

Sheng-Lung Peng
Margarita N. Favorskaya
Han-Chieh Chao *Editors*



Sensor Networks and Signal Processing

Proceedings of the 2nd Sensor Networks
and Signal Processing (SNSP 2019),
19–22 November 2019, Hualien, Taiwan

Smart Innovation, Systems and Technologies

Volume 176

Series Editors

Robert J. Howlett, Bournemouth University and KES International,
Shoreham-by-sea, UK

Lakhmi C. Jain, Faculty of Engineering and Information Technology,
Centre for Artificial Intelligence, University of Technology Sydney,
Sydney, NSW, Australia

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

**** Indexing: The books of this series are submitted to ISI Proceedings, EI-Compendex, SCOPUS, Google Scholar and Springerlink ****

More information about this series at <http://www.springer.com/series/8767>

Sheng-Lung Peng · Margarita N. Favorskaya ·
Han-Chieh Chao
Editors

Sensor Networks and Signal Processing

Proceedings of the 2nd Sensor Networks
and Signal Processing (SNSP 2019), 19–22
November 2019, Hualien, Taiwan

 Springer

Editors

Sheng-Lung Peng
Computer Science and Information
Engineering
National Dong Hwa University
Hualien, Taiwan

Margarita N. Favorskaya
Informatics and Computer Techniques
Reshetnev Siberian State University of
Science and Technology
Krasnoyarsk, Russia

Han-Chieh Chao
Department of Electrical Engineering
National Dong Hwa University
Hualien, Taiwan

ISSN 2190-3018

ISSN 2190-3026 (electronic)

Smart Innovation, Systems and Technologies

ISBN 978-981-15-4916-8

ISBN 978-981-15-4917-5 (eBook)

<https://doi.org/10.1007/978-981-15-4917-5>

© Springer Nature Singapore Pte Ltd. 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

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

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Organization

The 2nd Sensor Networks and Signal Processing (SNSP 2019)

General Chairs

Han-Chieh Chao, National Dong Hwa University, Taiwan
Huang Yen-Nun, Academia Sinica, Taiwan

Organizing Chair

Sheng-Lung Peng, National Dong Hwa University, Taiwan

Organizing Committee

Guan-Ling Lee, National Dong Hwa University, Taiwan
Shou-Chih Lo, National Dong Hwa University, Taiwan
Min-Xiou Chen, National Dong Hwa University, Taiwan
Yao-Chung Chang, National Taitung University, Taiwan

Publication Chair

Sheng-Lung Peng, National Dong Hwa University, Taiwan

Program Chairs

Sheng-Lung Peng, National Dong Hwa University, Taiwan
Suresh Chandra Satapathy, KIIT, India

Program Committee

Omar Arif Abdul-Rahman, IBM, Japan
Yogesh Beeharry, University of Mauritius, Mauritius
Xi Chen, Simons Foundation, USA
Mu-Song Chen, Da-Yeh University, Taiwan
Rex Cheung, San Francisco State University, USA

Domenico Ciuonzo, Network Measurement and Monitoring, Italy
Jun Duan, Stony Brook University, USA
Joydev Ghosh, National Research Tomsk Polytechnic, Russia
Anandakumar Haldorai, Sri Eshwar College of Engineering, India
Hamza Issa, Beirut Arab University, Lebanon
Jaichandran, Vinayaka Missions University, India
Hwang Hun Jeong, Korea Construction Equipment Technology Institute, Korea
Harish Krishnamoorthi, Cochlear Americas, USA
Grigorios Kyriakopoulos, National Technical University of Athens, Greece
Chih-Chin Lai, National University of Kaohsiung, Taiwan
Jiade Li, University of Florida, USA
Juane Li, University of California, USA
Yi-Ching Liaw, Nanhua University, Taiwan
Jiang Lu, University of Houston, USA
Chun-Lin Lu, Kun Shan University, Taiwan
Sarhan Musa, Prairie View A&M University, USA
Filipe Neves, Politechnic Institute of Leiria, Portugal
Ladislav Polak, Brno University of Technology, Czech Republic
V. A. Sankar Ponnappalli, Sreyas Institute of Engineering and Technology, India
P. Rama Koteswara Rao, Sri Vasavi College of Engineering and Technology, India
Gayathri S., Sri Jayachamarajendra College of Engineering, India
Dangdang Shao, Qualcomm Technologies Inc., USA
Oleksii K. Tyshchenko, University of Ostrava, Czech Republic
Ting Wang, Huawei Technologies Co. Ltd, China
Baiqiang You, Xiamen University, China
Connie Yuan, Hong Kong Institute of Vocational Education, China
Sai Zhang, Arizona State University, USA

2019 International Conference on Machine Learning and Intelligent Systems (MLIS 2019)

Honorary Chairs

Han-Chieh Chao, National Dong Hwa University
Lakhmi C. Jain, University of Canberra

General Chairs

Cheng-Chin Chiang, National Dong Hwa University
Chun-Shien Lu, Academia Sinica

Organizing Chair

Sheng-Lung Peng, National Dong Hwa University, Taiwan

Organizing Committee

Shi-Jim Yen, National Dong Hwa University
 Mau-Tsuen Yang, National Dong Hwa University
 I-Cheng Chang, National Dong Hwa University

Publication Chair

Sheng-Lung Peng, National Dong Hwa University, Taiwan

Program Chairs

Sheng-Lung Peng, National Dong Hwa University, Taiwan
 Nilanjan Dey, Techno India College of Technology

Program Committee

Mohamed Awad Awad Allah, Al-Aqsa University
 R. J. Anandhi, New Horizon College of Engineering
 Dinesh Kumar Anguraj, Koneru Lakshmaiah (KL) University
 Luis Anido Rifon, University of Vigo
 Chaodit Aswakul, Chulalongkorn University
 Chawalit Benjangkprasert, King Mongkut's Institute of Technology
 Giuseppe Bonifazi, University of Rome
 Noureddine Bouhmala, University of South-Eastern Norway
 Aleksander Cariow, West Pomeranian University of Technology
 Elena Casiraghi, University of Milan
 Cristina Cervelló-Pastor, Polytechnic University of Catalonia
 Ricardo Chalmeta Rosaleñ, Jaume I University
 I-Cheng Chang, National Dong Hwa University
 Qiang (Shawn) Cheng, University of Kentucky
 Cheng Siong Chin, Newcastle University in Singapore
 Claudio Cuevas, Federal University of Pernambuco
 Antonió Curierno, Università degli Studi di Salerno
 Tolga Ensari, Istanbul University
 Solomiia Fedushko, Lviv Polytechnic National University
 Feng Feng, Xi'an University of Posts and Telecommunications
 R. U. Gobithaasan, University Malaysia Terengganu
 Du Huynh, University of Western Australia
 Dmitry Kravchenko, Ben-Gurion University of the Negev
 Ryong Lee, Korea Institute of Science and Technology Information (KISTI)
 Kian-Guan Lim, Singapore Management University
 Songsong Liu, Swansea University
 Edwin Lughofer, Johannes Kepler University Linz
 Syed Tauseef Mohyud-Din, HITEC University Taxila Cantt, Pakistan
 Josefa Mula, Polytechnic University of Valencia
 Fusaomi Nagata, Sanyo-onoda City University
 Idelfonso Nogueira, University of Porto

Leonardo Barbosa e Oliveira, Stanford University
Ong Pauline, Universiti Tun Hussein Onn Malaysia (UTHM)
R. M. Chandima Ratnayake, University of Stavanger
B. S. Daya Sagar, Indian Statistical Institute-Bangalore Centre
Mohammad Salah, Hashemite University
Ahm Shamsuzzoha, University of Vaasa
Kittiwat Sirikasemsuk, King Mongkut's Institute of Technology Ladkrabang
Abdulhamit Subasi, Effat University
Yuriy Syerov, Lviv Polytechnic National University
Ze Tang, Jiangnan University
Gabriella Trucco, University of Milan
Hung-Hsu Tsai, National Formosa University
Athanasios V. Vasilakos, Luleå University of Technology
Anoop P. Verma, NextEra Energy
Bing Wang, Anhui University of Technology
Chia-Hung Wang, Fujian University of Technology
Zezhong Xu, Changzhou Institute of Technology
Qingzheng Xu, National University of Defense Technology
Jun-Juh Yan, Shu-Te University
Jasy Liew Suet Yan, Universiti Sains Malaysia
Selma Yilmazyildiz, Vrije Universiteit Brussel
Seung H. Baek, Hanyang University
Wen-Tsai Sung, National Chin-Yi University of Technology
Md. Hasinur Rahaman Khan, University of Dhaka
Oleksii Tyshchenko, University of Ostrava
Shin-Jer Yang, Soochow University

Preface

The 2nd Sensor Networks and Signal Processing (SNSP 2019) Conference is held in conjunction with the 2019 International Conference on Machine Learning and Intelligent Systems (MLIS 2019). With the support of National Dong Hwa University (NDHU), they are held together in National Dong Hwa University, Hualien, Taiwan, during November 19–22, 2019. It aims to present the developments in leading research and facilitate cross-disciplinary interaction in four main themes: sensor networks, signal processing, machine learning and intelligent systems.

SNSP 2018 was sponsored by the School of Communication and Information Engineering, Xi'an University of Science and Technology. Following the great success of the last conference, we received nearly 150 submissions from 14 different countries and constructed a strong technical program in six sessions for the two conferences. After an intense and strict peer review by the program committee, which is composed of by many experts, we decided to accept 42 papers, which represents an acceptance rate of about 30%.

On behalf of sponsors and conference committees, we want to express our special thanks to all the keynote and invited speakers, authors and attendees who put their effort in preparing at least a contribution for the conference. On the other hand, we are very grateful to the people, especially the program committee members and reviewers, who devoted time to evaluate the papers. In addition, we are grateful to Springer for publishing the conference proceedings.

Finally, we would like to thank all who have helped in making the two conferences a success. We hope that you will enjoy the technical program and the social events of the conference, and you can also discover many beautiful places in Hualien, Taiwan. Wishing you a fruitful and enjoyable SNSP 2019 and MLIS 2019.

Hualien, Taiwan
Krasnoyarsk, Russia
Hualien, Taiwan

Sheng-Lung Peng
Margarita N. Favorskaya
Han-Chieh Chao

Contents

Part I Sensor Networks and Communications

1	An Improved Clustering Routing Algorithm for Heterogeneous Wireless Sensor Network	3
	Huan Yao	
2	Improvement of RPL Routing Strategy Based on 6LoWPAN	21
	Hui Wang, Zhihui Fan, Xiangfu He, Peiyu Li, and Chaowu Zhang	
3	Design of a New Multi-output Constant Current Source Based on Power Allocation Control Strategy	37
	Hongli Cheng and Lei Wang	
4	Use Accelerometer to Monitoring the Exercise Status: Held the Push-Ups Movement as an Example	51
	Yi-Yang Chen, Li-Wa Sha, and Wen-Hsin Chiu	
5	Low-Coupling 2 * 1 Micro-strip Antenna Array Based on Defect Structure	63
	Ning Guo, Xinliang Liu, and Nana Bu	
6	Power Control in D2D Underlay Distributed Antenna Systems	73
	Gongbin Qian, Ce Zhang, Chunlong He, Xingquan Li, and Chu Tian	
7	Analysis of Transmission Efficiency of Magnetically Coupled Resonant Radio Energy	87
	Xiaohu Yin and Yue Zhao	
8	Remote Monitoring of River Water Pollution Using Multiple Sensor System of WSNs and IoT	99
	Evizal Abdul Kadir, Hitoshi Irie, Sri Listia Rosa, Bahruddin Saad, Sharul Kamal Abdul Rahim, and Mahmud Othman	

Part II Signal and Image Processing

- 9 Dimension Detection of Varistor Based on Random Hough Transform** 117
Wei Chen and Xueying Yang
- 10 Phase Retrieval Method Based on Transport of Intensity Equation with Microscope Single Field of View** 127
Hong Cheng, Rui Wang, Fen Zhang, Wenxia Bao,
and Quanbing Zhang
- 11 Graphic QR Code with the Second Hidden QR Code by Codeword Rearrangement** 137
Yi-Wei Juan, Tzren-Ru Chou, Chun-Shien Lu, and Hsi-Chun Wang
- 12 A Novel Position-Shift Method of Double-Phase Fresnel Hologram for Synthesizing a Complex Fresnel Hologram** 149
Chuan Shen, QinQin Zhu, QingQing Hong, Cheng Zhang,
and Sui Wei
- 13 An Examination of Obstacle Avoidance by Sound for Visually Impaired Children** 161
Yukiko Matsushima, Teruo Kimura, Feifei Cho, and Noboru Yabuki
- 14 Efficient Technique of Impulsive Noise Detection and Replacement in Color Digital Images** 171
Bogdan Smolka
- 15 Development of Walking Support System for Visually Impaired People** 187
Feifei Cho, Tatsuya Ohta, Yukiko Matsushima, Teruo Kimura,
and Noboru Yabuki
- 16 Research on Video Mosaic Technology Based on Fully Mechanized Face of Coal Mine** 203
Wei hu Zhang, Zhihui Tao, and Xu Li

Part III Data Processing and Security

- 17 Optimization Scheme for Traceability of Distributed Denial of Service Attacks Based on Dynamic Probability Packet Marking** 217
Li Chen and Jun Yao
- 18 Processing Initial Data for the Agent-Based Model of the Russian Federation Spatial Development** 227
Aleksandra L. Mashkova

19 Design and Implementation of Total Station Wireless Data Transmission System 241
 Xiaohu Yin and Mihuan Wang

20 The Comprehensive Evaluation of “Five Aspects” Based on Coefficient-of-Variation-Modified G1 Combination Weighting 253
 Wei Ren and Hao Jian

21 An Inertia Weight Variable Particle Swarm Optimization Algorithm with Mutation 269
 Mengying Zhao, Yuqi Ni, Tao Chao, and Ke Fang

Part IV Machine Learning

22 Performance of Probabilistic Approach and Artificial Neural Network on Questionnaire Data Concerning Taiwanese Ecotourism 283
 Vladislav Bína, Václav Kratochvíl, Lucie Váchová, Radim Jiroušek, and Tzong-Ru Lee

23 AC Operation Hardware Learning Neural Circuit Using V-F Converter System 297
 Masashi Kawaguchi, Naohiro Ishii, and Masayoshi Umeno

24 An Inequality for Linear Canonical Transform 311
 Mawardi Bahri and Ryuichi Ashino

25 A General Approach to Probabilistic Data Mining 325
 Radim Jiroušek and Václav Kratochvíl

26 A Novel Four-Dimensional Chaotic System with Four Cross Terms 341
 Jinmei Liu

27 Influence of the Optical Aspects of Photographic Composition on the User Experience in the Issues Related to Decision Making, Choices and Level of Visual Comfort 355
 Marcia Campos and Fabio Campos

28 Compositional Models: Iterative Structure Learning from Data 379
 Václav Kratochvíl, Vladislav Bína, Radim Jiroušek, and Tzong-Ru Lee

29 Consciousness Detection in Complete Locked-In State Patients Using Electroencephalogram Coherency and Artificial Neural Networks 397
 V. S. Adama and Martin Bogdan

30 Computer Vision-Based Demersal Fish Length Measurement Technology 411
Sheng-Wen Jeng, Chih-Kai Chiu, and Kai-Siang Gan

31 A Haze Removal Method Based on Additional Depth Information and Image Fusion 423
Tian Tian and Bin Zhang

32 Vehicle Detection Based on Area and Proportion Prior with Faster-RCNN 435
Hao Yuan, Bin Zhang, and Ming Liu

33 Mining High-Utility Itemsets of Generalized Quantity with Pattern-Growth Structures 447
Ming-Yen Lin, Tzer-Fu Tu, and Sue-Chen Hsueh

34 Large-Scale Instance Selection Using a Heterogeneous Value Difference Matrix 465
Chatchai Kasemtaweekchok, Nitiporn Sukkerd, and Chatchavin Hathorn

Part V Intelligent Systems

35 Bio-inspired Algorithms for Modeling and Control of Underwater Flexible Single-Link Manipulator 483
I. Z. Mat Darus and Ali A. M. Al-Khafaji

36 Application of Time Series Method to the Passenger Flow Prediction in the Intelligent Bus Transportation System with Big Data 497
Yinna Ye, Ruoxi Liu, and Feng Xue

37 Application of Sample Entropy to Analyze Consciousness in CLIS Patients 521
Shang-Ju Wu and Martin Bogdan

38 Intelligent Tuning of PID Controller for Double-Link Flexible Robotic Arm Manipulator by Artificial Bee Colony Algorithm 533
A. Jamali, I. Z. Mat Darus, M. H. A. Talib, H. M. Yatim, M. S. Hadi, and M. O. Tokhi

39 MG-CMF: A Multi-granularity Capture Matching Features Model for Text Matching 549
Liang Jin and Xiaopeng Cao

40 Detecting Domain Name System Tunneling and Exfiltration from Domain Name System Traffic 561
Yi-Chung Tseng, Ming-Kung Sun, and Wei-An Chen

**41 Activity Recognition Based on Latent Knowledge Mining
in Smart Home 573**
Yu Tong, Rong Chen, and Bo Yu

42 Pattern Retrieval on the Game of Go 587
Shi-Jim Yen and Yu-Chie Ho

About the Editors

Sheng-Lung Peng is a Professor of the Department of Computer Science and Information Engineering, National Dong Hwa University, Taiwan. He is an Honorary Professor at Beijing Information Science and Technology University, China, and a Visiting Professor at Ningxia Institute of Science and Technology, China. He is also an Adjunct Professor at Mandsaur University, India. He serves as the Director of the ICPC Asia Taipei-Hsinchu Site and as Director of the Institute of Information and Computing Machinery; of the Information Service Association of Chinese Colleges; and of the Taiwan Association of Cloud Computing. His research interests are in designing and analyzing algorithms for bioinformatics, combinatorics, data mining, and networks; areas in which he has published over 100 research papers.

Margarita N. Favorskaya is a Professor and Head of the Department of Informatics and Computer Techniques at Reshetnev Siberian State University of Science and Technology, Russian Federation. She has been a member of KES since 2010. She serves as a reviewer for several international journals (Neurocomputing, Knowledge Engineering and Soft Data Paradigms, Pattern Recognition Letters, Engineering Applications of Artificial Intelligence); as an Associate Editor for Intelligent Decision Technologies Journal, the International Journal of Knowledge-Based and Intelligent Engineering Systems, and the International Journal of Reasoning-based Intelligent Systems; and as an Honorary Editor for the International Journal of Knowledge Engineering and Soft Data Paradigms. Having published more than 200 papers, her main research interests are in digital image and video processing, remote sensing, pattern recognition, fractal image processing, artificial intelligence, and information technologies.

Han-Chieh Chao is currently a Professor at the Department of Electrical Engineering, National Dong Hwa University, where he also serves as President. He is also affiliated with the Department of Computer Science and Information Engineering and the Department of Electronic Engineering, National Ilan University, Taiwan; College of Mathematics and Computer Science, Wuhan

Polytechnic University, Wuhan, China; and Fujian University of Technology, Fuzhou, China. His research interests include high-speed networks, wireless networks, IPv6-based networks, digital creative arts, e-Government, and the digital divide. Having published over 400 research papers, he also serves as the Editor-in-Chief for the Institution of Engineering and Technology Networks, the Journal of Internet Technology, the International Journal of Internet Protocol Technology, and the International Journal of Ad Hoc and Ubiquitous Computing. He is a Fellow of the IET (IEEE) and a Chartered Fellow of the British Computer Society.

Part I
Sensor Networks and Communications

Chapter 1

An Improved Clustering Routing Algorithm for Heterogeneous Wireless Sensor Network



Huan Yao

Abstract To prolong the lifecycle of heterogeneous wireless sensor networks, an improved algorithm based on a distributed energy efficient clustering algorithm is proposed. Firstly, the proposed algorithm increases the possibility of the higher energy node to become the candidate cluster head based on the absolute value of the residual energy level. Secondly, the energy consumption rate, the ratio of residual energy, and the initial energy are added into the threshold that is used to control the probability of the node selected as the cluster head. Finally, the mixed data transmission mode is used in the data transmission phase to reduce the consumption of nodes to communicate with the base station. The simulation results show that the proposed algorithm can effectively prolong the network lifecycle and stability period.

1.1 Introduction

Wireless sensor networks (WSNs) are usually applied to battlefield surveillance, fire prevention, industrial management, and agriculture irrigation [1]. In WSNs, one of the main constraints is the battery power limit, which has a great impact on network lifecycle and quality [2]. There are two main strategies to solve this problem. One is to insert a certain proportion of heterogeneous nodes [3], and the other is to apply clustering technology [4]. Therefore, it is necessary to study heterogeneous clustering routing algorithm to prolong the network lifecycle [5].

In the literature, heterogeneous cluster routing technology is considered as one of the most mature energy-saving technologies [6]. In terms of energy heterogeneity, heterogeneous nodes have the ability to perform complex communicational tasks because the heterogeneous nodes have sustainable energy, meaning that it does not need frequent replacement [7]. While these heterogeneous nodes are expensive considering the comprehensive factors of energy efficiency and economy, how to deploy the least heterogeneous nodes reach the best effect is very important [8, 9]. In addition, cluster routing is another strategy that applies in WSNs [10]. However, it may lead

H. Yao (✉)

University of Northeast Electric Power Jilin, Jilin 132012, China
e-mail: lnsyaohuan_0226@163.com

© Springer Nature Singapore Pte Ltd. 2021

S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation, Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_1

to several cluster overloads, resulting in unreasonable cluster formation [11] leading the network to premature death. Therefore, how to design a reasonable heterogeneous clustering routing algorithm is a problem worth researching. The improved clustering routing protocol proposed in this paper is based on heterogeneous WSN (HWSN). The probability and threshold of cluster head selection is modified, and the modified hybrid transmission strategy can improve the case of path loss.

The following parts are shown as: Sect. 1.2 discusses the energy-saving work of heterogeneous cluster wireless sensor networks; Sect. 1.3 introduces the three-tier DEEC protocol; Sect. 1.4 details and describes the proposed protocol; Sect. 1.5 gives and analyzes several simulation results; Sect. 1.6 summarizes the results.

