	cartographic generalization
object-oriented	DEM	spatial data	DEM	remote sensing	electronic map	topological relationship	digital city	spatial analysis	big data
map	spatial database	geographic information	electronic map	spatial data	3DGIS	spatial data	geographic information	big data	cloud computing
spatial analysis	database	spatial analysis	map making	geographic information service	3D modeling	map	electronic map	digital city	smart city
spatial relationship	spatial relationship	DEM	map	ontology	spatial data mining	geographic information	mobileGIS	remote sensing	Tile Map
electronic map	map making	virtual reality	remote sensing image	spatio-temporal data model	3DGIS	geographic information service	spatial analysis	spatial data	geographic information service





References

1. Tang, G., Liu, Z., Sun, M.: Text visualization research review. J. Comput.-Aided Des. Graph. **25**(3), 273–285 (2013)
2. Liu, H., Sun, Q., Xiao, Q., et al.: In the digital mapping of multi-source data (information) comprehensive application. J. Surveying Mapp. Sci. Technol. **23**(3), 161–164 (2006)
3. WINKLER: Evaluation of Scientific Research Scientific Metrology Index. Science and Technology Literature Press, Beijing, December 2014. Translated by M.A. Zheng
4. Jiang, Y.: The Research on the Literature Metrology in the Field of Humanities and Social Sciences. Social Science Literature Press, Beijing
5. Lamport, L.: LaTeX, A Document Preparation System, 2nd edn. Addison-Wesley, Reading (1994)
6. Chen, S., Zeng, X., Liang, J.: Based on statistical data of GIS visualization study. Comput. Eng. Des. **29**(14), 3757–3759 (2008)
7. Wang, Y.: The comprehensive research of literature metrology and content analysis. Nanjing University of science and technology (2007)
8. Lin, H., Gao, T.: The visual representation of Chinese text. J. Northeast. Univ.: Nat. Sci. Ed. **21**(5), 501–504 (2000)
9. Jiang, C., Liu, S., Ding, K.: Research hotspot and the evolution of knowledge map in China of the journal of science and technology. J. Environ. Sci. **06**, 954–958 (2008)

Author Index