1.2 Related Work

WSNs are popular with many researchers on account of its potential application values. Aiming at the most effective clustering protocol in WSNs, a classical low-energy adaptive clustering hierarchy (LEACH) [12] algorithm is proposed, which is the earliest widely applied clustering algorithm that extends the lifetime of WSNs. In this algorithm, each cluster has a key node, called cluster heads (CHs). All nodes do not transmit data directly to the base station, but some nodes convey information to the local CHs. In each round, the CHs have a certain probability that is computed into circulation. Each round is defined by the establishment phase and the stable phase. This paper also discusses several LEACH-based adaptive clustering protocols, such as LEACH-c [13] and LEACH-m [14]. These methods are isomorphic, but are not suitable for heterogeneous wireless sensor networks. Stable election protocol (SEP) [15] was designed by Smaragdakis in 2004, which is suitable for heterogeneous networks and shows the process of cluster head selection including advanced nodes and normal nodes. It is the earliest protocol that refers to heterogeneity. According to the initial energy of each node, the probability of each node becoming a cluster head is weighted. Therefore, compared with ordinary nodes, the probability of high-level nodes becoming cluster heads is more likely, and the performance of SEP is better than LEACH. Li et al. discussed the distributed efficient clustering (DEEC) algorithm [16] about the two-level and multi-level heterogeneous network. This is developed based on SEP. When filtering cluster heads, nodes with larger initial energy and residual energy are more likely to be selected as cluster heads.

However, in DEEC protocol, it is complicated to calculate the average energy consumption. Elbhiri [17] proposes a developed DEEC (DDEEC) for HWSN. DDEEC introduces the concept of threshold to ensure that the residual energy of high-level nodes after a certain number of rounds is equal to that of ordinary nodes and will not be repeated as CHs. Saini et al. [18] proposed enhanced DEEC (EDEEC) protocol, which extends DEEC to a three-level network by adding super nodes, but the probability of CHs selection is the same as before. Javaid proposes a new clustering routing protocol: an improved distributed energy-saving clustering scheme (EDDEEC) [19] for HWSNs, which improves the probability of cluster head (CH) selection based on

dynamic changes. Xie proposed an improved distributed energy-saving clustering algorithm (IDEEC) [20] for HWSNs in 2017. IDEEC takes into account the multi-level energy model, simplifies the threshold, improves the probability of cluster head selection, and optimizes the average energy in the network. All the above algorithms only change the cluster head selection, without considering the comprehensive factors. The data transmission stage is single hop, which consumes a lot of energy and is not conducive to prolonging the life cycle of the network.

On the basis of the previous research results, the existing DEEC improvement project still has some shortcomings. Its algorithm is more complex in practical application, without considering all the important factors. Therefore, this paper proposes an improved heterogeneous WSN routing protocol, which can effectively utilize energy and then prolong the lifetime of the network.

1.3 The Three-Level Protocol of DEEC

In this section, a three-level distributed efficient clustering (DEEC) algorithm is introduced, which contributes to the research of the subsequent routing protocols. Now, we will discuss energy heterogeneity protocols in detail.

1.3.1 Network Model

The rational assumptions have the following attributes:

- (1) The sole base station has an unlimited supply of power. Therefore, there are no other limitations.
- (2) After deployment, all sensor nodes and base stations are static, and they all have their own identifier (ID) number
- (3) Radio is symmetrical, so data transmission nodes consume the same power as node A.
- (4) The distance between nodes can be calculated based on the received signal strength.
- (5) Three-level nodes are deployed in the network. In addition, they have similar storage, processing, sensing, and communication functions

1.3.2 Energy Model

There are N nodes made up of three types deployed in the network that differ in their initial energy. Normal nodes contain an energy level of E_0 , and the advanced nodes of fraction m have a times extra energy than the normal nodes equal to $E_0 \cdot (1 + a)$.

However, super nodes of fraction m_0 have a factor of b times more energy than the normal nodes. Therefore, $N \cdot m \cdot m_0$ is the total number of super nodes, and $N \cdot m \cdot (1 - m_0)$ is the total number of advanced nodes. The total initial energy of the three-level heterogeneous networks is given by:

$$\begin{aligned} E_{\text{total}} &= N \cdot (1 - m) \cdot E_0 + N \cdot m \cdot (1 - m_0) \cdot (1 + a) \cdot E_0 + N \cdot m \cdot m_0 \cdot (1 + b) \cdot E_0 \\ &= N \cdot E_0 \cdot (1 + m \cdot (a + m_0 \cdot b)) \end{aligned} \quad (1)$$

We can see from (1) that the three-level heterogeneous networks contain $m \cdot (a + m_0 \cdot b)$ times more energy as compared to homogeneous WSNs.

As the author proposed in [21, 22], a node transmits l bit messages to a distance d and the equation to calculate the energy consumption is given by:

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

Furthermore, when a node receives l bit messages, the equation to calculate the energy consumption is given by

$$E_{Rx}(l, d) = lE_{\text{elec}} \quad (3)$$

where E_{elec} signifies the energy dissipation per bit in the transmitter and receiver circuitry, the parameters ε_{fs} and ε_{mp} are the energy consumption per bit in the radio frequency amplifier, d signifies the transmission distance, and d_0 signifies the threshold distance, whose value is given by

$$d_0 = \sqrt{\varepsilon_{\text{fs}} / \varepsilon_{\text{mp}}} \quad (4)$$

The average energy of r round from [22] is given as:

$$\bar{E}(r) = \frac{1}{N} E_{\text{total}} \left(1 - \frac{r}{R}\right) \quad (5)$$

where R denotes the total rounds during the network lifetime and can be estimated as:

$$R = \frac{E_{\text{total}}}{E_{\text{round}}} \quad (6)$$

where E_{round} is the energy dissipated in a network during a single round and it can be calculated as:

$$E_{\text{round}} = l(2NE_{\text{elec}} + NE_{\text{DA}} + k\varepsilon_{\text{mp}}d_{\text{toBS}}^4 + N\varepsilon_{\text{mp}}d_{\text{toCH}}^2) \quad (7)$$

where k is the number of clusters, E_{DA} is the data aggregation cost expended in CH, d_{toBS} is the average distance between CH to BS, and d_{toCH} is the average distance between cluster members to CH. Now, d_{toBS} and d_{toCH} can be calculated as:

$$d_{\text{toCH}} = \frac{M}{\sqrt{2\pi k}} \quad (8)$$

$$d_{\text{toBS}} = 0.765 \frac{M}{2} \quad (9)$$

Through finding the derivative of E_{round} with respect to k , and setting the derivative as zero, we can get the optimal number clusters k_{opt} as:

$$k_{\text{opt}} = \frac{\sqrt{N}}{\sqrt{2\pi}} \sqrt{\frac{\varepsilon_{\text{fs}}}{\varepsilon_{\text{mp}}} \frac{M}{d_{\text{toBS}}^2}} \quad (10)$$

Hence, we can calculate the energy dissipated per round by substituting Eqs. (8), (9), and (10) into (7). When the number of the clusters is k_{opt} , the total energy consumption of the entire network is minimal. So, we can determine the optimal cluster head ratio P_{opt} by

$$P_{\text{opt}} = \frac{k_{\text{opt}}}{N} \quad (11)$$

1.3.3 Cluster Heads Selection Method

In DEEC, each round node decides whether to become a cluster head based on the threshold calculated by the suggested percentage of cluster heads for the network. This decision is that the nodes automatically generate a random number between 0 and 1. If the number is less than the threshold $T(S_i)$, then the node becomes a cluster head for the current round. The threshold is set as:

$$T(S_i) = \begin{cases} \frac{p_{\text{norm}}}{1 - p_{\text{norm}} \left(r \bmod \frac{1}{p_{\text{norm}}} \right)} & S_{\text{norm}} \in G' \\ \frac{p_{\text{adv}}}{1 - p_{\text{adv}} \left(r \bmod \frac{1}{p_{\text{adv}}} \right)} & S_{\text{adv}} \in G'' \\ \frac{p_{\text{sup}}}{1 - p_{\text{sup}} \left(r \bmod \frac{1}{p_{\text{sup}}} \right)} & S_{\text{sup}} \in G''' \\ 0 & \text{Otherwise} \end{cases} \quad (12)$$

where p_{norm} , p_{adv} , and p_{sup} are the weighted election probabilities for normal, advanced, and super nodes. G' , G'' , and G''' stand for the sets for which relevant

nodes have not become cluster heads within the last $1/p_{\text{norm}}$, $1/p_{\text{adv}}$, and $1/p_{\text{sup}}$ rounds.

The weighted election probabilities p_{norm} , p_{adv} , and p_{sup} are given by

$$p_i = \begin{cases} \frac{P_{\text{opt}} E_i(r)}{(1+m(a+m_0b))\bar{E}(r)} & \text{for Nrm nodes, if } E_i(r) > T_{\text{abs}} \\ \frac{P_{\text{opt}}(1+a)E_i(r)}{(1+m(a+m_0b))\bar{E}(r)} & \text{for Adv nodes, if } E_i(r) > T_{\text{abs}} \\ \frac{P_{\text{opt}}(1+b)E_i(r)}{(1+m(a+m_0b))\bar{E}(r)} & \text{for Sup nodes, if } E_i(r) > T_{\text{abs}} \end{cases} \quad (13)$$

We can substitute (13) into (12), and the threshold formula of the final cluster head is obtained.

where T_{abs} is the threshold of energy, which can be expressed by:

$$T_{\text{abs}} = zE_0 \quad (14)$$

where E_0 is the initial energy of the normal node and the range of $z = 0$ is 0 to 1. If $z = 0$, we will have the previous DEEC algorithm, the threshold to be a cluster head will be equal to zero for all nodes, and all advanced nodes and super nodes have few opportunities to be a cluster head. So, they go to transmit data directly to the base station in a large part; thus, they die quickly. Through lots of simulations with a random topology, the value z directly controls the cluster heads number. If z is higher, the cluster heads will increase. But, it cannot increase without limit, and it will lead to the phenomenon that cluster head overloads. To solve this problem, put the reasonable value that it is worth researching. And in the following experiment, it is researched that when z is 0.74, the network performance is optimal.

1.3.4 Setup of the Cluster and Data Transmission

First filter the candidate CH in the first round. When the selected candidate CH is formed, it will send the data to the competing CH within the competition radius. At the same time, the member node can get the distance to join the cluster according to the received signal strength and can get the minimum value according to the received strongest signal information. The cluster member nodes send data to the cluster head according to the time slot of TDMA. Then, each cluster head receives data, aggregates it, and forwards it to the base station. The data transmission mode is single hop mode. After completing one round of data transmission, new cluster head selection will begin in the next round.

1.4 Proposed Algorithm

1.4.1 Cluster Head Selection

1.4.1.1 The Improvement of Probability

From (13), it can be seen that nodes with large residual energy on the circle are more likely to become cluster heads. Therefore, nodes with higher residual energy or higher nodes are more likely to be selected as cluster heads than those with lower energy or normal nodes. However, when the residual energy of super node, advanced node, and ordinary node is almost the same, the probability of choosing super node or advanced node as a cluster head is still higher than that of choosing ordinary node. This makes super and advanced nodes to have less residual energy than normal nodes, and they will die faster than normal nodes. To avoid this unbalanced problem, the probability formulas are changed. All the nodes will use their respective probability formula until they reach a threshold energy level T_{abs} . The paper adds a condition that when the residual energy of all three-level nodes reaches the threshold T_{abs} or below, the cluster head elections can still work efficiently, and then, all the nodes will use a common probability formula as given below:

$$p_i = c \frac{P_{\text{opt}}(1+b)E_i(r)}{(1+m(a+m_0b))\bar{E}(r)}, \text{ for Nor, Adv, Sup, if } E_i(r) \leq T_{\text{abs}} \quad (15)$$

where c is a positive value which adjusts the probability of cluster head election, the range of c is 0 to 1, and it will be shown in Part 5 (Simulation and Discussion). Formula (15) is an improvement for the election probability of cluster heads. It can well improve the trend of rapid decline of energy of high-energy nodes when the residual energy of all nodes is less than the threshold value. Therefore, the probability of improved nodes makes the energy consumption more balanced.

Thus, the whole improved probability is expressed as (16):

$$p_i = \begin{cases} \frac{P_{\text{opt}}}{(1+m(a+m_0b))} & \text{for Nrm nodes, if } E_i(r) > T_{\text{abs}} \\ \frac{P_{\text{opt}}(1+a)E_i(r)}{(1+m(a+m_0b))\bar{E}(r)} & \text{for Adv nodes, if } E_i(r) > T_{\text{abs}} \\ \frac{P_{\text{opt}}(1+b)E_i(r)}{(1+m(a+m_0b))\bar{E}(r)} & \text{for Sup nodes, if } E_i(r) > T_{\text{abs}} \\ c \frac{P_{\text{opt}}(1+b)E_i(r)}{(1+m(a+m_0b))\bar{E}(r)} & \text{for Nor, Adv, Sup, if } E_i(r) \leq T_{\text{abs}} \end{cases} \quad (16)$$

Formula (16) completely describes the cluster head election probability under different conditions for the three node types, where P_{opt} represents the optimal cluster head election probability, and a and b , respectively, represent the multiples of the energy of advanced node and super node higher than that of ordinary node.

1.4.2 The Improvement of Threshold

Starting from (12), the threshold depends only on probability and integer. However, due to the heterogeneity of nodes in the network, the initial energy and energy consumption rate of nodes are different. Therefore, it is unreasonable to choose a cluster head based on the above factors. We consider natural factors, including the speed of node energy consumption and residual energy.

We believe that the initial energy of nodes should have different probability of cluster head selection, and the corresponding threshold should reduce the lower initial energy of premature death to some extent. However, in the actual operation of the network, only considering the initial energy cannot solve the problem of unbalanced energy consumption. The residual energy of nodes in the network undergoes dynamic changes. Some nodes usually act as cluster heads, which makes the stability period of the network very short. Therefore, we need to consider the residual energy factor. In addition, the initial energy of the node is higher, but the energy consumption is higher. Therefore, if only considering the high residual energy factor, the nodes will consume energy quickly in the next round, which is not the best strategy to select cluster heads.

From another point of view, if only one factor of node energy consumption rate is considered, there is a problem of loss of energy consumption rate of the node in the first round. In this case, the threshold will become zero. As a result, all nodes become cluster heads, and network performance decreases dramatically.

In order to realize the former method without considering node energy consumption rate and residual energy consumption, we modify the threshold, which adds influence factor η , which is multiplied by the original probability, and proposes that:

$$T(S_i) \begin{cases} \frac{p_{nm}}{1-p_{nm}\left(r \bmod \frac{1}{p_{nm}}\right)} * \eta_{nm} & p_{nm} \in G' \\ \frac{p_{adv}}{1-p_{adv}\left(r \bmod \frac{1}{p_{adv}}\right)} * \eta_{adv} & p_{adv} \in G'' \\ \frac{p_{sup}}{1-p_{sup}\left(r \bmod \frac{1}{p_{sup}}\right)} * \eta_{sup} & p_{sup} \in G''' \\ 0 & \text{Otherwise} \end{cases} \quad (17)$$

The η is a factor, and the formula is given by

$$\eta = \alpha \frac{E_i(r)}{E_{ini}} + \beta \frac{r-1}{E_{ini} - E_i(r)} \quad (18)$$

There is also the practical significance of $\frac{r-1}{E_{ini} - E_i(r)}$, which is the reciprocal of the energy consumption rate. In addition, E_{ini} , $E_i(r)$, and r represent the initial energy, residual energy, and the current round, while α and β are the weight factor which adjust the factors that $\frac{E_i(r)}{E_{ini}}$ and $\frac{r-1}{E_{ini} - E_i(r)}$.

We can also draw the forecast conclusion by determining that the cluster head selection in the next round is based on the η of the previous round. Then, the node

will have a greater chance to become a cluster head in the next round. The η is an advantage over the previous round, which contributes to filter the cluster head.

1.4.3 State Transition

According to the characteristic that the node energy in the wireless sensor network is limit, considering that it will consume plenty of energy since each node needs to join in the cluster header's election, we adopt reasonable dormancy mechanism during the process of cluster head election.

In the process of forming clusters, there is the dedupe-aware neighboring node relationship between the node and node in the network. In our protocol, each node is determined by its state of transition with respect to its own energy state and the state of its neighboring node in the future. Therefore, we describe the dormancy mechanism in detail. When the node is not dead, the state of the node is alive, and the state flag of node is "active." Otherwise, the state flag of node is "dead." Once it is found that the neighbor node is in the dead state, we will set the neighbor node to sleep state, which meant that the neighboring node turned off its radio for the same time period to save energy. If a node was isolated, it would remain in active mode for the entire network lifetime. Based on the collaboration of neighboring nodes, we propose Algorithm I to describe how to execute the dormancy mechanism.

Algorithm I. The demonstrate of dormancy mechanism.

Initialization: all the nodes

```

1: if (node.neighbour_flag==1)
2:   if (node.state==Active)
3:     send the data to the base station;
4:     if (node.neighbour!=dead)
5:       node.state=Sleep;
6:     end;
7:   end;
8:   if (node.state==Sleep)
9:     node.state==Active;
10:  end
11: end
12: if (node.neighbour_flag==0)
13:   send the data to the base station;
14: end;

```

1.4.4 Data Transmission

For data transmission, the previous algorithms used single hop mode. When the node is far away from the base station, it consumes unnecessary energy, so we adopt hybrid transmission mode. When member nodes are close to base stations, single hop mode should be used. When it is far away from the base station, multi-hop mode is adopted.

In the actual situation of cluster routing protocol, multi-hop transmission mode can share energy consumption with multiple clusters instead of focusing on one cluster. It can effectively balance the energy consumption inside the cluster.

In our improved protocol, each normal node chooses a routing path to send data to the base station. The path choice depends on the weight of distance between nodes. To reduce the overhead and delay, each node sends sensing data to the base station in the path of minimum energy consumption. Paths should be established directly or through transit nodes to forward aggregated data to the base station.

Each node first estimates the communication energy consumed by sending l bit messages directly to the base station. Its value depends on:

$$E_{CHi_to_BS} = E(l, d_{CHi_to_BS}) \quad (19)$$

where $d_{CHi_to_BS}$ is the distance between the node i and the BS(base station). The value of $d_{CHi_to_BS}$ is given by:

$$d_{CHi_to_BS} = \sqrt{(x_{BS} - x_{CHi})^2 + (y_{BS} - y_{CHi})^2} \quad (20)$$

, each node decides whether to find an intermediate node or not in the current round. This decision depends on the following conditions:

$$E_{CHi_to_BS} \geq E(k, d_{CHi_to_CHj}) + E(k, d_{CHj_to_BS}) \quad (21)$$

where $d_{CHi_to_CHj}$ is the distance between the node i and the node j , and $d_{CHj_to_BS}$ is the distance between node j and the base station. The value of $d_{CHi_to_CHj}$ and $d_{CHj_to_BS}$ is given by

$$d_{CHi_to_CHj} = \sqrt{(x_{CHj} - x_{CHi})^2 + (y_{CHj} - y_{CHi})^2} \quad (22)$$

$$d_{CHj_to_BS} = \sqrt{(x_{BS} - x_{CHj})^2 + (y_{BS} - y_{CHj})^2} \quad (23)$$

The mixed data transmission scheme is presented in Fig. 1.1.

1.5 Simulation and Discussion

In our simulations, we deployed 100 nodes in a 100×100 square meter region in which the base station was located in the center of the region. We initialized p_{opt} to 0.1 depending on k_{opt} , which was given by Eq. (11). The initialized energy of the normal node is 0.5 Joules. Through these simulations, we can obtain the appropriate parameter that α is 0.35 and β is 0.65. We simulated the new protocol by using MATLAB. The values of radio characteristics are set as same as [16]. Those

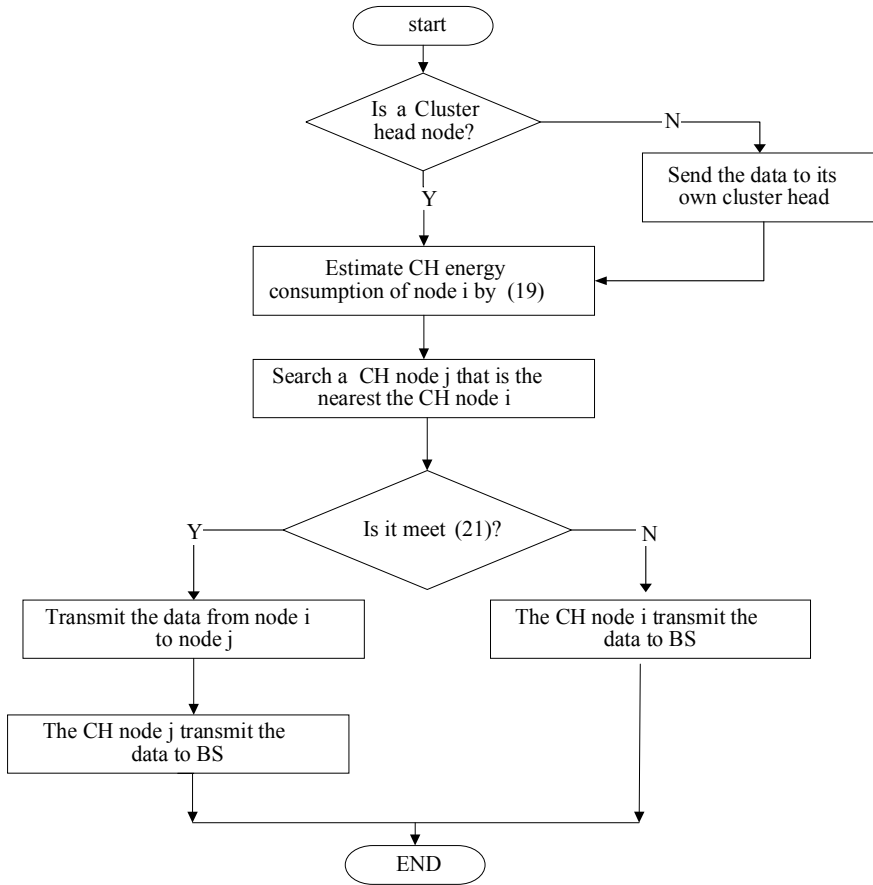


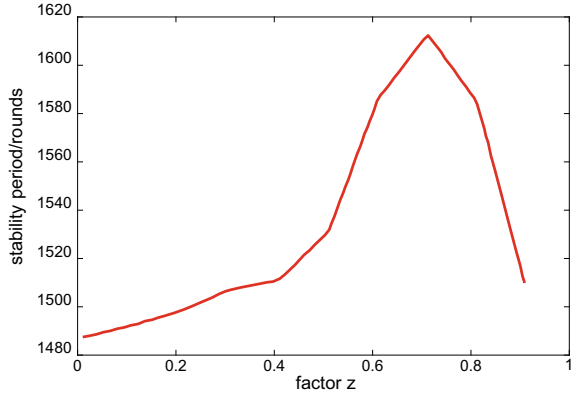
Fig. 1.1 Flowchart of mixed data transmission scheme

parameters are given in Table 1.1. We employ 20% advanced nodes and 20% super nodes. The energy of advanced nodes is 1.5 times more than normal nodes, while the

Table 1.1 Simulation Parameters

Parameters	Values
d_0	87 m
ϵ_{fs}	10 pJ/bit/m ²
ϵ_{mp}	0.0013 pJ/bit/m ⁴
E_{elec}	50 nJ/bit
E_{DA}	5 nJ/bit
l	4000 bit

Fig. 1.2 Relationship between the stability period and the parameter z



energy of super nodes is three times more than normal nodes ($m = 0.50, m_0 = 0.50, a = 1.50, b = 3.0$).

We will now discuss the performance of our proposed DEEC and DDEEC methods on the basis of several evaluation indicator. This work uses four different performance indicators to evaluate the performance of our algorithm. The variances of stability period, network life cycle, throughput, and cluster head number are described as follows:

- (1) Stability period: This is the death time of the first node.
- (2) Network life cycle: Usually designated as the total time for the full operation and functionality of the network to perform dedicated tasks. Take the integer of all nodes to die, that is, the last node in the network to die.
- (3) Packets sent to base stations: This is the total number of messages received by base stations over a period of time, during which each node sends a message to the cluster head or base station in each round.
- (4) Number of cluster heads: This is the number of cluster heads in each round.

Figure 1.2 illustrates the relationship between the stability period and the parameter z in the proposed DEEC. As can be seen from Fig. 1.2, the parameter z varies with the stability period, when it is near to 0.74, at the same time as the network performance—the stable period reaches the optimal level. In order to simplify follow-up research, we can set the parameter z to 0.74.

Figure 1.3 illustrates the relationship between the network lifetime and the parameter c in the proposed DEEC. As can be seen from Fig. 1.3, the parameter varies with the stability period, when it near to 0.02, at the same time as the network performance—the network lifetime reaches the optimal level. In order to simplify follow-up research, we can set the parameter c to 0.02.

Figure 1.4 illustrates the relationship between the number of active nodes and the number of rounds. As can be seen from Fig. 1.4, the first nodes of DEEC, DDEEC, and the proposed method begin to die at 1095, , and 1493, respectively. This means that this method has a longer stability period than other methods. As can be seen

Fig. 1.3 Relationship between the network lifetime and the parameter c

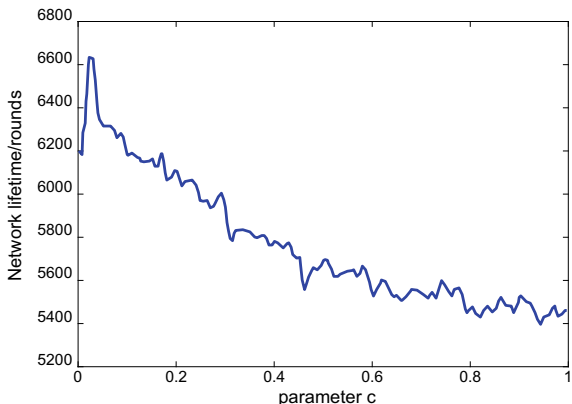
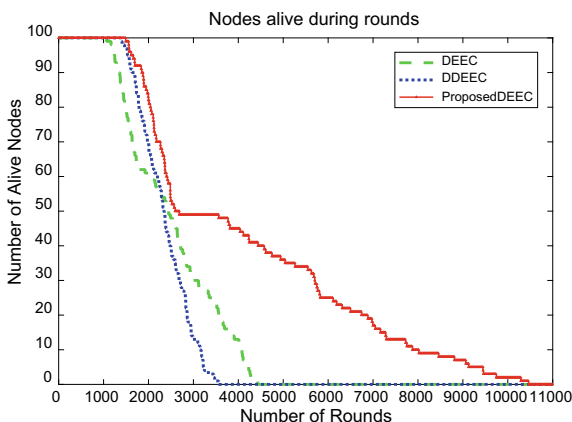


Fig. 1.4 Number of nodes alive



from Fig. 1.4, the last nodes of DEEC, DDEEC, and the algorithm in this paper begin to die in 3569, 4350, and 10,501 rounds, respectively. This means that this method has a longer network life cycle than other methods.

Figure 1.5 illustrates the relationship between the packets sent to the base station and the round. From Fig. 1.5, the packets sent to the base station for DEEC, DDEEC, and the proposed method are 0.75×10^5 , 1.2×10^5 , and 4.2×10^5 . We can see that the packets sent to the base station for the proposed algorithm are still more than the others in the whole network lifetime. This means that the proposed method is more efficient than the others in terms of data transmission. It is necessary for the improvement measures that we adopt the mixed transmission mode in the data transmission phase. These cluster heads will deliver packets to the base station and thereby increase throughput.

Figures 1.6, 1.7, and 1.8 show the relationship between the number of cluster heads generated by DEEC, DDEEC, and the proposed algorithm throughout the network time. As can be seen from Fig. 1.4, the stabilization periods of DEEC, DDEEC, and

Fig. 1.5 Number of messages

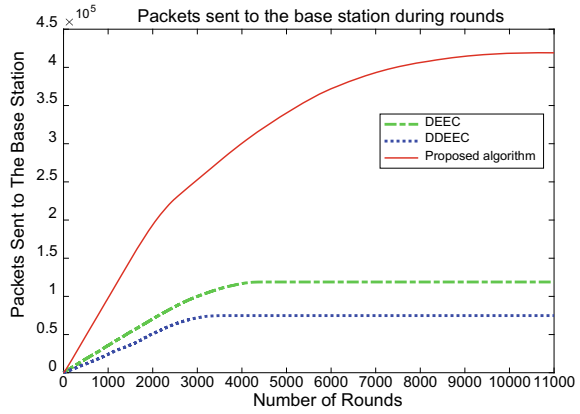


Fig. 1.6 Number of cluster heads of DEEC

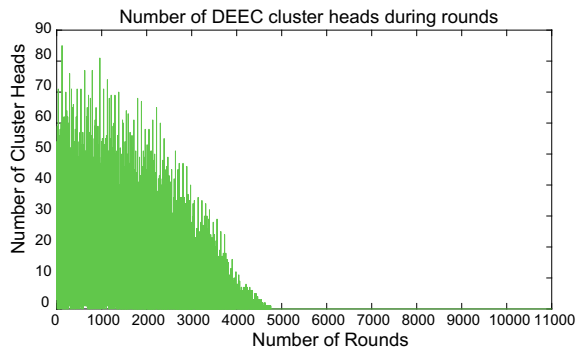
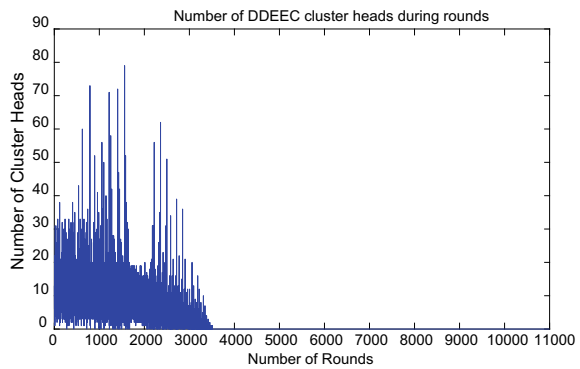
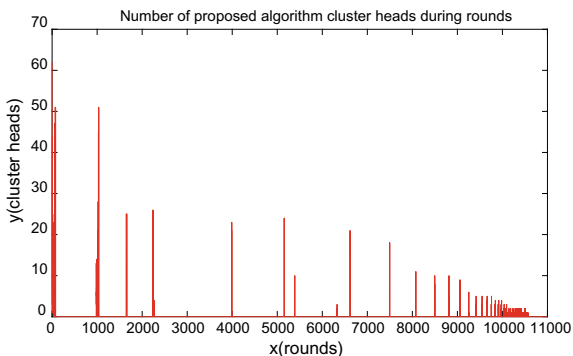


Fig. 1.7 Number of cluster heads of DDEEC



the proposed method are 0-1095, 0-1350, and 0-1493 rounds, respectively. Based on the above stable period, we can see that the number of cluster head DEEC is more balanced, followed by DDEEC and stable period. The method and number of cluster heads are more stable and unstable than other clusters. This means that the energy

Fig. 1.8 Number of cluster heads of proposed method



consumption of the method is more balanced than that of the method during the unstable period. The reason is that improved clustering techniques, such as adaptive weighted selection probability by adding conditions and positive calculation, save a lot of energy in the protocol. Therefore, we can conclude that the algorithm can save more energy and prolong the network life, and the stability is more effective.

It can be seen from the figure that the algorithm is superior to DEEC and DDEEC in energy efficiency and energy balance. This is because our improved algorithm reduces the fluctuation of cluster heads. It needs to consider the speed of residual energy and energy consumption to improve the probability threshold to select cluster heads, and also uses hybrid transmission mode in the data transmission phase. Therefore, the protocol helps prolong the network lifetime.

From all the figures, we can see that they show that the proposed algorithm is better in terms of energy efficiency and energy balance than DEEC and DDEEC. This is because our improved algorithm reduces the fluctuation of the cluster heads by taking residual energy and the rate of energy consumption into consideration to improve the probability threshold during the period of choosing cluster heads, and also by adopting the mixed transmission mode in the data transmission phase. Thus, the proposed protocol is helpful in extending the network lifetime and stability time.

1.6 Conclusion

An improved, distributed and efficient clustering algorithm for HWSNs was proposed. The probability threshold and cluster head selection probability are improved, and the hybrid data transmission mode is adopted in the data transmission stage, which reduces the energy consumption of nodes. The simulation results show that when compared with DEEC and DDEEC, the algorithm can effectively prolong the stability time and lifetime of the network.

References

1. Jingxia, Z., Junjie, C.: An adaptive clustering algorithm for dynamic heterogeneous wireless sensor networks. *Wireless Netw.* **25**(2), 455–470 (2019)
2. Tungchun, C., Chihan, L., Kate Chingju, L.: Traffic-aware sensor grouping for IEEE 802.11ah networks: regression based analysis and design. *IEEE Trans. Mobile Comput.* **99**(1):674–687 (2019)
3. Rani, R., Kakkar, D., Kakkar, P.: Distance based enhanced threshold sensitive stable election routing protocol for heterogeneous wireless sensor network. *Comput. Intell. Sensor Netw.* **776**(5), 101–122 (2019)
4. Yang, L., Yinzi, L., Yuanchang, Z.: An unequal cluster-based routing scheme for multi-level heterogeneous wireless sensor networks. *Telecommun. Syst.* **68**(10), 1–16 (2018)
5. Kiyumi, A., Raja, F., Chuan, H.: Fuzzy logic-based routing algorithm for lifetime enhancement in heterogeneous wireless sensor networks. *IEEE Trans. Green Commun. Netw.* **2**(5), 517–532 (2018)
6. Huarui, W., Huaji, Z., Yiheng, M.: An energy efficient cluster-head rotation and relay node selection scheme for farmland heterogeneous wireless sensor networks. *Wireless Pers. Commun.* **101**(5), 1639–1655 (2018)
7. Li, C., Bai, J., Gu, J.: Clustering routing based on mixed integer programming for heterogeneous wireless sensor networks. *Ad Hoc Netw.* **72**(3), 81–90 (2018)
8. Naranjo, P., Shojafar, M., Mostafaei, H.: P-SEP: a prolong stable election routing algorithm for energy-limited heterogeneous fog-supported wireless sensor networks. *J. Supercomput.* **73**(2), 733–755 (2017)
9. Shrivastav, K., Kulat, K.D.: Energy efficient clustering of statistically distributed heterogeneous wireless sensor networks for Internet-of-Things. *SSRN Electron. J.* **32**(12), 45–50 (2018)
10. Sharma, D., Bhondekar, A.P.: Traffic and energy aware routing for heterogeneous wireless sensor networks. *IEEE Commun. Lett.* **99**(20), 1 (2018)
11. Suniti, D., Agrawal, S., Vig, R.: Cluster-head restricted energy efficient protocol (CREEP) for routing in heterogeneous wireless sensor networks. *Wireless Pers. Commun.* **100**(7), 1–21 (2018)
12. Lin, C., Ruolin, F., Kaigui, B.: On heterogeneous neighbor discovery in wireless sensor networks. *Ad Hoc Netw.* **44**(7), 693–701 (2018)
13. Xiangning, F., & Yulin, S.: Improvement on LEACH protocol of wireless sensor network. *Appl. Mech. Mater.* 341–342 (2013)
14. Liao, Q., Zhu, H.: An energy balanced clustering algorithm based on LEACH protocol. *Appl. Mech. Mater.* 347–350 (2013)
15. Smaragdakis, G., Matta, I., Bestavros, A.: SEP: a stable election protocol for clustered heterogeneous wireless sensor networks. In: 2nd international workshop proceeding on sensor and actor network protocol and applications, pp. 1–2. SANPA (2004)
16. Twinkle, T., Roy, N.R.: Modified DEEC: a varying power level based clustering technique for WSNs. In: 2015 international conference on computer and computational sciences (ICCCS), pp. 27–29. Noida, Inida (2015)
17. Elbhiri, B., Saadane, R., Fldhi, S.E.: Developed distributed energy-efficient clustering (DDEEC) for heterogeneous wireless sensor networks. In: International Symposium on I/v Communications & Mobile Network (2010)
18. Saini, P., Sharma, A.K.: E-DEEC: enhanced distributed energy efficient clustering scheme for heterogeneous WSN. In: International Conference on Parallel Distributed & Grid Computing, pp. 28–30, Solan, India (2010)
19. Javaid, N., Qureshi, T.N., Khan, A.H.: EDDEEC: enhanced developed distributed energy-efficient clustering for heterogeneous wireless sensor networks. *Procedia Comput. Sci.* 914–919 (2013)
20. Benyin, X., Chaowei, W.: An improved distributed energy efficient clustering algorithm for heterogeneous WSNs. In: IEEE Wireless Communications and Networking Conference (WCNC), pp. 19–22, San Francisco, CA (2017)

21. Rehman, O., Javaid N., Manzoor, B.: Energy consumption rate based stable election protocol (ECRSEP) for WSNs. *Procedia Comput. Sci.* 932–937 (2013)
22. Dhand, G., Tyagi, S.S.: SMEER: secure multi-tier energy efficient routing protocol for hierarchical wireless sensor networks. *Wireless Pers. Commun.* **105**(19), 17–35 (2019)

Chapter 2

Improvement of RPL Routing Strategy Based on 6LoWPAN



Hui Wang, Zhihui Fan, Xiangfu He, Peiyu Li, and Chaowu Zhang

Abstract The routing protocol for low-power and lossy networks (RPL) has become the standard routing protocol for the Internet of things (IoT). This paper studies the application of RPL in the 6LoWPAN. In the topology construction process of the existing R_RPL routing strategy, the problem of the network depth of the node is not considered. This paper proposes an objective function optimization scheme named RR_RPL that combines energy consumption, hop count, ETX, and other metric factors and sets different weights for the parent nodes of different energy consumption. Simulations are done using Cooja with random waypoint mobility scenario, and the results show that the RR_RPL has higher network lifetime and PDR compared to related existing protocols.

2.1 Introduction

Due to the unstable link characteristics of the 6LoWPAN network and the need for low power consumption, the traditional routing protocol is not suitable for 6LoWPAN, so a special network protocol must be developed. To this end, the IETF established a special working group routing over low-power and lossy networks (ROLL) for the development of a low-power lossy network routing protocol (RPL) in compliance with the 6LoWPAN specification and introduced the RPL routing protocol in 2012. The first specification is RFC6550. The specification mainly describes the message frame format, network creation mode, and communication type of the 6LoWPAN network [1]. In addition, the ROLL working group also agreed on the corresponding routing specifications for practical applications in different fields [2]. The RPL routing protocol mainly solves three major problems:

H. Wang · Z. Fan (✉) · P. Li
Network Communication Technology Institute, Henan University of Science and Technology,
Luoyang, China
e-mail: fzh@haust.edu.cn

X. He · C. Zhang
School of Information Engineering, Henan University of Science and Technology, Luoyang, China

- (a) Load balancing ensures that most nodes in the PAN maintain low energy consumption.
- (b) Loop avoidance and ability to converge quickly.
- (c) Set congestion control mechanism to control traffic transmission.

Load balancing is achieved by designing the corresponding objective function. The loop avoidance mechanism is implemented by the rank mechanism and DODAG. Congestion control is implemented by trickle.

The RPL routing construction process is the DODAG construction process. The DIO, DIS, DAO, and DAO-ACK messages are used to complete the networking. The routing metric determines the information carried by these types of messages. The routing metric consists of the RPL routing strategy. It is also implemented by the objective function. According to the strategy, OF can have multiple definitions, but the same type of OF can only be used in the same RPL instance. The objective function has always been the focus of people's research. By combining different metrics and routing constraints to form different objective functions (object functions, OF), a destination-oriented directed acyclic graph (DODAG) is constructed to achieve the improved link transmission efficiency and energy consumption reduction. Since the energy consumption and link quality of nodes in 6LoWPAN have always been the focus of attention, this paper will also reduce the energy consumption, extend the network life cycle, and reduce the packet loss rate. Existing objective functions are researched and improved.

In this paper, network depth is used as a routing metric to balance the relationship between node energy consumption and link quality, and finally create a load-balanced topology. The improvements brought by this paper are:

1. In this paper, expected transfer count (ETX), energy consumption, and hops are used as the metrics of the objective function to construct an objective function of multiple metrics.
2. The value of the hop count is used as a weight to measure the depth of the network where the node is located. Eventually, the node can select the parent node according to the depth.
3. The new routing strategy proposed in this paper balances link quality and node residual energy to achieve load balancing.

2.2 Related Work

2.2.1 *Basic Objective Function in Routing Protocol for Low-Power and Lossy Networks*

The objective function is not a single function, but a collection of functions. All nodes in the same RPL instance use the same objective function. The main task is to establish the selection of the parent node and the calculation of the rank value through

the objective function in the process of establishing the RPL routing topology, so as to achieve different networking purposes. This process is triggered when the node detects an event that may occur with neighbor node information updates (node failure, timer expiration, DIO message received).

The calculation method of the rank value is shown in Eq. 2.1:

$$\text{Rank}_c = \begin{cases} \text{Rank}_p + \text{Rank}_{\text{increase}}, & \text{non - root} \\ 0, & \text{root} \end{cases} \quad (2.1)$$

Rank_c is the rank value of the child node, Rank_p is the rank value of the candidate parent node, and $\text{Rank}_{\text{increase}}$ is the rank increment.

When the network starts, the root node RANK value is defined as 0, and the non-root node calculates its own rank value by obtaining the parent node's rank value and then adding the rank increment (depending on its own DODAG), and the rank increment cannot exceed threshold, if the new rank calculation result is greater than the existing rank value, the parent node will not be changed.

1. OF0

OF0 [3] is the default routing function of RPL. It can be known from RFC6552 that it uses the hop count as the main metric to make the decision of the parent node. The smaller the hop count of the candidate parent node from the root node, the more likely it is to become. The optimal parent node of the child node calculates the possible rank value of the child node after selecting the candidate parent node and takes the smaller one as the best parent node. The value can approximate the relative distance between the node and the root node. The calculation of the rank value R_c of the child node is 1 and Eq. 2.2:

$$\text{Rank}_{\text{Increase}} = (R_f * S_p + S_r) * \text{MinHopRank}_{\text{Increase}} \quad (2.2)$$

where R_f is the expansion factor, in order to adjust the weight of S_p , the default is 1; S_p is the ladder value, which can be set to be related to the link state between the node and the parent node, the default is 3; the maximum value of S_r is the largest S_p value. The amount of growth, this variable is set to maintain routing diversity, to ensure that there is at least one valid next node, the default is 0, the maximum is 5; $\text{MinHopRank}_{\text{Increase}}$ is set to the minimum increment between the child node and the parent node in DODAG, the value is provided by the root node, the default is 1.

It is not difficult to see from the formula that the method of calculating the hop count alone may result in the selection of the node with poor link quality as the parent node. Although OF0 provides the expansion factor such as S_p , if it is dynamically assigned, the routing may be increased. Diversity, but it is also very likely to cause many unstable factors to the network, so it needs to be verified by enough experiments to modify it [4].

2. ETX [5]

The goal of the ETX design is to calculate the expected number of times the node needs to successfully send data to the destination. The total ETX of a link is the sum

of ETX between the nodes on the link. The ETX value of the link is calculated by the DIO message. When a node receives a DIO message from a node of another link, and the ETX included in the DIO message changes, the ETX value of the link is recalculated and compared with the ETX of the current link. Finally, select the best parent node.

The calculation formula of ETX is shown as Eq. 2.2:

$$ETX = \frac{1}{PRR_{up} * PRR_{down}} \quad (2.3)$$

PRR_{down} is the delivery rate of the downlink direction of the data packet (the reception rate of the ACK packet), and PRR_{up} is the delivery rate of the uplink reverse direction of the data packet. It can be seen that the higher the two-way delivery rate, the smaller the value of ETX, indicating that the quality of the link is higher. The two nodes periodically send probe data packets to complete the PRR data collection. The PRR calculation formula is shown in Eq. 2.4.

$$PRR = \frac{S_{a \rightarrow b}}{T_{a \rightarrow b}} \quad (2.4)$$

where S is the number of packets successfully delivered between nodes a and b , and T is the total transmitted packet data.

3. MRHOF [5]

Unlike OF0, MRHOF is an objective function based on the hysteresis mechanism. By default, the selected metric is the ETX value of the link, which is related to the path cost. In the case of network stability, the node with the lowest link cost is selected as the node, parent node. However, MRHOF weakens the response speed of the objective function of pure ETX. On the relatively stable link of ETX, the hysteresis mechanism is used to weaken the influence of small changes of ETX on topology reconstruction.

The idea of MRHOF is that when a new candidate parent node appears, only the link cost is less than the current optimal link, and the difference of the rank value is greater than the set threshold, the current node will select the node as its own optimal parent. Node, this is the lag mechanism of MRHOF, which can reduce the jitter of the network to a certain extent, thus meeting the characteristics of the scene [6].

2.2.2 Routing Decision Research

In terms of loop avoidance, a routing loop avoidance mechanism is mentioned in [7]. Although this scheme has certain effects on loop avoidance under certain conditions, after the sensor network scale is further increased. The loop avoidance mechanism that continuously detects the loop has a more serious impact on the network than the

route loop. Therefore, for an RPL network with a large number of nodes, the loop avoidance mechanism is unrealistic.

In terms of load balancing, Michel et al. combine the opportunity routing mechanism to adjust the wake-up period of the intermediate nodes in the link, thereby achieving the goal of reducing node energy consumption [8]. However, the algorithm does not take into account other situations in the packet transmission process, so there is no significant optimization effect on the life cycle of the entire network.

In the introduction of new metrics, Capone et al. added the transmission power factor of the node, combined with the original link quality and the energy consumption of the nodes, combined with functional operations to form a new metric L2AM, which is selected during node decision making. The metric is smaller as the parent node [9]. Lova et al. join the data transmission factor between nodes, combined with the original link quality and node energy consumption, and selects the smaller measurement as the decision basis [10].

In terms of routing strategy, Chang et al. proposed the energy-efficient algorithm, which combines the remaining power and the expected number of transmissions with the candidate parent nodes [11], but there is no mechanism to dynamically adjust the weights of the two.

In terms of multi-path transmission, Lova et al. improve the objective function of the ELT in the literature [12] and determine the parent node according to the ELT value. This scheme can improve the utilization of the node to a certain extent, but it is easy to cause the ring. The road, in turn, causes a packet loss. In [13], Gaddour et al. proposed an objective function which combining the multi-path mechanism. Although the experimental results show that the scheme has a significant improvement on the packet delivery rate and the overall energy consumption of the network under a small-scale network, network congestion will still occur frequently.

In terms of reducing node power consumption, Rukpakavong et al. proposed a routing algorithm based on node transmission power consumption, which calculates the minimum power consumption consumed by a sub-node to transmit a data packet, and uses this as a measure factor to select the node on the path with the lowest energy consumption which is the parent node [14], so as to achieve the effect of prolonging the network life cycle. Lova et al. proposed a multi-parent path selection scheme, and the multi-parent node strategy improves the network stability and reduces the overall energy consumption of the network [15]. However, when the node leaves or joins in the network and causes topology changes, incoming to repair and rebuild the routing state will cause a large number of data packets to be filled in the network, thereby increasing the burden on the node. Therefore, the solution is only suitable for a network with stable node conditions.

Chang et al. proposed an energy-oriented RPL routing mechanism R_RPL [16], combining the residual energy index with ETX and setting the fixed weight to make the network more balanced. According to Lassouaoui L et al., the results show that the R_RPL protocol has obvious advantages compared to the objective function OF0 or ETX of a single metric factor [17, 18].

2.2.3 RPL Performance Test

Hakeem et al. conducted a simulation test on the performance of the RPL routing protocol in packet forwarding and networking which is based on the Cortex M3-nodes node [19]. In the network with about 150 nodes, the experimental results show that the actual application requirements can be met, and the network operation is basically stable.

Gaddour et al. carried out simulation tests on the packet delay, packet loss rate, and energy consumption of the RPL protocol [20]. The results show that as the network scale increases and the number of nodes increases, the protocol cannot cope with the impact of packet growth, and resulting in a sudden drop in performance.

Lova et al. make a detailed test analysis of the RPL's DIO packet transmission and the end-to-end delay in the network [21]. The results show that the network is started when there is a problem with the link in the network. The repair mechanism will cause a large number of control messages to be flooded on the link. The growth of control packets will greatly reduce the stability of the network and increase the overall energy consumption of the network, resulting in a shorter network life cycle.

Sharma et al. quantitatively test and analyze the ETX-based and RANK-based objective function OF0 in RPL, and find that the objective function with ETX as the measure factor will be based on the minimum route during packet transmission [22]. ETX selects the routing path. Through simulation experiments, it is concluded that the ETX-based objective function is stronger than OF0 in network throughput [23], network delay, and link transmission rate.

In summary, based on the low power consumption and unstable characteristics of the wireless sensor network in the 6LoWPAN system, most of the research on RPL is to improve the network life cycle and the delivery rate of data packets. This paper will also present its own research program for this purpose.

2.3 RR_RPL Description

As shown in Fig. 2.1, in the ideal 6LoWPAN network, the remaining energy consumption level in operation should be gradually reduced according to the distance from the root node. In reality, the node closer to the root node tends to bear more forwarding tasks, so the energy will consume faster and prematurely exit the network, while the leaf nodes in the edge part may still have more remaining power after the parent node runs out of power. Therefore, we need to optimize the routing decision of the 6LoWPAN network by designing a reasonable objective function to achieve more balanced energy consumption. The ultimate goal is to extend the life cycle of the entire network.

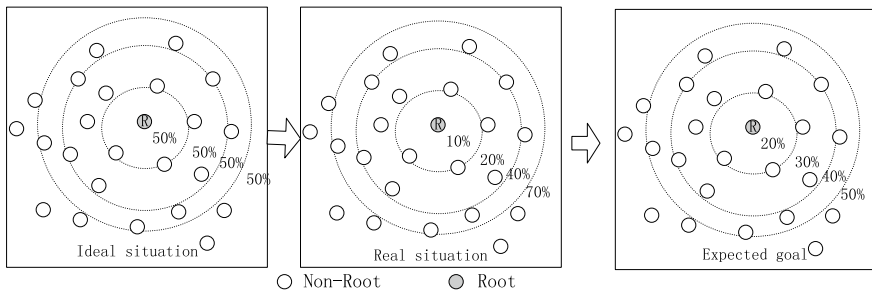


Fig. 2.1 Remaining energy of 6LoWPAN

2.3.1 RR_RPL Calculation Model

Load balancing is critical for RPL networks. In the long-term stable operation of the network, maintaining low packet loss is a necessary function. Therefore, ETX metrics must be taken into account in routing decisions. In actual testing, rank comparison A node with a lower node than a high rank value needs to forward more routing information, which means that the energy consumption is correspondingly increased. In order to prevent the intermediate node in the root node to the leaf node transmission link from depleting energy prematurely, we also need to consider the remaining energy of the parent node in the process of selecting the parent node, so that the network has a longer life cycle. Since the rank value reflects the network level at which the node is located, the rank value is also incorporated into our routing algorithm as an important link metric.

This paper proposes a new RPL routing protocol: RR_RPL (Rank_Remain Energy RPL). The protocol uses the value of the node rank as the weight, combined with the ETX and the remaining energy of the node, aiming at balancing the energy consumption of each node and ensuring the optimal link condition to achieve the purpose of extending the network life cycle. The RR_RPL calculation model is as shown in Eq. 2.5:

$$M = \frac{\text{Rank}_p}{\text{Rank}_{\max}} \left(\frac{\text{ETX}_{(n,p)}}{\text{ETX}_{\max}} \right) + \left(1 - \frac{\text{Rank}_p}{\text{Rank}_{\max}} \right) \left(1 - \frac{\text{Energy}_p}{\text{Energy}_{\max}} \right). \quad (2.5)$$

M is the routing metric when the child node selects the parent node. In the routing decision, the child node selects the parent node with the smaller M value among the candidate parent nodes. Rank_p is the rank value of the parent node, Rank_{\max} is the maximum number of hops allowed, Energy_p is the remaining energy value of the parent node, Energy_{\max} is the initial energy value, and $\text{ETX}_{(n,p)}$ is the expected number of transmissions from the child node to the candidate parent node. ETX_{\max} is the sum of the most ETXs in the DIO message sent to the alternate parent.

The Rank_p value is calculated using the Formula (3.1) in the OF0 objective function, where $\text{Rank}_{\text{Increase}}$ is set to 1 in the Formula (3.1).

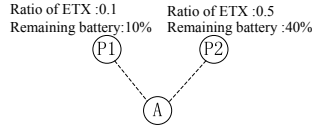


Fig. 2. Parent node selection

2.3.2 Comparison of M Metrics and R Metrics

In the R metric, since the b value is a fixed value as the weight, when the different depth nodes are faced, the weight ratio of the two metrics cannot be dynamically adjusted according to the actual situation of the node. Figure 2.2 shows an example of the decision to select a parent node.

Suppose $P1$ and $P2$ are the potential parent nodes of the A node. By calculating the parameters in Fig. 2.2, you can get:

1. Under the R metric, $RP1 = 0.5$ and $RP2 = 0.55$. As a result, the $P1$ node is selected as the parent node.
2. Under the M metric, if the parent node is at a higher rank level, it is assumed that the rank ratio is 0.2, then $RP1 = 0.74$, $RP2 = 0.58$, and the $P2$ node is selected as the parent node; if the parent node is at the lower rank level, for example, the rank ratio is 0.8, then $RP1 = 0.80$, $RP2 = 0.88$, then the $P1$ node is selected as the parent node.

Therefore, it can be seen that the M metric is a routing decision mechanism that uses different weights for nodes of different rank levels, so that the network can dynamically adjust the routing strategy according to the depth of the network where the node is located and is more flexible.

2.3.3 RR_RPL Pseudocode Implementation

The source location of the RPL protocol is in the core \rightarrow net \rightarrow rpl directory of the Contiki operating system, which contains the source code used for the objective function and topology construction. Table 2.1 is the pseudocode implementation of RR_RPL.

2.4 Simulation Results and Analysis

In this paper, Contiki's own Cooja simulator is used to simulate and analyze the ETX-based objective function, R_RPL objective function and improved RR_RPL objective function in several different nodes.

Table 2.1 Pseudo code of RR_RPL

```

Pseudo code of RR_RPL :
Define R_weight=Rankp/Rankmax /* R value as weight */
If(ParentA && ParentB != NULL)
{
    R1=BestNextHop(ParentA);
    R2=BestNextHop(ParentB);
    Return R1<R2?ParentA:ParentB;
}
else return ParentA;
BestNextHop(Node)
{
    M=R_weight(ETX(n,p) / ETXmax) + (1-R_weight)(1-Energyp / Energymax);
    Return M;
}
    
```

2.4.1 Simulation Setup

As shown in Fig. 2.3, the nodes are distributed in a square area of 250 m * 300 m. In this area, All nodes are randomly distributed. Three simulations were performed under the same number of nodes, and the results were averaged. The following is the selection and setting of the simulation environment and parameters to compare the network status before and after the improvement. Two dry batteries have a power

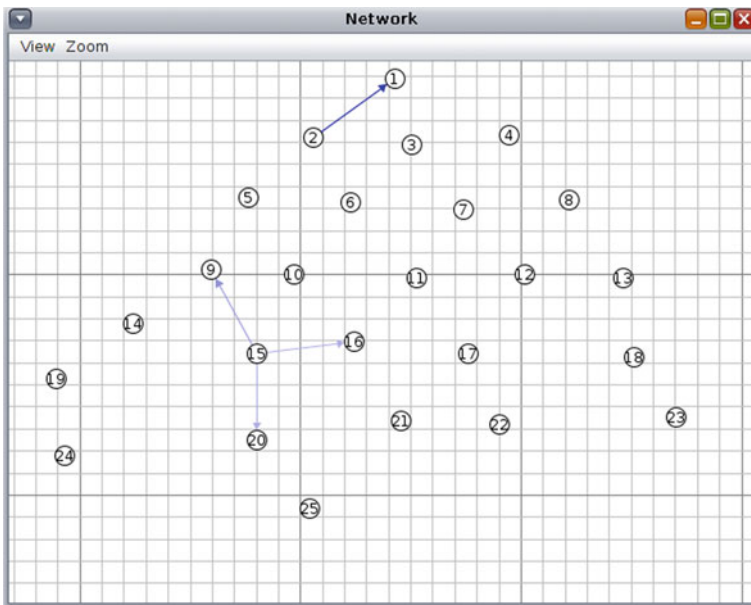


Fig. 2.3 Node distribution

Table 2.2 Simulation parameters

Variable name	Value
The operation system	Contiki-3.0
Simulation range	250 m * 300 m
Number of nodes	25, 50, 75, 100, 125, 150, 175
Transmission distance setting	50 m
Send rate	2 pkt/s
Packet size	127 bytes
MAC layer	IEEE802.15.4 MAC
Routing protocol	RPL
Initial power of the node	6 J
Simulation cycle	100 min

of about 7500 J, and the two power supply cycles are about 150 days. Complete the simulation test in time, set node 1 as the root node, and set it to the continuous power supply. The initial power of the remaining nodes is set to 6 joules, and the simulation period is 6000 s. The simulation parameters are shown in Table 2.2.

Figure 2.3 is an example of 25 nodes as an example to analyze the structure of the network topology.

2.4.2 Simulation Results

After comparing the simulations under the two objective functions, the topology generated in the initial state of the network is shown in Figs. 2.4 and 2.5.

The objective function is different. The decision of the node to select the parent node is different, and the naturally generated topology is different. However, since RR_RPL and R_RPL are combined with ETX, the remaining power, and the remaining power of each node is sufficient in the initial stage of network construction, the topology structure under these two objective functions is not much different. However, as time passes, the energy of nodes closer to the node decreases rapidly. The improved objective function will replan the topology according to different rank values. The topology of the two objective functions will be greatly different.

Figures 2.6 and 2.7 show the average energy consumption of each node in the simulation environment.

As can be seen in Figs. 2.6 and 2.7, the improved RR_RPL objective function is superior to the R_RPL objective function in terms of overall energy consumption, and intuitively, the average energy consumption of each node is relatively balanced. The simulation results of the life cycle are analyzed.

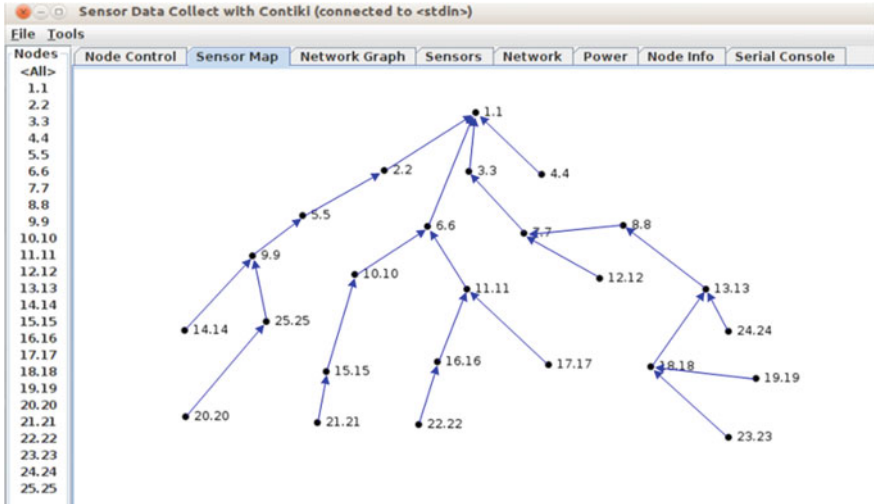


Fig. 2.4 Network topology of R_RPL

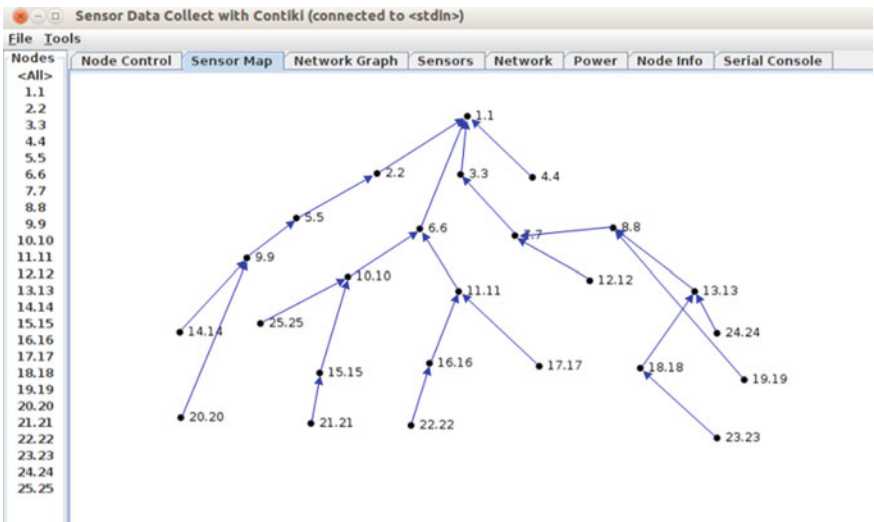


Fig. 2.5 Network topology of RR_RPL

2.4.3 Simulation Analysis

In the simulation environment, we simulated the placement of 25, 50, 75, 100, 125, 150, and 175 nodes and performed three different objective functions (RR_RPL, R_RPL, and ETX) for each layout. In contrast, the focus is on the measurement of the life cycle and throughput of the entire network.

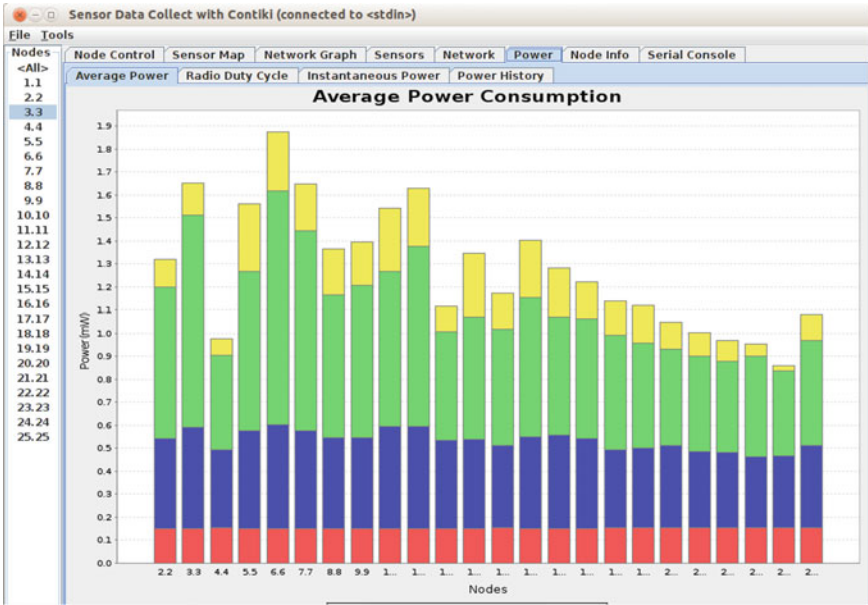


Fig. 2.6 Average energy consumption of R_RPL

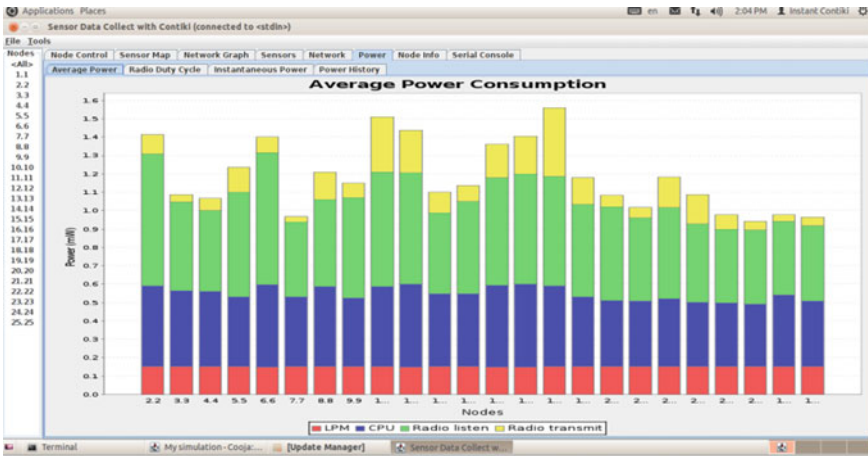


Fig. 2.7 Average energy consumption of RR_RPL

As can be seen from Fig. 2.8, as the number of nodes in the network increases, the life cycle of the entire network decreases. The objective function measured by ETX uses only a single metric of link quality as the routing decision of the objective function. Therefore, the routing decision of the node is not changed because the energy consumption of the individual node is high throughout the network cycle.

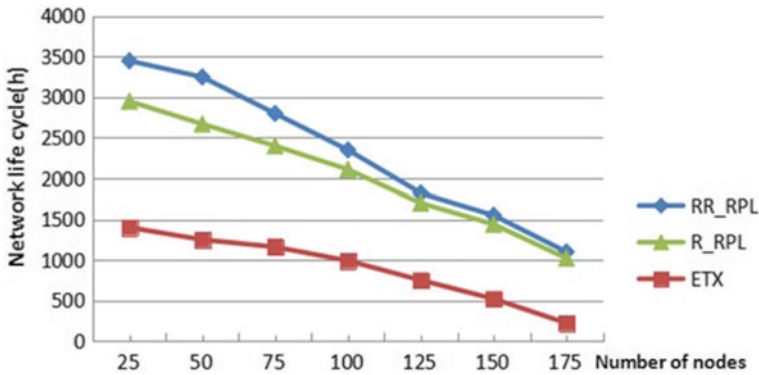


Fig. 2.8 Comparison of network lifetime

As a result, the network life cycle drops sharply as the number of nodes continues to increase. However, the R_RPL objective function does not take into account the dynamic increase or decrease of the residual energy and ETX weight of the node, which leads to the delay of updating the topology. While the RR_RPL can dynamically set the dynamic according to the three metrics during the network operation, the weights can be dynamically set according to the three metrics to select a better parent node, thus extending the network life cycle accordingly.

Figure 2.9 shows the comparison of packet delivery rates under the three objective functions. The ETX-based objective function uses link quality as the only indicator. Therefore, when the number of nodes is lower than 120, the performance is optimal. With the growth of network scale, nodes with better link quality will undertake

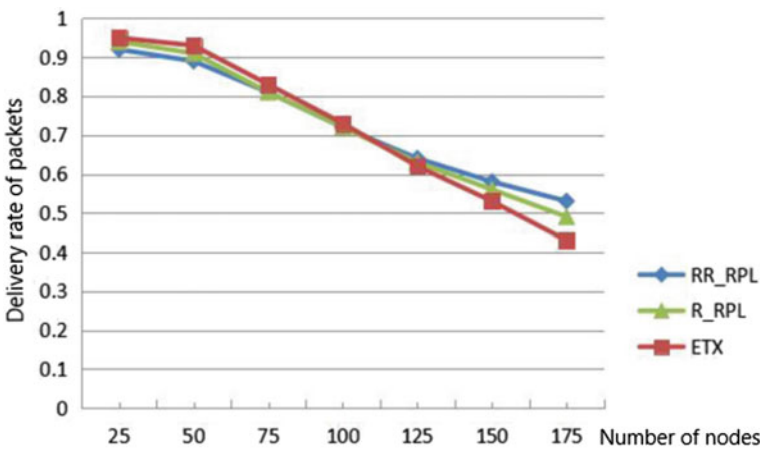


Fig. 2.9 Comparison of packet delivery rate

more packet forwarding tasks, and the resulting network congestion will increase the number of retransmissions and packet loss rate.

R_RPL and RR_RPL take into account ETX and residual energy. When the network scale is small, due to the small number of hops and insufficient network depth level, both objective functions consider the residual energy of the node more, so it is not sufficient to show the superiority of routing decisions. As the scale grows, more and more congestion begins to occur in networks based on the three objective functions, so the packet delivery rate is similar at this stage. In the case of further growth of the network scale, the nodes with smaller levels begin to consider the energy of the parent node more, while the nodes with more distant ones consider the link quality more, and the optimization effect of RR_RPL is reflected.

2.5 Conclusion and Recommendations

This paper introduces the routing process and routing strategy of RPL routing, focuses on the objective function of determining the parent node strategy, and analyzes the shortcomings of the R_RPL metric function.

Furthermore, an objective function optimization scheme RR_RPL based on load balancing and multi-measure factor is proposed, and the specific objective function design is given. Enable the node to select the best parent node according to the network level at which it is located.

After simulation test analysis under different network scales, RR_RPL shows flexible adaptive ability. Compared with R_RPL, the improved routing strategy has optimization effects on life cycle and packet delivery rate.

References

1. Iova, O., Picco, G.P., Istomin, T., Kiraly, C.: RPL, the routing standard for the Internet of Things . . . Or is it? *IEEE Commun. Mag. Inst. Electr. Electron. Eng.* **54**(12), 16–22 (2016). <https://doi.org/10.1109/mcom.2016.1600397CM.hal-0164715>
2. Brandt, A., Baccelli, E., Cragie, R., et al.: Applicability statement: the use of the routing protocol for low-power and lossy networks(RPL) protocol suite in home automation and building control. *Organic Biomol. Chem.* **9**(7), 2274–2278 (2016)
3. RFC 6552[S]: Objective function zero for the routing protocol for low-power and lossy networks (RPL). Internet Engineering Task Force: P. Thubert, Mar 2012
4. Umamaheswari, S., Negi, A.: Internet of Things and RPL routing protocol: a study and evaluation. In: 2017 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, pp. 1–7 (2017)
5. RFC 6719[S]: The minimum rank with hysteresis objective function. Internet Engineering Task Force: O. Gnawali and P. Levis, Sept 2012
6. Semedo, F., Moradpoor, N., Rafiq, M.: Vulnerability assessment of objective function of RPL protocol for Internet of Things. In: Proceedings of the 11th International Conference on Security of Information and Networks (2018)
7. Gaddour, O., Koubaa, A.: RPL in a nutshell: a survey. *Comput. Netw.* **56**(14), 3163–3178

8. Michel, M., DuQuennoy, S., Quoitin, B., et al.: Load-balanced data collection through opportunistic routing. In: Proceedings of the 2015 International Conference on Distributed Computing in Sensor Systems, pp. 62–70. IEEE, Piscataway (2015)
9. Capone, S., Accettura, N.: An energy efficient and reliable composite metric for RPL organized networks. In: 12th IEEE International Conference on Embedded and Ubiquitous Computing, pp. 178–184. IEEE Press, Milano (2014)
10. Iova, O., Theoleyre, F., Noel, T., et al.: Improving the network lifetime with energy-balancing routing : application to RPL. In: 7th IFIP Wireless and Mobile Networking Conference, pp. 1–8. IEEE Press, Vilamoura (2014)
11. Chang, L.H., Lee, T.H., Chen, S.J., et al.: Energy-efficient oriented routing algorithm in wireless sensor networks. In: IEEE International Conference on Systems, Man, and Cybernetics, Manchester, pp. 3813–3818 (2013)
12. Iova, O., Theoleyre, F., Noel, T.: Exploiting multiple parents in RPL to improve both the network lifetime and its stability. In: IEEE International Conference on Communications (ICC), London, pp. 610–616 (2015)
13. Gaddour, O., Koubaa, A., Rangarajan, R., et al.: Co-RPL: RPL routing for mobile low power wireless sensor networks using Corona mechanism. In: 2014 9th IEEE International Symposium on industrial embedded systems (SIES 2014). IEEE, pp. 200–209 (2014)
14. Rukpakavong, W., Phillips, I., Guan, L., et al.: RPL router discovery for supporting energy-efficient transmission in single-hop 6LoWPAN. In: IEEE International Conference on Communications, pp. 5721–5725. IEEE Press, Ottawa (2012)
15. Iova, O., Theoleyre, F., Noel, T.: Using multiparent routing in RPL to increase the stability and the lifetime of the network. *Ad Hoc Netw.* **29**, 45–62 (2015)
16. Chang, L.H., Lee, T.H., Chen, S.J., et al.: Energy-efficient oriented routing algorithm in wireless sensor networks. In: IEEE International Conference on Systems, Man, and Cybernetics, Manchester, pp. 3813–3818 (2013)
17. Lassouaoui, S., Rovedakis, S., Sailhan, F., Wei, A.: Evaluation of energy aware routing metrics for RPL[C]. In: IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), New York, NY, pp. 1–8 (2016)
18. Lassouaoui, L., Rovedakis, S., Sailhan, F., Wei, A.: Comparison of RPL routing metrics on grids. In: Zhou, Y., Kunz, T. (eds.) *Ad Hoc Networks*. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 184. Springer, Cham (2017)
19. Hakeem, S.A.A., Barakat, T.M., Seoud, R.A.A.: New real evaluation study of Rpl routing protocol based on cortex M3 nodes of Iot-lab test bed. *Middle-East J. Sci. Res.* **23**(8): 1639–1651 (2015)
20. Gaddour, O., Koubaa, A.: RPL in a nutshell: a survey. *Comput. Netw.* **56**(14), 3163–3178 (2012)
21. Iova, O., Theoleyre, F., Noel, T.: Stability and efficiency of RPL under realistic conditions in wireless sensor networks. In: IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). IEEE, pp. 2098–2102 (2013)
22. Sharma, R., Jayavignesh, T.: Quantitative analysis and evaluation of RPL with various objective function for 6LoWPAN. *Indian J. Sci. Technol.* **8**(19) (2015)
23. Janani, P., Diniesh, V.C., Auxilius Jude, M.J.: Impact of path metrics on RPL's performance in low power and lossy networks. In: 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, pp. 0835–0839 (2018)

Chapter 3

Design of a New Multi-output Constant Current Source Based on Power Allocation Control Strategy



Hongli Cheng and Lei Wang

Abstract In order to reduce the cross-regulation rate of multi-output constant current source and further improve the stability and accuracy of output current, a new design method based on power distribution control strategy is proposed. The system adopts the flyback converter as the main topology of the multi-output constant current source. The ARM system samples the real-time load value of the multi-output and combines the target current to determine the total power of the multi-output at this time. The total output power is the total input power. Ideally, combining the energy transfer of each output switch state of the multi-output flyback converter, the conduction time of the main switch transistor and the secondary side switch transistor of the flyback converter can be calculated, so as to distribute the power rationally on the transformer, thereby achieving reduced cross-regulation and improved current accuracy and stability. The experimental results show that a new multi-output constant current source system based on power distribution control has lower cross-regulation rate and the output current accuracy is within $\pm 2\%$ mA, which improves the accuracy and stability of the output current.

3.1 Introduction

At present, how to reduce the cross-regulation rate of multi-output constant current sources and improve the stability and accuracy of multi-output currents is becoming hot topic. In this regard, a lot of research and experiments have been carried out. The main reason that affects the stability and accuracy of multi-output constant current source is the existence of cross-regulation rate. The multi-output flyback converter will only use feedback regulation for the main output, while the auxiliary output will use open-loop non-feedback [1]. The reason for the deterioration of the cross-regulation rate is the non-ideality of the transformer (leakage inductance and diode drop) [2]. The literature [3] proposes a new TDK-Lambda solution to

H. Cheng · L. Wang (✉)

College of Communication and Information Engineering, Xi'an University of Science and Technology, Xi'an 710000, China
e-mail: 13149169707@163.com

improve the multi-output cross-regulation rate. The literature [4–8] proposes the use of weighted voltage feedback control method, passive lossless box circuit and optimized transformer, magnetic amplification, etc. These methods are used to improve the cross-regulation rate of the converter. Although these methods reduce the cross-regulation rate, they do not reduce the total amount of error, and the algorithm and circuit are complicated. In [9], a design method based on power distribution control strategy is proposed. This method fundamentally solves the problem of multi-output cross-regulation rate caused by the existence of leakage inductance and improves the accuracy and stability of the output voltage. However, this design method is not suitable for high-precision multi-channel constant current sources. On the one hand, the previous control strategy is multi-channel constant voltage output instead of multi-channel constant current output. On the other hand, two methods are used in the literature to prove that the multi-channel design method lacks generality, and the accuracy of multi-channel output is low. The main reason for low accuracy is that the on-time calculation of the main switch and the output switch is too simplified in the power distribution control strategy firstly. Secondly, the mutual inductance between the primary and secondary inductors of the flyback converter after the main switch transistor is turned off is not considered, so it is difficult to achieve the accuracy required by the design.

Based on the original power allocation control strategy, this paper proposes a new power distribution control-based flyback converter secondary side switch transistor on-time calculation method to reduce cross-adjustment rate for multi-channel constant current source. The control strategy further improves the current accuracy and the stability of the multi-output constant current source.

3.2 System Composition and Control Strategy

3.2.1 System Composition

The main structure of the circuit of the multi-output constant current source is shown in Fig. 3.1. If the output voltage and output current value of each channel are measured, the real-time load value of each channel and the rated output total power required for each channel can be determined. The total rated output power is equal to the input power, and the input end of the transformer determines the input power by controlling the conduction time of the primary synchronous rectifier. And the power on the transformer is distributed by controlling the conduction time of the secondary switch transistor. The system adopts multi-output flyback converter as the main topology. The primary side of the flyback converter includes the input DC power supply $U_1(t)$ and the synchronous rectifier S that controls the power level. The secondary side of the flyback converter includes three outputs corresponding synchronous rectifiers S_1 , S_2 , and S_3 , filter capacitors C_1 , C_2 , and C_3 , three output loads R_{L1} , R_{L2} , and R_{L3} . The main control part of the system uses ARM as the control

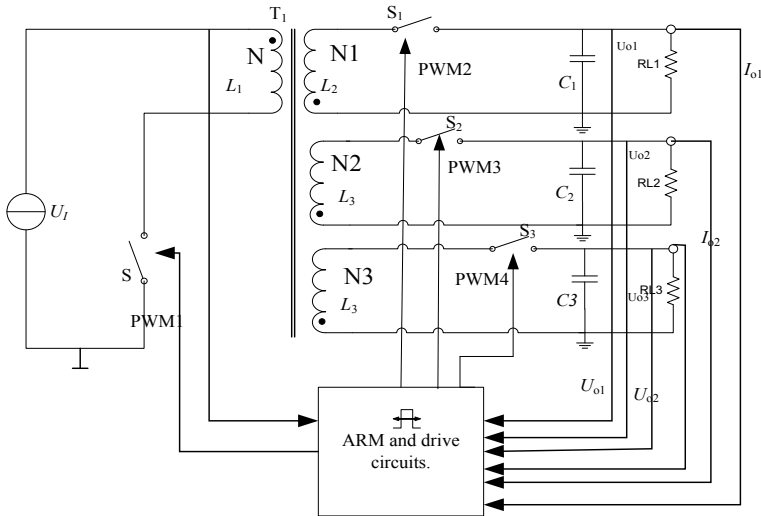


Fig. 3.1 Main circuit structure diagram of the system

core to process the collected data. The signal output by the ARM is controlled by the driving circuit to control the on and off states of several synchronous rectifiers.

ARM samples the input voltage $U_1(t)$, the output voltages $U_{o1}(t)$, $U_{o2}(t)$, and $U_{o3}(t)$, and the load currents $I_{o1}(t)$, $I_{o2}(t)$, and $I_{o3}(t)$ in real time to obtain the real-time loads R_{L1} , R_{L2} , and R_{L3} . ARM calculates the rated power of each output combined with the rated current and derives the on-time of each switch according to the new design method based on the power distribution control strategy. The ratio of the on-time to the period is four PWM waves. The requirements of the three outputs for the rated power can be satisfied by controlling the duty cycle of the four PWM waves.

3.2.2 System Control Strategy

Equation (3.1) is the real-time load value expression sampled by the ARM system. In the formula, R_{Li} is the i th real-time load value of the flyback converter output, V_{oi} is the i th output voltage value of the flyback converter, and I_{oi} is the real-time current value of the i th of the flyback converter. The three-way output real-time loads R_{L1} , R_{L2} , and R_{L3} can be calculated using Eq. (3.1).

$$R_{Li} = \frac{v_{oi}}{i_{oi}} \tag{3.1}$$

Equation (3.2) is the output rated power expression of each load of the flyback converter. In Formula (3.2), P_{oi} is the rated output power of the i th path of the flyback converter, and I_{Ei} is the target current of the i th path of the flyback converter. The real-time load obtained by Formula (3.1) can obtain the rated output power of the i th channel, and the three-way rated output power P_{o1} , P_{o2} , P_{o3} can be calculated by Formula (3.2). At this time, the total output rated power of the flyback converter is the sum of the rated output power of each output.

$$P_{oi}(t) = I_{Ei}^2 R_{Li} \quad (3.2)$$

Ideally, the total output rated power of the flyback converter is the total rated power of the input. However, the primary and secondary energy of the flyback converter cannot be completely transferred due to the leakage inductance and mutual inductance. Assume that the conversion efficiency of the high-frequency transformer, the rectifier switch, and the snubber circuit is η , and the real-time input power of the transformer is Eq. (3.3).

$$p_i(t) = \eta P_O(t) \quad (3.3)$$

The flyback converter designed by this system works in discontinuous conduction mode (DCM) [10]. According to the working principle of DCM, the input power obtained by the flyback converter is calculated as Eq. (3.4) in one switching cycle.

$$p_i(t) = \frac{U_1^2(t)t^2}{2LT} \quad (3.4)$$

In Eq. (3.4), L is the primary inductance value of the flyback converter, $U_1(t)$ is the input real-time voltage, T is the main switching period, and t is the conduction time of the main switch transistor in one switching period. According to Eqs. (3.3) and (3.4), the on-time t of the main switch S in one switching period can be obtained as Eq. (3.5).

$$t = \sqrt{\frac{P_O(t) \cdot 2LT}{\eta U_1^2(t)}} \quad (3.5)$$

Let $n_1 = N_1/N$, $n_2 = N_2/N$, $n_3 = N_3/N$, $V_{E1} = I_{E1}R_{L1}$, $V_{E2} = I_{E2}R_{L2}$, $V_{E3} = I_{E3}R_{L3}$, N , N_1 , N_2 , N_3 are the primary turns and secondary of the flyback converter, respectively. The number of turns n_1 , n_2 , and n_3 is the primary and secondary ratios of the flyback converter, respectively. I_{INMAX} is the peak value of the primary inductor current. I_{1MAX} , I_{2MAX} , and I_{3MAX} are the inductor current peaks of the first-stage secondary side, respectively. According to the above and multi-output output flyback converter ampere-conservation principle, Formulas (3.6), (3.7), and (3.8) can be obtained early.

$$I_{INMAX} = n_1 \cdot I_{1MAX} + n_2 \cdot I_{2MAX} + n_3 \cdot I_{3MAX} \quad (3.6)$$

$$\frac{I_{1MAX} \cdot I_{E1} \cdot R_{L2}}{I_{2MAX} \cdot I_{E2} \cdot R_{L2}} = \frac{N_1}{N_2} = \frac{n_1}{n_2} \quad (3.7)$$

$$\frac{I_{1MAX} \cdot I_{E1} \cdot R_{L2}}{I_{3MAX} \cdot I_{E3} \cdot R_{L3}} = \frac{N_1}{N_3} = \frac{n_1}{n_3} \quad (3.8)$$

The three peak currents of the secondary side inductance of the flyback converter obtained by Eqs. (3.6), (3.7), and (3.8) are expressed by Eqs. (3.9), (3.10), and (3.11).

$$I_{1MAX} = \frac{n_1 \cdot V_{E2} \cdot V_{E3} \cdot I_{INMAX}}{n_1^2 V_{E2} V_{E3} + n_2^2 V_{E1} V_{E3} + n_3^2 V_{E1} V_{E2}} \quad (3.9)$$

$$I_{2MAX} = \frac{n_2 \cdot V_{E1} \cdot V_{E3} \cdot I_{INMAX}}{n_1^2 V_{E2} V_{E3} + n_2^2 V_{E1} V_{E3} + n_3^2 V_{E1} V_{E2}} \quad (3.10)$$

$$I_{3MAX} = \frac{n_3 \cdot V_{E1} \cdot V_{E2} \cdot I_{INMAX}}{n_1^2 V_{E2} V_{E3} + n_2^2 V_{E1} V_{E2} + n_3^2 V_{E1} V_{E2}} \quad (3.11)$$

After the main switch transistor is turned off, the secondary side three channels simultaneously release current. When the system is operating in the first stage, the inductor current on the three secondary windings is shown in Fig. 3.2.

Assume that the three inductor currents reach their respective rated output current requirements after the times t_{s1} , t_{s2} , and t_{s3} , respectively, and the times at t_{s1} , t_{s2} , and t_{s3} can be found. According to the inductor current waveform in the first stage of Fig. 3.2. The corresponding inductor current values I_1' , I_2' , and I_3' are shown in Eq. (3.12).

$$\begin{cases} I_1' = I_{1MAX} - \frac{V_{E1}}{L_1} t_{s1} \\ I_2' = I_{2MAX} - \frac{V_{E2}}{L_2} t_{s2} \\ I_3' = I_{3MAX} - \frac{V_{E3}}{L_3} t_{s3} \end{cases} \quad (3.12)$$

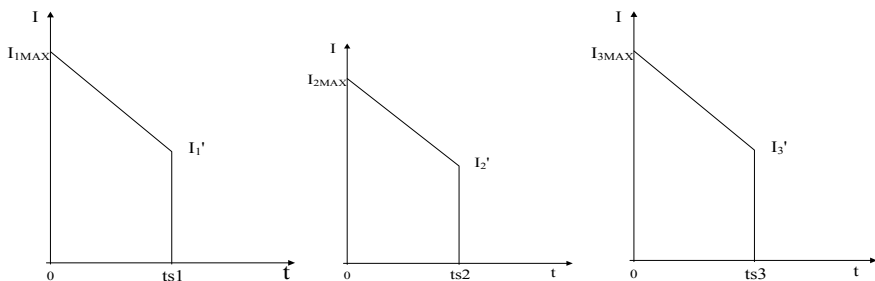


Fig. 3.2 Second-stage secondary inductance current diagram of the system

At this time, the average output current of the three channels is the rated current. The output rated current can be expressed as Eq. (3.13).

$$I_{E1} = \frac{1}{2T} (I_{1\text{MAX}} + I'_1) \cdot t_{s1} \quad (3.13)$$

$$I_{E2} = \frac{1}{2T} (I_{2\text{MAX}} + I'_2) \cdot t_{s2} \quad (3.14)$$

$$I_{E3} = \frac{1}{2T} (I_{3\text{MAX}} + I'_3) \cdot t_{s3} \quad (3.15)$$

Bringing Formula (3.12) into Eqs. (3.13), (3.14), and (3.15) can solve Eqs. (3.16), (3.17), and (3.18) when the respective outputs reach the respective rated current requirements in the first stage.

$$t_{s1} = \frac{2I_{1\text{MAX}} \cdot L_1 \pm \sqrt{4I_{1\text{MAX}}^2 L_1^2 - 8L_1 I_{E1} V_{E1} T}}{2V_{E1}} \quad (3.16)$$

$$t_{s2} = \frac{2I_{2\text{MAX}} \cdot L_2 \pm \sqrt{4I_{2\text{MAX}}^2 L_2^2 - 8L_2 I_{E2} V_{E2} T}}{2V_{E2}} \quad (3.17)$$

$$t_{s3} = \frac{2I_{3\text{MAX}} \cdot L_3 \pm \sqrt{4I_{3\text{MAX}}^2 L_3^2 - 8L_3 I_{E3} V_{E3} T}}{2V_{E3}} \quad (3.18)$$

Equations (3.16), (3.17), and (3.18) are the times when the outputs of the respective circuits reach their respective rated currents, and the on-times t_{s1} , t_{s2} , and t_{s3} of the first-stage switches are solved, and three times and minimum values are compared. It is the first turn-on time of the switch, and the remaining values are discarded. It is assumed that the on-time of the switch S_1 is the shortest at this time, and the minimum time $t = t_{s1}$ is the on-time of the corresponding switch S_1 , thereby obtaining the on-time of the switch S_1 .

The second stage is the switch with the shortest turn-off, and the other switches continue to conduct their rated current requirements. After the switch S_1 is turned off, the switches S_2 and S_3 are continuously turned on, and the other two inductor currents jump when the switch S_1 is turned off. Figure 3.3 shows the change in secondary inductor current in the second stage.

According to the change of the second stage inductor current in Fig. 3.3, the peak values of the inductor currents of the remaining paths of the first switch can be obtained as shown in Eqs. (3.19), (3.20), and (3.21).

$$I'_1 = I_{1\text{MAX}} - \frac{V_{E1}}{L_1} \cdot t_{s1} \quad (3.19)$$

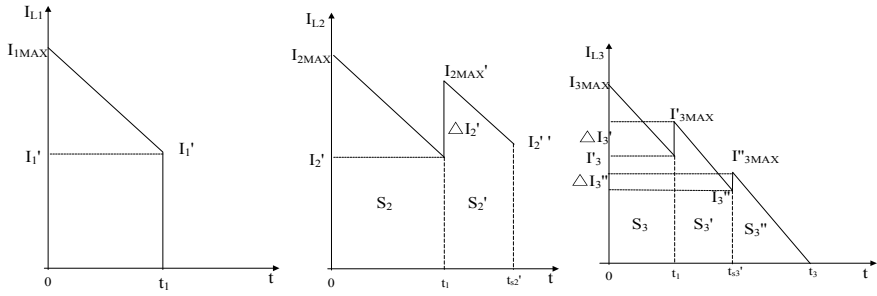


Fig. 3.3 System second-stage secondary inductor current diagram

$$I_2' = I_{2MAX} - \frac{V_{E2}}{L_2} \cdot t_{s1} \quad (3.20)$$

$$I_3' = I_{3MAX} - \frac{V_{E3}}{L_3} \cdot t_{s1} \quad (3.21)$$

$$I_{2MAX}' = I_{2MAX} - \frac{V_{E2}}{L_2} \cdot t_{s1} + \Delta I_2' \quad (3.22)$$

$$I_{3MAX}' = I_{3MAX} - \frac{V_{E3}}{L_3} \cdot t_{s1} + \Delta I_3' \quad (3.23)$$

In Eqs. (3.22) and (3.23), $\Delta I_2'$ and $\Delta I_3'$ are transitions of the residual energy of the inductor on the inductors L_2 and L_3 .

Let $n_{21} = N_2/N_1$, $n_{31} = N_3/N_1$, and refer to Formula (3.6) to obtain:

$$I_1' = n_{21} \cdot \Delta I_2' + n_{31} \cdot \Delta I_3' \quad (3.24)$$

$$\frac{\Delta I_2' \cdot V_{E2}}{\Delta I_3' \cdot V_{E3}} = \frac{n_{21}}{n_{31}} \quad (3.25)$$

According to Eqs. (3.24) and (3.25), the jump values of the remaining two inductor currents are obtained as Eqs. (3.26) and (3.27).

$$\Delta I_2' = \frac{n_{21} \cdot V_{E3}}{n_{21}^2 V_{E3} + n_{31}^2 V_{E2}} \cdot I_1' \quad (3.26)$$

$$\Delta I_3' = \frac{n_{31} \cdot V_{E2}}{n_{21}^2 V_{E3} + n_{31}^2 V_{E2}} \cdot I_1' \quad (3.27)$$

Assume that the other two outputs reach their rated current values after the t_{s2}' , t_{s3}' time, and the expressions of the remaining two output rated currents according to Fig. 3.2 are (3.28) and (3.29).

$$I_{E2} = \frac{S_2 + [1/2(I'_{2\text{MAX}} + I'_{2\text{MAX}} - V_{E2}/L_2 \cdot (t'_{s2} - t_1)) \cdot (t'_{s2} - t_1)]}{T} \quad (3.28)$$

$$I_{E3} = \frac{S_3 + [1/2(I'_{3\text{MAX}} + I'_{3\text{MAX}} - V_{E3}/L_3 \cdot (t'_{s3} - t_1)) \cdot (t'_{s3} - t_1)]}{T} \quad (3.29)$$

According to Eqs. (3.28) and (3.29), the on-time of the switch when the remaining two paths of the second stage reach the output rated current requirement can be solved, and S_2 and S_3 are the total output current of the two channels when the first stage. Suppose $t_2 = \min\{t_{s2}', t_{s3}'\}$ is the on-time of switch S_2 , and find the value of the inductor current of the remaining two channels at time t_2 and the total output current of the second stage as Eqs. (3.30), (3.31), (3.32).

$$I''_2 = I'_{2\text{MAX}} - \frac{V_{E2}}{L_2} (t'_{s2} - t_1) \quad (3.30)$$

$$I''_3 = I'_{3\text{MAX}} - \frac{V_{E3}}{L_3} (t'_{s2} - t_1) \quad (3.31)$$

$$S'_3 = \frac{1}{2} (I'_{3\text{MAX}} + I''_3) \cdot (t'_{s2} - t_1) \quad (3.32)$$

The remaining inductor current jumps, and the system enters the third stage when the switch S_2 is turned off.

$$\Delta I'_3 = \frac{N_2}{N_3} \cdot I''_2 \quad (3.33)$$

$$I''_{3\text{MAX}} = I''_3 + \Delta I'_3 \quad (3.34)$$

$$S''_3 = \frac{1}{2} \cdot I'_{3\text{MAX}} \cdot (t'_{s3} - t_2) \quad (3.35)$$

$$I_{E3} = \frac{S_3 + S'_3 + S''_3}{T} \quad (3.36)$$

Equation (3.33) is the jump value of the inductor current at this time. Equation (3.34) is the peak value of the inductor current at time t_2 , and Eq. (3.35) is the total output current of the switch S_3 . The conduction time t'_3 of the third switch can be obtained in Eq. (3.37) by Eqs. (3.33), (3.34), (3.35), and (3.36) when the switch S_2 is turned off.

$$t'_3 = \frac{2 \cdot (I_{E3} \cdot T - S_3 - S'_3)}{I''_{3\text{MAX}}} \quad (3.37)$$

At this time, t_3 is the on-time of the switch S_3 . At this time, if $t_3 > t_{\text{off}}$, t_{off} is the main switch off time, the switch S_3 is turned on until the next cycle comes that means

the third time is as shown in Eq. (3.38).

$$t_3 = T - t \quad (3.38)$$

So far, a new three-way switch on-time of the multi-output constant current source system based on the power distribution control strategy design method is derived. The design method is based on the power allocation control strategy, and the way of calculation is accurate on the original basis. This method avoided the error caused by the inaccurate on-time of the switch conduction time.

3.3 System Programming

The ARM Cortex-M3 processor-based STM32 microcontroller is widely used for its high performance, high compatibility, easy development, and low power consumption. This design selects STM32F103RCT6 as the main control chip. It integrates a wealth of peripheral functions such as ADC, DMA, TIM, and GPIO. After frequency doubling, the system clock is 72 MHz, and two 12-bit ADC analog-to-digital conversion modules are integrated internally, which have up to 16 external signal sampling channels and can realize multi-channel data sampling. The input voltage is sampled in real time by ADC, the duty cycle of driving waveform is calculated by outputting three voltage and current values, and the timer compares and outputs four PWM signals. STM32F103RCT6 is a commonly used microcontroller of ST Company, which is suitable for: power electronic control, PWM motor drive, industrial application control, medical system, handheld equipment, PC game peripherals, GPS platform, PLC, frequency converter control, scanner, printer, alarm system, video intercom, air-conditioning heating system, and so on. This topic belongs to the application of power electronic control system, and the processing speed and function are much faster than the 51 and AVR used before. Figure 3.4 is the block diagram of the system. Firstly, the system configures the system clock and initializes each peripheral, and then the ARM samples and calculates the real-time load value after the program is initialized. Using the above derivation, the real-time output power and the on-time of each output switch transistor are calculated. The ratio of the on-time of the switch transistor to the period T is the duty ratio of each channel, and four PWM waves are output. Because the actual circuit parameters are not ideal, the primary and secondary leakage inductances of the transformer exist. The calculated secondary control PWM wave duty cycle needs to be adjusted by the program to stabilize the three output currents. The det in the flowchart is 0.01. When the flyback converter is operating, the duty cycle of the main switch is less than 0.5. Duty cycle is taken less than 0.45 to ensure that the flyback converter operates in DCM in this design.

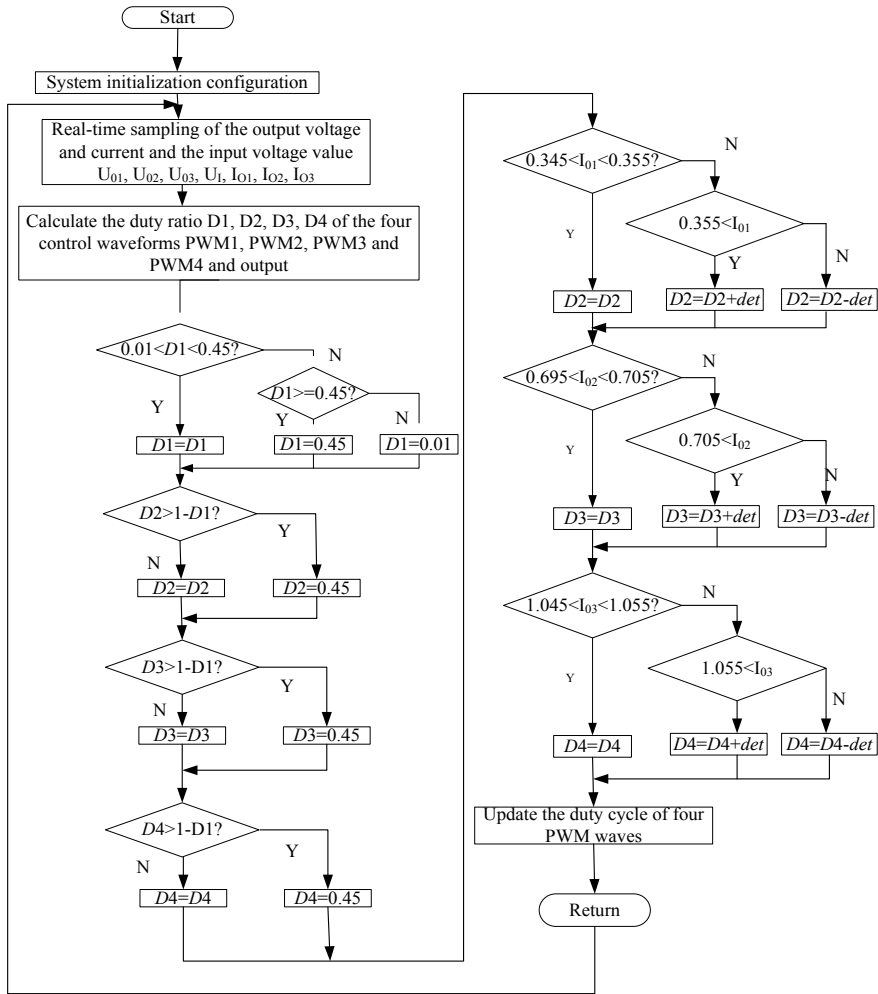


Fig. 3.4 System program flowchart

3.4 System Simulation Results Analysis

Saber simulation software is an EDA software from Synopsys of the USA. It can be used in hybrid system simulation of different types of systems such as electronics, power electronics, mechatronics, mechanical, optoelectronics, optics, and control. It provides complex mixed-signal design and verification. A powerful mixed-signal simulator is compatible with analog, digital, and control-mode hybrid simulations to solve a range of problems from system development to detailed design verification. Its main applications are power converter design, servo system design, circuit simulation, power supply and distribution design, bus simulation, and many other

fields. It has powerful and obvious advantages. The simulation of this system mainly adopts Saber.

A three-output flyback converter is designed according to the above derivation ideas and simulated by Saber.

(1) Simulation requirements

Input voltage: $U_1(t) = 60 \text{ V}$

Output power: $P_{\text{omax}} = 48 \text{ W}$

Output rated current: $I_{E1} = 0.35\text{A}; I_{E2} = 0.7\text{A}; I_{E3} = 1.05\text{A}$

Switching frequency: $f = 20 \text{ kHz}$

Maximum duty cycle: $D_{\text{max}} = 0.45$

(2) Parameter calculation

The simulation test input DC voltage $U_1(t) = 60 \text{ V}$ and the period is $50 \mu\text{s}$. The calculated primary inductance value of the flyback converter is $L_p = 250 \mu\text{H}$, the secondary inductance values $L_1 = 85 \mu\text{H}, L_2 = 72 \mu\text{H}$, and $L_3 = 50 \mu\text{H}$ from the literature [11–13]. The value of the output current is measured according to the load parameters listed in Table 3.1. The experimental test results are shown in Table 3.1.

Figure 3.5 is a waveform diagram of the three-way inductor current of the transformer secondary side obtained by saber simulation in the whole cycle, and Fig. 3.6 is a simulation output three-way current diagram. It can be seen that the output three-way inductor current is simultaneously released when the main switch transistor is turned off, and the other two inductor currents jump when the first switch is turned off, and the third inductor current jumps when the second switch transistor is turned off, which is consistent with the theoretical derivation. The slope of the second and third inductor currents that jump after the transition is ideally the same as the slope of the previous stage.

The Saber software is used to simulate the control method of the system. The duty cycle of the four-way PWM is calculated according to the design goal, and then the measured output current is shown in Table 3.1. According to the table calculation, when the load changes in any way, the other road loads are unchanged. The load adjustment rate of each channel does not exceed 2%, and the cross-regulation rate does not exceed 2%, which improves the cross-regulation rate. According to Table 3.1, it can be seen that when the three rated currents are 0.35, 0.7, and 1.05 A,

Table 3.1 Simulation measured database on modified power distribution control

U_1	R_{L1}	R_{L2}	R_{L3}	$T \text{ (}\mu\text{s)}$	$t_1/T \text{ (}\%)$	$t_2/T \text{ (}\%)$	$t_3/T \text{ (}\%)$	$I_{O1} \text{ (A)}$	$I_{O2} \text{ (A)}$	$I_{O3} \text{ (A)}$
60	44	34	22	21.13	9.8	44	57.7	0.349	0.701	1.046
60	35	34	22	21.24	10.6	42.5	57.5	0.352	0.703	1.05
60	44	30	22	20.68	10.6	38.6	61	0.346	0.699	1.052
60	44	20	22	20.45	11.1	35.7	64.3	0.348	0.701	1.051
60	38	34	21	20.7	9.1	44.2	56.8	0.349	0.698	1.048
60	38	34	15	20.45	9.4	41.8	58.2	0.351	0.702	1.045

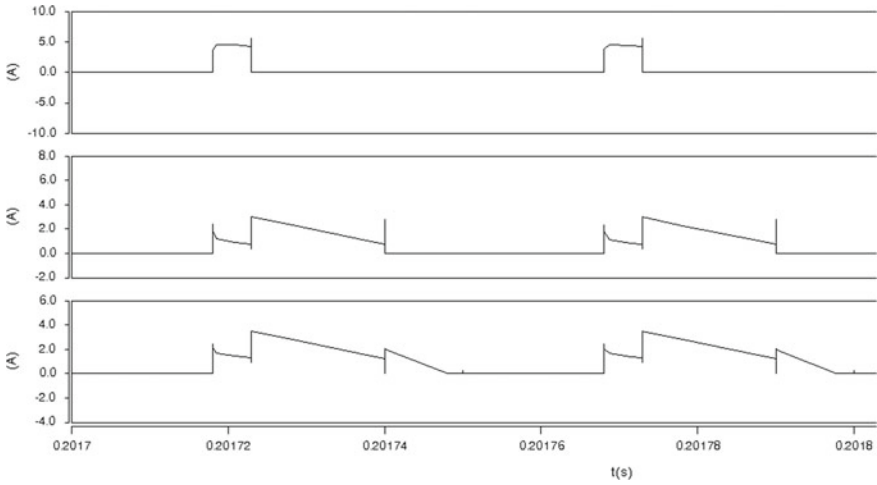


Fig. 3.5 Secondary side three-way inductor current waveform

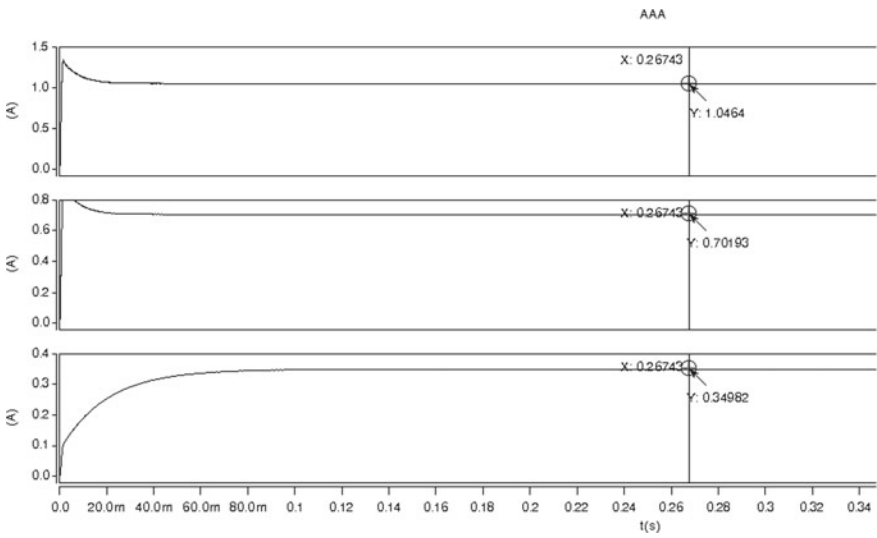


Fig. 3.6 Output current waveform measured by simulation

and the measured final output current accuracy is within $\pm 2\%$ mA, which effectively improves the accuracy of the output current. The experimental results showed that the improved method also improves the stability of multiple outputs.

3.5 Conclusion

In this paper, a new multi-output constant current source based on power distribution control strategy is designed. The system starts from the perspective of power distribution control and proposes a new method to solve the problem of cross-adjustment rate fundamentally. The design structure is simple and easy, high reliability, especially in low-power applications and low-cost multi-output power supply applications. It can meet the requirements of general multi-output power supply cross-adjustment rate. It also improved the accuracy and stability of multi-output constant current source at the same time. The overall power control is actually digital, giving the appropriate input power according to the rated power on the load and giving the appropriate power distribution according to each desired current. This paper makes the output current more accurate on the original basis by rigorous theoretical derivation. At the end of the paper, the performance parameters of the power supply are simulated and verified. The simulation results showed that the new power-based control strategy based on the secondary side conduction time is more reliable and reduced the cross-regulation rate of multiple outputs that improved the stability and accuracy of multiple output currents.

References

1. Mao, X., Chen, W.: Improving the cross-regulation rate of flyback converter by designing transformer. *Low Volt. Appar.* **2007**(23), 8–12 + 55
2. Ji, C., Smith, M., Smedley, K.M., King, K.: Cross regulation in flyback converters: analytic model and solution. *IEEE Trans. Power Electron.* **16**(2), 231–239 (2001)
3. Wang, H., Wang, Y.: A new type of solution to improve the cross-regulation rate of multi-output power supply. *Power World* **2015**(01): 25–27 + 24
4. Chen, Q., Lee, F.C., Jovanovic, M.M.: Analysis and design of weighted voltage-mode control for a multiple-output converter. *IEEE*, 449455 (1993)
5. Ji, C., Smith, M.J., Smedley, K.M.: Cross regulation in flyback converters: analytic model and solution. *IEEE Trans. Power Electron.* **16**(2), 231–239 (2001)
6. Chalermyanont, K., Sangampai, P., Prasertsit, A., Theinmontri, S.: High Frequency Transformer Designs for Improving Cross-Regulation in Multi-Output Flyback Converters. *IEEE PEDS*, pp. 53–56 (2007)
7. Jovanovic, M.M., Huber, L.: Small-signal modeling of nonideal magamp PWM switch. *IEEE Trans. Power Electron.* **14**(5), 882–889 (1999)
8. Wang, R., Wang, J., Zhang, J.: Multi-channel independent controllable constant current output LED flyback drive power supply based on post-stage adjustment. *Mech. Elect. Eng.* **33**(01), 101–105 (2016)
9. Wu, C.H., Guo, Y.J.: Multi-output flyback converter based on power distribution control. *Electron. Dev.* **40**(02), 471–475 (2017)
10. Ren, X., Ren, S., Fan, Q.: Influence of excitation current characteristics on sensor sensitivity of permeability testing technology based on constant current source. In: 2018 7th International Conference on Energy, Environment and Sustainable Development (ICEESD 2018) 2018
11. Zhao, T., Liu, H., Huang, J., et al.: Modeling and design of current mode DCM flyback converter circuit. *Power Supply Technol.* 2014(11), 2122–2124

12. Liu H, Yuan H, Shi Z, Sui S, Wang K.: Design of high precision digital AC constant current source. In: 2017 3rd International Forum on Energy, Environment Science and Materials (IFEESM 2017) (2018)
13. Zhang, Z., Cai, X.: Principle and Design of Switching Power Supply (Revised Edition). Publishing House of Electronics Industry, Beijing (2007)

Chapter 4

Use Accelerometer to Monitoring the Exercise Status: Held the Push-Ups Movement as an Example



Yi-Yang Chen, Li-Wa Sha, and Wen-Hsin Chiu

Abstract The present study employed the common push-up exercise to evaluate the feasibility of using accelerometers to monitor exercise status. A total of 12 adult male participants were recruited to engage in a two-staged experiment with an accelerometer attached to their body. In Stage 1, they were first required to perform 10 push-ups with fresh muscle strength to establish the acceleration and time-axis validity criteria. In Stage 2, they performed numerous push-ups, the data of which were compared with the validity criteria established in Stage 1 for a paired sample t test, with the level of significance set at $\alpha = 0.05$. The results indicated that the Stage 2 acceleration values of the third to seventh push-ups were significantly faster than the acceleration validity criterion, which was the mean of the 10 push-ups performed in Stage 1. However, the acceleration values of the twenty-second to thirtieth push-ups were significantly lower than the acceleration validity criterion. Regarding the time-axis readings, in Stage 2, the nineteenth to thirtieth push-ups took significantly more time than the mean of the 10 push-ups performed in Stage 1. These results indicate that acceleration and time-axis readings can be used to effectively monitor exercise status during push-up exercise and to help reduce possible injuries; however, time-axis readings without the acceleration values cannot be regarded as a full representation of exercise status.

Y.-Y. Chen

Department of Physical Education, National Taiwan Normal University, Taipei City, Taiwan

L.-W. Sha

Jilin Sport University, Changchun, China

Department of Educational Administration and Management, National Dong Hwa University, Hualien, Taiwan

W.-H. Chiu (✉)

Department of Physical Education, National Tsing Hua University, Hsinchu City, Taiwan

e-mail: whchiu@mx.nthu.edu.tw

Center for Sport Technology, Hsinchu City, Taiwan

© Springer Nature Singapore Pte Ltd. 2021

S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation, Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_4

4.1 Introduction

Muscle fatigue in sustained exercise can lower the benefits of exercise and raise the risk of injuries [1]. From the perspective of a trainer, his or her duty is to enhance an athlete's skills, and more importantly, to supervise the athlete's status while in exercise to prevent injuries [2]. Taking the push-up exercise as an example, it can effectively train the muscles in the upper extremities and torso [3]. However, as the number of push-ups increases, muscle strength wanes and fatigue sets in, lengthening the time required to complete a push-up. Consequently, muscle groups that are not supposed to be used in push-ups may be activated to help maintain balance and complete the move. This behavior, known as "muscular compensation," can deteriorate exercise status. Without the supervision of a trainer, this can lead to injuries [4]. How modern sensors and associated technologies can be applied to help trainers monitor exercise status, keep athletes safe, and prevent sports injuries has become an extensively researched topic in recent years [5].

The rapid evolution of wearable devices has made triaxial accelerometers much more affordable. With qualities such as being small, portable, and highly reliable, accelerometers have been widely applied in smartphones and fitness trackers. Numerous studies have employed gas analyzers to compare accelerometers' readings with actual energy consumption in bodily exercise, and all have concluded that accelerometers that generally provide reasonably accurate are reasonably reliable [6, 7]. Acceleration measured by accelerometers can also be used to measure activities of daily living; for example, it can be used to assess older adults' levels of physical activity as well as the exercise status of children. Moreover, time-axis readings can be used to determine differences in older adults' stride frequency, and the amount of time children remain sedentary, active, and asleep. This means these two sets of parameters can be useful when monitoring the walking and exercise statuses and lifestyles of older adults and children [8–10].

Some studies have used acceleration and the time axis to monitor athletes in ball games and swimming in order to determine skill level, frequency of strength training, and motor time [11, 12]. The rationale is that accelerometers primarily rely on piezoelectric crystals to monitor the acceleration of a movement and provide readings as output voltage values. When an athlete moves while wearing an accelerometer, a larger movement requires greater acceleration, and this exhibits a linear relationship with actual energy consumption. The level of correlation is particularly high for repetitive movements, such as walking and running. Therefore, it is "indeed feasible" to assess exercise status, level of physical activity, frequency of physical activity, and motor time using accelerometers [13, 14].

Although the aforementioned studies have pioneered the use of accelerometers in monitoring exercise status and sports skills, whether accelerometers can be used to monitor fatigue remains unclear. The push-up is one of the most popular body-weight exercises. However, doing an insufficient number of push-ups can reduce the effectiveness of training, whereas doing too many can result in delayed onset muscle

soreness. Therefore, the feasibility of applying accelerometers during highly repetitive push-up sessions to help trainers monitor exercise status and prevent injuries should be determined. Accordingly, this study conducted a two-staged experiment, first monitoring push-ups when participants' muscle strength was fresh and unspent to establish validity criteria for acceleration and the time axis (Stage 1), followed by consecutive push-up sessions (Stage 2). Subsequently, the differences in acceleration and time-axis readings were compared between the two stages.

4.2 Methodology

4.2.1 Participants

The participants were 12 adult men with no particular proficiency for any given sport but who were familiar with the push-up exercise (age, 21.3 ± 1.2 years; height, 173.2 ± 2.2 cm; weight, 69.3 ± 5.1 kg). None of the participants had sustained any serious joint or muscle injuries to the upper or lower extremities in the past year. The researchers briefed them on the procedures of the experiment, any matters of note, and their rights, after which they were asked to sign a letter of consent for participating in the experiment. This study was approved by the Tsing Hua University of Taiwan institutional review board.

4.2.2 Phases of a Push-up

A push-up can be divided into upward and downward phases [4], as muscle strength gradually weakens after numerous push-ups, muscular compensation sets in, with the upward phase being more prone to muscular compensation than the downward phase is [15]. In the present study, the push-up was also broken into upward and downward phases, and the signals detected by an infrared grating were used for movement phasing. This was to ensure that there was a standard for counting the upward phase, which is a critical reference for the calculation of acceleration. Figure 4.1 illustrates the downward and upward phases of a push-up.

4.2.3 Defining the Validity Criteria for Exercise Status

This study defined exercise status as the movement speed of the human body in a specific direction in a reference frame with respect to time. An accelerometer and a metronome (SEIKO-DM51, Japan, 60 bpm) were employed to monitor exercise status during the push-up session, with the acceleration and time-axis readings as the

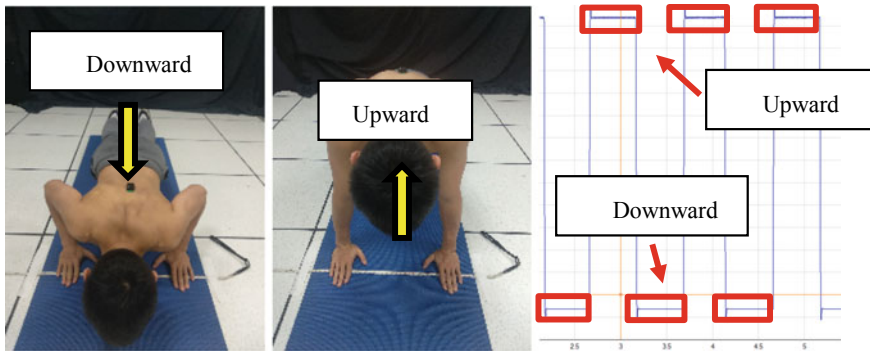


Fig. 4.1 Phases of a push-up

validity criteria. The literature indicates that the positioning of the accelerometer is unimportant as long as it does not impede movement, because no matter which part of the human body the accelerometer is attached to, the generated acceleration readings are similar [16, 17]. Considering that a push-up primarily works the pectoralis major and arm muscles, the wireless accelerometer was attached to the participants' trapeziuses at a position close to the tenth thoracic vertebra. The participants performed push-ups with fresh muscle strength to set up the validity criteria for the acceleration and time-axis readings.

4.2.4 Experimental Control

To prevent irregularities in the posture and frequency of the push-ups from affecting the results, the participants were required to place both arms directly under the shoulders, with both palms pointing toward the front, the torso straight, legs closed, and both hands shoulder-distance apart, which was defined as the distance between both acromion processes. The upper arms had to be parallel to the floor when reaching the end of the downward phase. Throughout the whole process, the participants had to keep their heads steady and eyes looking at the floor while they followed the beat of the metronome (60 bpm) to perform one push-up per second [18].

4.2.5 Procedure

In Stage 1, the participants were required to perform push-ups with fresh muscle strength while wearing an accelerometer (Delsys, three-axis, 16 bit \pm 5 g, 150 Hz), which took the readings of the first 10 push-ups to establish the validity criteria for the acceleration and time-axis readings. In Stage 2, the accelerometer took the vertical

acceleration of 30 consecutive push-ups. In both stages, the push-ups were performed to the rhythm of a metronome, and the acceleration and time-axis readings taken in Stage 1 served as the validity criteria for the readings taken in Stage 2 to determine the feasibility of using accelerometers to monitor people's exercise status when performing push-ups. A measuring tape was used to determine each participant's shoulder width, which was measured as the distance between both acromion processes. This width was then adopted as the participants' standard distance between two upper extremities when performing push-ups [19]. The height of a participant's abdominal external oblique muscles when he was at the end of the upward phase was set as the height of the infrared gratings for that individual. In Stage 1, a wireless accelerometer was attached to the participant's tenth thoracic vertebra, after which he was required to assume a prone position on a yoga mat with both hands apart at the same width as his shoulders and his legs closed. After the accelerometer was activated and the infrared gratings synchronized, the participant was required to perform 10 push-ups to the rhythm of the metronome to ensure that the acceleration could be collected. The participant was required to perform the push-ups strictly according to the study protocol. Following Stage 1, rest for 3 min, after which Stage 2 began, in which he was asked to perform 30 consecutive push-ups in the same manner as in Stage 1. To determine the feasibility of using accelerometers to monitor exercise status while performing push-ups, regardless of whether the participants could maintain the established rhythm due to weakened muscle strength, they were asked to keep performing push-ups until they reached the required number. The participants only completed Stage 1 and Stage 2 once each. Figure 4.2 illustrates the experiment setting.

The acceleration collected in the experiments was analyzed using Delsys EMGworks Analysis software. The data were smoothed by a low-pass filter with a cutoff frequency of 13 Hz, after which the vertical acceleration was calculated by applying the root mean square formula [20]. Subsequently, each participant's infrared gratings signals and time-axis readings were synchronized and compared with the vertical

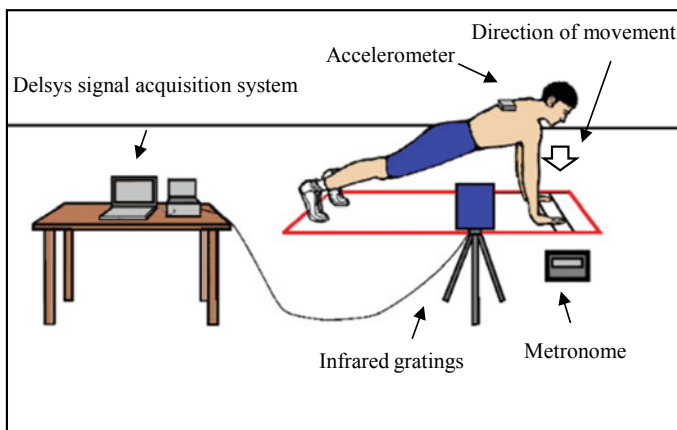


Fig. 4.2 Experiment setting

acceleration in each upward phase, thus obtaining the vertical-axis (Z-axis) acceleration for use as the acceleration standards in this experiment (unit: g). Regarding the time-axis readings, the validity criterion was 60 bpm. The time-axis readings from the infrared gratings signals for one downward movement to one upward movement recorded in Stage 2 were compared to determine the time each participant spent completing a push-up. The mean of the readings was adopted as the standard for the time-axis readings. Subsequently, the vertical acceleration and time-axis validity criteria determined in Stage 1 and the mean vertical accelerations and mean time-axis readings (i.e., the time required to complete a push-up) in Stage 2 were used to conduct a paired-sample *t* test using IBM SPSS 23 software. The changes in the participants' mean vertical accelerations and time-axis readings revealed how acceleration varied with time, which was used as the basis for monitoring exercise status. The level of significance was set at $\alpha = 0.05$.

4.3 Results

Table 4.1 presents the results for acceleration, showing that the third to seventh push-ups in Stage 2 were noticeably faster than the acceleration validity criterion obtained in Stage 1. This indicates that at the time the participants still had fresh muscle strength, enabling greater acceleration than the validity criterion. However, as the number of push-ups increased, the acceleration became considerably lower than the validity criterion, as shown from the results for the twenty-second to thirtieth push-ups. This suggests that, by the time participants reached the twenty-second push-up, their muscle strength had gradually waned, causing acceleration to slow down. The

Table 4.1 Comparison of the acceleration validity criterion and mean vertical acceleration

Acceleration validity criterion	Push-ups 1–10	<i>T</i>	Push-ups 11–20	<i>t</i>	Push-ups 21–30	<i>t</i>
1.02 ± 0.10	1.01 ± 0.11	1.256	1.07 ± 0.13	-1.799	0.92 ± 0.13	2.029
	1.08 ± 0.15	-1.921	1.05 ± 0.15	-0.806	0.88 ± 0.09	0.473*
	1.09 ± 0.14	-2.622*	1.04 ± 0.16	-0.658	0.82 ± 0.07	6.412*
	1.12 ± 0.18	-2.396*	1.04 ± 0.12	-0.63	0.77 ± 0.05	10.677*
	1.13 ± 0.18	-2.823*	1.03 ± 0.12	-0.614	0.82 ± 0.09	7.124*
	1.11 ± 0.10	-3.506*	1.01 ± 0.12	-0.214	0.78 ± 0.05	7.141*
	1.09 ± 0.12	-2.611*	1.01 ± 0.13	-0.213	0.75 ± 0.05	8.508*
	1.07 ± 0.16	-1.171	1.02 ± 0.12	-0.426	0.73 ± 0.03	8.061*
	1.06 ± 0.13	-1.076	0.99 ± 0.17	-0.491	0.73 ± 0.04	7.189*
	1.04 ± 0.13	-0.638	0.98 ± 0.15	-0.729	0.71 ± 0.05	9.441*

Note **p* < 0.05; unit: g

apparently lowered acceleration indicates that fatigue set in after the twenty-second push-up.

Several studies have measured the acceleration values generated by exercises and used these values as standards to monitor the physical activities of adult women and children. In the event that the measured acceleration of a certain physical activity is greater than that of the relevant validity criterion, the signal is considered a valid datum for the physical activities of adult women and children [14, 21, 22].

From the perspective of the present study, if the acceleration measured in Stage 2 did not exhibit a significant difference with the acceleration validity criterion from Stage 1 (e.g., the results for the first to second and eighth to twenty-first push-ups in Table 4.1), the participants' muscle strength was still fresh and unspent. Conversely, if the acceleration measured in Stage 2 exhibited a significant difference with the acceleration validity criterion from Stage 1, there were two possible meanings: (1) If the acceleration measured in Stage 2 was significantly greater than the acceleration validity criterion (e.g., the third to seventh push-ups in Table 4.1), the participants' muscle strength was still fresh. (2) If the measured acceleration was significantly lower than the acceleration validity criterion (e.g., the twenty-second to thirtieth push-ups in Table 4.1), the participants were fatigued and their muscle strength was waning. This lowered their acceleration and could possibly compel their bodies to resort to muscular compensation, raising the risk of injuries. In situations such as this, exercise should be stopped immediately [2, 4].

Table 4.2 presents the results for the time-axis readings. From the first to the tenth push-ups in Stage 2, the participants' exhibited no significant difference from the time-axis validity criterion in Stage 1, suggesting that the participants could complete each push-up within the time limit defined in the Experimental Control Section. This also suggests that the participants' exercise statuses had not yet deteriorated.

Table 4.2 Comparison of the time-axis validity criterion and mean time of push-ups

Time-axis validity criterion	Push-ups 1–10	<i>t</i>	Push-ups 11–20	<i>t</i>	Push-ups 21–30	<i>t</i>
1	1.02 ± 0.11	-0.544	1.07 ± 0.06*	-3.261	1.21 ± 0.20*	-2.854
	1.08 ± 0.11	-1.778	1.07 ± 0.14	-1.406	1.28 ± 0.10*	-7.012
	1.05 ± 0.09	-1.335	1.04 ± 0.11	-0.948	1.35 ± 0.14*	-6.759
	1.01 ± 0.10	0.001	1.02 ± 0.07	0.095	1.34 ± 0.07*	-12.69
	0.99 ± 0.05	0.386	1.06 ± 0.10	-1.499	1.33 ± 0.10*	-8.624
	1.01 ± 0.08	-0.176	1.06 ± 0.12	-1.386	1.41 ± 0.12*	-9.197
	1.02 ± 0.09	-0.559	1.07 ± 0.12	-1.682	1.39 ± 0.12*	-8.328
	1.05 ± 0.13	-1.049	1.07 ± 0.13	-1.371	1.41 ± 0.15*	-7.134
	1.05 ± 0.08	-1.614	1.15 ± 0.12*	-2.503	1.42 ± 0.16*	-6.931
	1.04 ± 0.08	-1.341	1.15 ± 0.17*	-2.376	1.21 ± 0.20*	-2.854

Note **p* < 0.05; unit: seconds

However, the time the participants spent for the eleventh to nineteenth push-ups was considerably longer than the validity criterion, suggesting that their exercise statuses had started to deteriorate, causing them to spend more time on each push-up. This became even more noticeable between the twenty-first and thirtieth push-ups. Therefore, the time-axis readings indicate that, when the participants engaged in consecutive push-ups, their exercise statuses gradually deteriorated and by the time they reached the nineteenth push-up, the time they spent on one push-up had become considerably longer than the validity criterion, which is indicative of much lower acceleration.

The time-axis readings of an accelerometer can be used to record the stride frequency, steps per minute, and swing time during running exercise [22]. Furthermore, the readings can serve as reference for the estimation of exercise status, and after comparison with the measured acceleration values, can be utilized for precise assessment of physical activity level and exercise status. This enables monitoring of exercise intensity [21, 23–26]. This concept was applied in the present study in the form of validity criterion that required the participants to perform one push-up per second. By examining the accelerometer's time-axis readings, the time the participants spent on one push-up was monitored. The results indicated that, if the time the participants spent did not exhibit a significant difference from the time-axis validity criterion (e.g., the first to tenth and the twelfth to eighteenth push-ups in Table 4.2), the participants still had fresh muscle strength. However, as the number of push-ups increased, their muscle strength gradually became spent, and the time they took to complete a push-up lengthened. When the time they took became significantly longer than the time-axis validity criterion (e.g., the eleventh and the nineteenth to thirtieth push-ups in Table 4.2), it was indicative of deteriorated muscle strength and exercise status, which increased the time for each movement due to lowered acceleration. When this occurs, exercise should be stopped immediately to prevent injury [2, 4].

This study cross-compared the Stage 2 data that exhibited a significant difference with the acceleration and time-axis validity criteria from Stage 1. The findings revealed that when the participants were performing the third to seventh push-ups in Stage 2, their acceleration values were significantly greater than the acceleration validity criterion (Table 4.1), but no significant difference was noted in the time they spent versus the time-axis validity criterion (Table 4.2). This proves that the participants' muscle strength was still fresh at the time. When the participants were performing the eleventh and nineteenth to thirtieth push-ups in Stage 2, no significant difference was noted between the acceleration values and the acceleration validity criterion (Table 4.1), but the time spent was significantly longer than the time-axis validity criterion (Table 4.2). This suggests that the longer time taken to complete the press-ups at this point could not be entirely attributed to lowering of acceleration. However, by the time the participants reached the twenty-second to thirtieth push-ups in Stage 2, their acceleration values had become significantly lower than the acceleration validity criterion, and the time they spent was significantly longer than the time-axis criterion. Moreover, as the number of push-ups increased, their acceleration values dropped and the time they spent increased, suggesting that their muscle strength was waning. This observation agrees with another study that when

muscle strength deteriorates, measured acceleration decreases, but the time spent on each movement increases.

4.4 Conclusion

This study examined whether accelerometers can be used to monitor exercise status. Pre-established validity criteria were employed to assess performance during push-ups based on acceleration and time-axis readings measured by an accelerometer, and these validity criteria were compared with actual measurements while participants engaged in push-ups. This approach was designed with the goal of precisely monitoring athletes' exercise statuses, thereby helping trainers to execute their training programs. The results proved that acceleration values measured by accelerometers could be used to monitor exercise status, and time-axis readings could be used to determine the time spent on each movement. Notably, the results revealed that the length of time spent on each move could not accurately represent exercise status alone without the relevant acceleration data. Specifically, an individual's exercise status could only be regarded as significantly deteriorated when the acceleration became significantly lower than the acceleration validity criterion and the time spent on a movement became significantly longer than the time-axis validity criterion. Only when this prerequisite was satisfied could exercise status monitoring be considered effective. In summary, acceleration and time-axis readings measured by accelerometers can be used in conjunction to help trainers monitor athletes' exercise statuses and to prevent possible injuries.

References

1. Montalvo, A.M., Shafer, H., Rodriguez, B., Li, T., Epnere, K., Myer, G.D.: Retrospective injury epidemiology and risk factors for injury in crossfit. *J Sport Sci Med* **16**(1), 53–59 (2017). <https://doi.org/10.1249/01.mss.0000487599.90040.1b>
2. Liu, H., Garrett, W.E., Moorman, C.T., Yu, B.: Injury rate, mechanism, and risk factors of hamstring strain injuries in sports: a review of the literature. *J Sport Health Sci* **1**(2), 92–101 (2012). <https://doi.org/10.1016/j.jshs.2012.07.003>
3. Anderson GS, Gaetz M, Holzmann M, Twist P (2011) Comparison of EMG activity during stable and unstable push-up protocols. *Eur J Sport Sci* 42–48. <https://doi.org/10.1080/17461391.2011.577240>
4. Mok, K.M., Ho, C., Yung, P.S.H., Chan, K.M.: Are the muscle activations different in various type of push-up exercise? *Br J Sport Med* **51**(4), 363–364 (2017). <https://doi.org/10.1136/bjsports-2016-097372.202>
5. Ma, J., Kharboutly, H., Benali, A., Benamar, F.: Joint angle estimation with accelerometers for dynamic postural analysis. *J. Biomech.* **48**(13), 3616–3624 (2015). <https://doi.org/10.1016/j.jbiomech.2015.08.008>

6. Balsalobre-Fernández, C., Kuzdub, M., Poveda-Ortiz, P., Campo-Vecino, J.: Validity and reliability of the push wearable device to measure movement velocity during the back squat exercise. *J Strength Cond Res* **30**(7), 1968–1974 (2016). <https://doi.org/10.1519/JSC.0000000000001284>
7. Miller, N.E., Strath, S.J., Swartz, A.M., Cashin, S.E.: Estimating absolute and relative physical activity intensity across age via accelerometry in adults. *J Aging Phys Act* **18**(2), 158–170 (2010). <https://doi.org/10.1123/japa.18.2.158>
8. Barnett, A., Van den Hoek, D., Barnett, D., Cerin, E.: Measuring moderate-intensity walking in older adults using the actigraph accelerometer. *BMC Geriatrics* **211**(16), 1–9 (2016). <https://doi.org/10.1186/s12877-016-0380-5>
9. Vanhelst, J., Béghin, L., Turck, D., Gottrand, F.: New validated thresholds for various intensities of physical activity in adolescents using the actigraph accelerometer. *Int. J. Rehabil. Res.* **34**(2), 175–177 (2011). <https://doi.org/10.1097/MRR.0b013e328340129e>
10. Verloigne M, lippevelde WV, Maes L, Yildirim M, Chinapaw M, Manios Y, Androutsos O, Kovács É, Bringolf-Isler B, Brug J, De Bourdeaudhuij ID (2011) Levels of physical activity and sedentary time among 10- to 12-year-old boys and girls across 5 european countries using accelerometers: an observational study within the energy-project. *Int J Behav Nutr Phys Act* **34**(9): 1–8. <https://doi.org/10.1186/1479-5868-9-34>
11. Stamm, A., James, D.A., Burkett, B.B., Hagem, R.M., Thiel, D.V.: Determining maximum push-off velocity in swimming using accelerometers. *Procedia Eng* **60**, 201–207 (2013). <https://doi.org/10.1016/j.proeng.2013.07.067>
12. Nurwanto F, Ardiyanto I, Wibirama S (2016) Light sport exercise detection based on smartwatch and smartphone using k-nearest neighbor and dynamic time warping algorithm. In: 2016 8th international conference on information technology and electrical engineering, Yogyakarta, Indonesia. <https://doi.org/10.1109/icitheed.2016.7863299>
13. Khoshnoud, F., De Silva, C.W.: Recent advances in MEMS sensor technology-mechanical applications. *Instrum Meas Mag, IEEE* **15**(2), 14–24 (2012). <https://doi.org/10.1109/MIM.2012.6174574>
14. Shiang, T.Y., Shih, S., Ho, C.S.: The applications of sensor technology for exercise and sport science. *Phys Educ J* **45**(1), 1–12 (2012). <https://doi.org/10.6222/pej.4501.201203.0701>
15. Harris, S., Ruffin, E., Brewer, W., Ortiz, A.: Muscle activation patterns during suspension training exercises. *Int J Sport Phys Ther* **12**(1), 42–52 (2017). <https://doi.org/10.1249/01.mss.0000479278.25611.a6>
16. Nilsson, A., Ekelund, U., Yngve, A., Sjöström, M.: Assessing physical activity among children with accelerometers using different time sampling intervals and placements. *Hum Kinet J* **14**(1), 87–96 (2002). <https://doi.org/10.1123/pes.14.1.87>
17. Phillips, L.R.S., Parfitt, G., Rowlands, A.V.: Calibration of the GENEa accelerometer for assessment of physical activity intensity in children. *J Sci Med Sport* **16**(2), 124–128 (2013). <https://doi.org/10.1016/j.jsams.2012.05.013>
18. Marcolin, G., Petrone, N., Moro, T., Battaglia, G., Bianco, A., Paoli, A.: Selective activation of shoulder, trunk, and arm muscles: a comparative analysis of different push-up variants. *J Athl Train* **50**(11), 1126–1132 (2015). <https://doi.org/10.4085/1062-6050-50.9.09>
19. Batbayar, Y., Uga, D., Nakazawa, R., Sakamoto, M.: Effect of various hand position widths on scapular stabilizing muscles during the push-up plus exercise in healthy people. *J Phys Ther Sci* **27**(8), 2573–2576 (2015). <https://doi.org/10.1589/jpts.27.2573>
20. Zijlstra, W., Bisseling, R.W., Schlumbohm, S., Baldus, H.: A body-fixed-sensor-based analysis of power during sitto-stand movements. *Gait Posture* **31**, 272–280 (2010). <https://doi.org/10.1016/j.gaitpost.2009.11.003>
21. Neil-Sztramko, S.E., Rafn, B.S., Gotay, C.C., Campbell, K.L.: Determining activity count cut-points for measurement of physical activity using the Actiwatch2 accelerometer. *Physiol. Behav.* **173**, 95–100 (2017). <https://doi.org/10.1016/j.physbeh.2017.01.026>
22. Shull, P.B., Jirattigalachote, W., Hunt, M.A., Cutkosky, M.R., Delp, S.L.: Quantified self and human movement: a review on the clinical impact of wearable sensing and feedback for gait analysis and intervention. *Gait Posture* **40**(1), 11–19 (2014). <https://doi.org/10.1016/j.gaitpost.2014.03.189>

23. Jarning, J.M., Mok, K.M., Hansen, B.H., Bahr, R.: Application of a tri-axial accelerometer to estimate jump frequency in volleyball. *Sport Biomech* **14**(1), 95–105 (2015). <https://doi.org/10.1080/14763141.2015.1027950>
24. Park, J., Ishikawa-Takata, K., Tanaka, S., Bessyo, K.: Accuracy of estimating step counts and intensity using accelerometers in older people with or without assistive devices. *J Aging Phys Act* **25**(1), 41–50 (2017). <https://doi.org/10.1123/japa.2015-0201>
25. Thomas E, Bianco A, Raia T, Messina G, Messina G, Tabacchi G, Bellafiore M, Paoli A, Palma, A (2018) Relationship between velocity and muscular endurance of the upper body. *Hum Mov Sci* **60**: 175–182. <https://doi.org/10.1016/j.humov.2018.06.008>
26. Wang, R., Hoffman, J.R., Sadres, E., Bartolomei, S., Muddle, T.W.D., Fukuda, D.H., Stout, J.R.: Evaluating upper-body strength and power from a single test: the ballistic push-up. *J Strength Cond Res* **31**(5), 1338–1345 (2017). <https://doi.org/10.1519/JSC.0000000000001832>

Chapter 5

Low-Coupling 2 * 1 Micro-strip Antenna Array Based on Defect Structure



Ning Guo, Xinliang Liu, and Nana Bu

Abstract The periodic defected ground structure (PDGS) has been proposed in this work to diminish high mutual coupling of antenna components. The PDGS is always periodically etched on the bottom ground plane along x -axis, placed at the center of the two neighboring patches with side-by-side distance of 0.025λ (λ free space wavelength). Both PDGS and antenna elements are not coplanar, working at the center frequency 5.8 GHz. By inserting a compact PDGS structure at the ground plane facing the center of the two patches, over 50 dB of coupling between adjacent antennas is suppressed within the effective operating frequency range of 5.7–5.9 GHz, achieving low coupling and compact type in the antenna arrays.

5.1 Introduction

With the development of integration, the array antenna is required not only to have high gain but also to meet the requirements of miniaturization. Intensive array antenna will be widely used [1], which is the development trend of array antenna. The closer the antenna elements are, the higher the coupling will be, which will distort the antenna performance. The farther the antenna elements are, the lower the coupling is, but it cannot be infinitely far away, which also distorts the antenna performance. By studying the principle of coupling formation, scientists successively proposed defect ground structure (DGS) [2], electrical tape gap structure (EBG) [1, 3], and stop-band filter structure [4–6] to suppress the coupling between antenna elements. In order to meet the growing demand for antenna performance, array antennas provide an irreplaceable solution for most wireless communication systems. Nevertheless, the narrow spacing of antenna elements always results in strong mutual coupling between adjacent patches, which can result in poor antenna performance [7]. Therefore, low mutual coupling of antenna components plays a significant role in wireless communication systems [8]. In previous studies, some have conducted to effectively

N. Guo (✉) · X. Liu · N. Bu
College of Communication and Information Engineering, Xi'an University of Science and Technology, 710000, Xi'an, China
e-mail: guoning123123@163.com

decrease the high coupling of antenna units in antenna array. Among them, the special band-gap structure [9] is an effective method to reduce the mutual coupling, which uses the band-stop characteristic to change the current distribution to reduce the strong mutual coupling of the antenna component. Defective ground structures (DGS) are also one of the most common structures for reducing strong mutual coupling. It has been used to degrade the mutual coupling of antenna components through interference field to change the current distribution of the antenna components [10, 11]. Another common way to reduce mutual coupling is to use an electromagnetic band-gap structure (EBG) [12], which can change the direction of wave propagation to decrease the mutual coupling of two adjacent patches. Other special structures can also achieve low coupling. In reference [13], the U-shaped micro-strip line structure embedded between the coupling elements reduces the strong mutual coupling of adjacent patches through interfering with a field between the two patches. In reference [14], an innovative H-type conductive wall structure is proposed which restrain mutual coupling and ameliorates isolation of the tightly spaced error antennas (MSA). Furthermore, in reference [15], mutual coupling reduction is achieved by controlling the polarization of the coupled field using a novel polarization conversion isolator (PCI). Although these proposed methods can reduce coupling, there are still drawbacks in terms of compactness and applicability. In any case, researchers still need to work hard on the decoupling between the antenna arrays, propose a better solution, and ensure low-coupling problems without affecting the compactness of the antenna array [16] and high gain. In this paper, the periodic defect structure (PDGS) is used to change the distribution of effective dielectric constant of the circuit substrate, thereby changing the distributed inductance and distributed capacitance of transmission line based on the medium, so that the transmission line has the characteristics of band gap and slow wave, and then suppress high coupling of adjacent patches without affecting the resonant frequency of the antenna array itself. By using the proposed periodic defect structure (PDGS), mutual coupling is greatly reduced, and the maximum value of mutual coupling reduction is greater than 50 dB. In addition, by comparing the un-decoupled antenna array with the decoupled antenna array, the gain of the decoupled antenna array proposed in this paper is increased about 2 dB in the range of effective frequency.

5.2 Decoupling Structure

Defect ground structure (DGS) is in the micro-strip and coplanar waveguide transmission line earth metal plate etching of periodic or aperiodic various grid structures, in order to change the distribution circuit substrate effective dielectric constant, which changes based on the medium distribution inductance and distributed capacitance of transmission line, makes this kind of transmission line with band-gap characteristics and slow wave. DGS is simple in structure, simple in manufacture, easy to integrate, and can be used in array antenna array effectively to realize the function of restraining coupling. As the number of DGS structures increases, the coupling factor of the coupler will increase [1]. The change of DGS cell size will cause the change

of equivalent inductance and capacitance of the micro-strip line, so as to realize the desired band resistance characteristics [2].

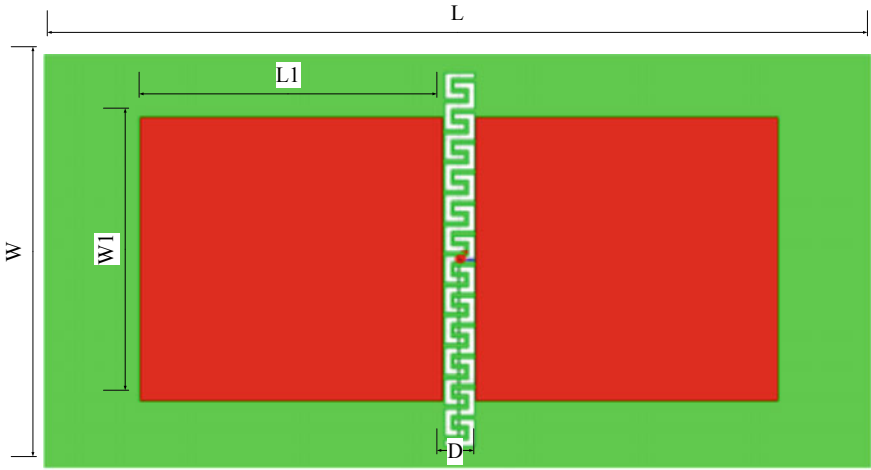
EBG, a periodic structure, can control the propagation of electromagnetic waves by correctly selecting the size, material, and shape of the scattering medium. EBG structure uses its surface wave band-gap characteristics to suppress the propagation of surface waves between antenna elements, so as to reduce the mutual coupling between antenna elements and improve the performance of antenna array. EBG improves antenna direction diagram, realizes micro-strip antenna miniaturization and low profile. The research shows that [3], the micro-strip antenna array is made of EBG structure as the substrate to replace the metal base plate of the antenna, which effectively improves the performance of the micro-strip antenna array. The results show that the array performance of micro-strip antenna with EBG structure is much better than that of ordinary array.

By comparison, the theoretical basis of defect ground structure (DGS) and electrical tape gap structure (EBG) is mature, which is widely used in micro-strip antenna, micro-strip filter, and other structures, and its special properties are used to make the antenna have the advantages of wide bandwidth, low coupling, and small size. It plays an important role in satellite communication, mobile communication, radar navigation, and other fields. However, there are still defects in compactness and applicability. So in this article, using periodic defect ground structures (PDGS) to change the distribution circuit substrate effective dielectric constant, which changes based on the medium distribution inductance and distributed capacitance of transmission line, make this kind of band-gap characteristics and slow potter, a transmission line, in turn, to curb high coupling between adjacent tiles, and shall not affect the resonance frequency of the antenna array itself.

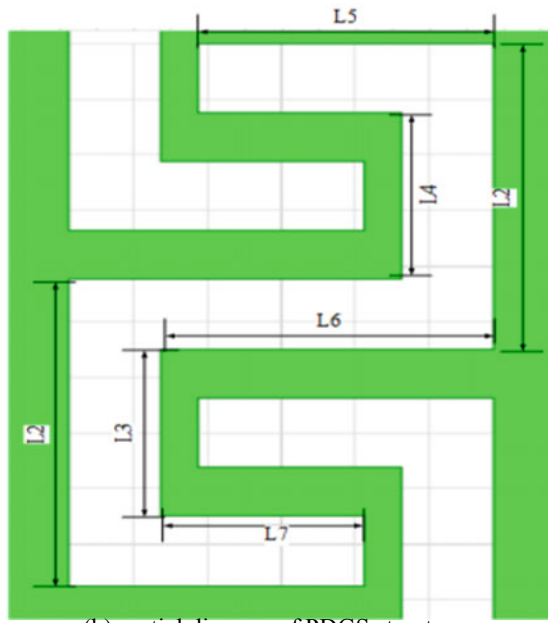
5.3 Antenna Design

Figure 5.1 shows the antenna array (Fig. 5.1a) and the decoupled structure PDGS unit (the model of Fig. 5.1b). The 2 * 1 antenna array and the PDGS structure (green part) are designed to work at the center frequency of 5.8 GHz, and working in the same frequency range. Two micro-strip patches (red part) are arranged along the x -axis with a distance $D = 1.3 \text{ mm}$ (0.025λ) between adjacent edges for good compactness. FR4 is adopted as the substrate of the antenna array, which has the dielectric constant of 4.4 and the loss tangent of 0.02. The overall size of the 2 * 1 patch array antenna is 35 mm (length) \times 22 mm (width) with the substrate thickness of 1.6 mm. the length ($L1$) and width ($W1$) of each patch are 11.68 mm and 11.8 mm, respectively.

The structure is formed by etching an empty white region on the ground plane to form a defect as shown in Fig. 5.1b, which minimizes mutual coupling between two adjacent patches. The PDGS structure suppresses surface current by changing the current distribution and isolates the patches finely at the target frequency. The optimization parameters of the DGS unit are: $L2 = 1.1 \text{ mm}$, $L3 = 0.6 \text{ mm}$, $L4 = 0.6 \text{ mm}$, $L5 = 0.8 \text{ mm}$, $L6 = 0.9 \text{ mm}$, and $L7 = 0.55 \text{ mm}$. The width of the PDGS structure is always 0.1 mm.



(a) Top view of antenna array



(b) partial diagram of PDGS structure

Fig. 5.1 Model of $2 * 1$ patch array antenna with DGS

5.4 Decoupling Performance Analysis

The simulated S parameters of an antenna array having a PDGS structure and no PDGS structure are shown in Fig. 5.2. The figure shows clearly that the resonant frequencies at S_{11} and S_{22} always maintain the same target frequency. In other words, compared to the decoupled antenna array, there is no difference in the resonant frequency of the 2 * 1 antenna array based on the PDGS structure, which indicates that the decoupling structure is effective. From Fig. 5.2, it can be clearly observed that the introduction of the decoupling structure PDGS will sharply reduce S_{12} from -14.8 dB to -51.5 dB, that is, the coupling between adjacent patches is reduced.

The current distribution is shown in Fig. 5.3. As shown in Fig. 5.3a, a large amount of energy accumulate between adjacent patches in 2 * 1 micro-strip antenna array without a PDGS structure, which results in the strong mutual coupling between the patches. On the contrary, the model of Fig. 5.3b, that is, the 2 * 1 antenna array based on the PDGS structure, the current density between adjacent patches is significantly reduced. The main reason is that when the left port is energized and the right port is terminated by 50Ω, it inhibits propagation from the left port to the right port, allowing energy to be concentrated on one side to minimize mutual coupling of the antenna components.

Figure 5.4a, b show the radiation patterns of 2 * 1 antenna array having a PDGS structure and no PDGS structure on the xz plane and yz plane. There is no meaningful change in the radiation pattern of the antenna array with PDGS structure and without the PDGS structure. That is to say, the PDGS structure proposed in the paper is effective for suppressing mutual coupling of antenna units. And it is worth noticing that the gain of the antenna is slightly enhanced compared to the gain before the decoupling. As shown in Fig. 5.5, the gain is increased by about 2 dB.

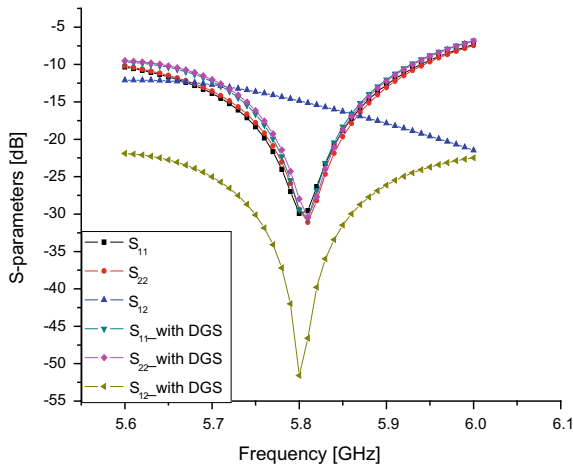


Fig. 5.2 S parameters of antenna model

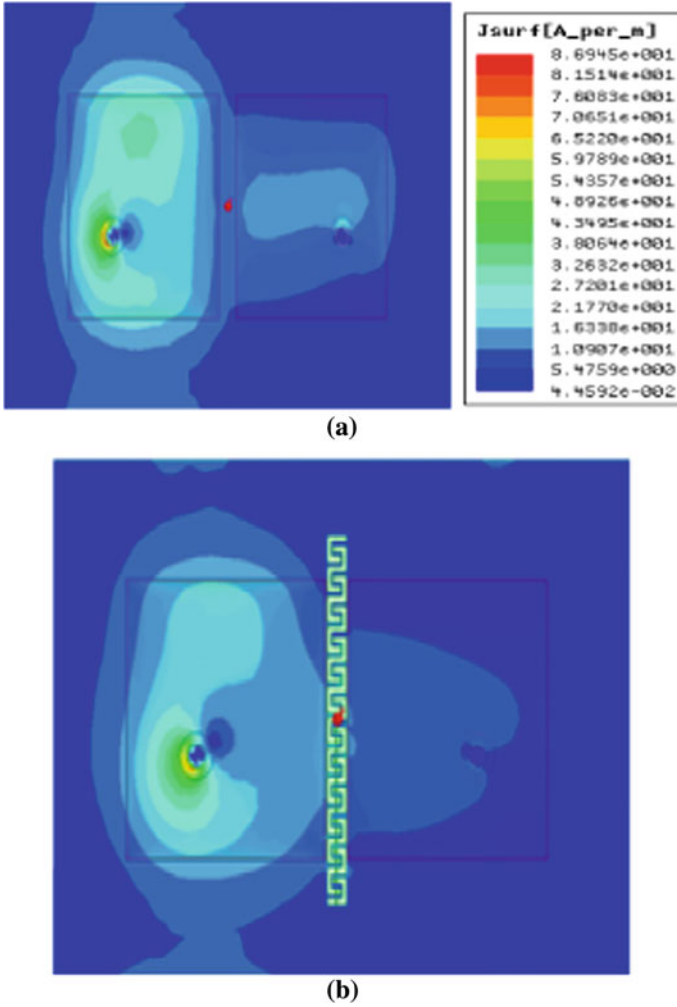
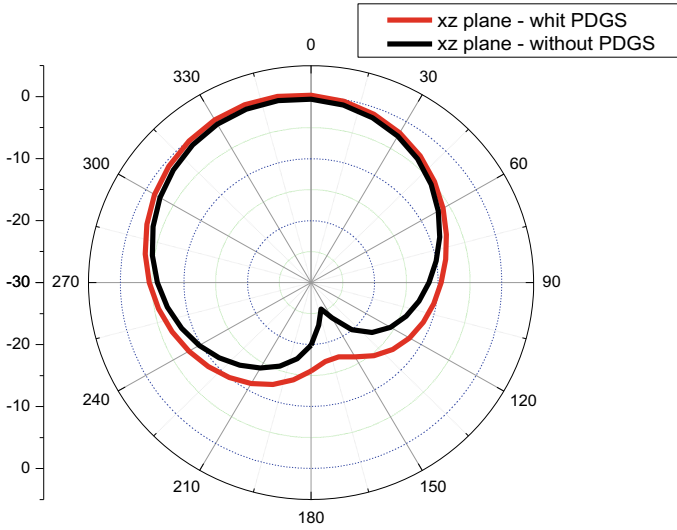
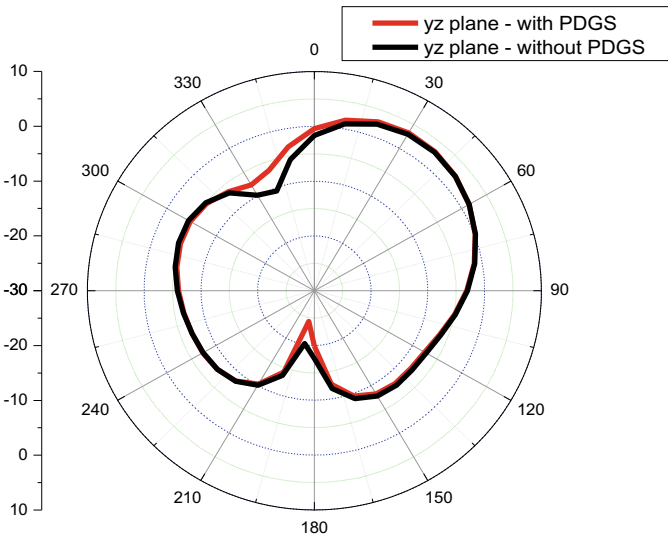


Fig. 5.3 Distribution of surface current on the ground level

As shown in Table 5.1, by comparing the decoupling structure proposed in this paper with the decoupling effect and other properties of the decoupling structure in the reference, it can be clearly obtained that the PDGS structure proposed in this paper is more compact and has higher isolation that is more applicable. In addition, there is no change in the resonant frequency, which means that the size of the PDGS structure can be adjusted to be suitable for antenna array decoupling in other frequency bands.



(a) The plane in xz



(b) The plane in yz

Fig. 5.4 Radiation pattern of models with and without PDGS

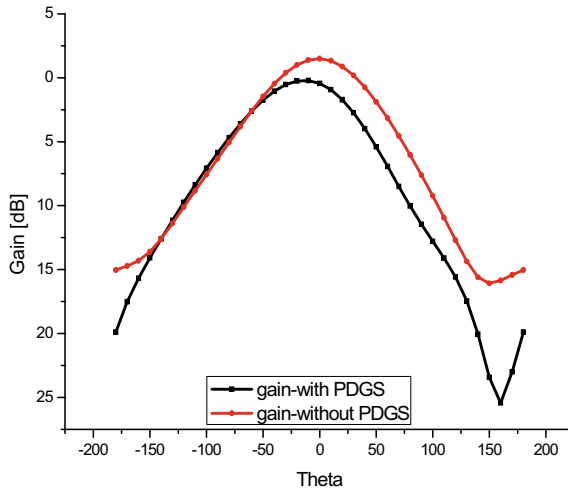


Fig. 5.5 PDGS structure and antenna gain without PDGS structure

Table 5.1 Comparison of performance of the proposed antenna with known research

References	Resonant frequency [GHz]	Adjacent spacing	Improve S_{12} [dB]	Resonant frequency shift [MHZ]	Decoupling structure
[3]	2.57	0.14λ	40	30	DGS
[5]	5.59	0.4λ	25.35	Not mentioned	EBG
[8]	5.8	0.18λ	22.3	20	PCI
This paper	5.8	0.025λ	35–45	0	PDGS

5.5 Conclusion

In this paper, periodic defect ground structure (PDGS) is used to change the distributed inductance and distributed capacitance of the transmission line based on this medium, so that such transmission line has band-gap characteristics and slow baud property, so as to suppress the high coupling between adjacent patches without affecting the resonant frequency of the antenna array itself. Through adopting the proposed periodic defect ground structure (PDGS), the maximum mutual coupling reduction is more than 50 dB. In addition, by comparing the un-decoupled antenna array with the decoupled antenna array, the gain of the decoupled antenna array proposed in this paper is increased by about 2 dB in the effective frequency range. The periodic defect ground structure (PDGS) designed in this paper can not only achieve low coupling between adjacent patches, but also meet the requirements of compact structure, and at the same time adapt to low-coupling antenna arrays of different frequency bands. In addition, the model presented in this paper shows that there is a smaller separation distance and greater isolation between adjacent patches than

the model mentioned in the reference. Furthermore, there is no marked difference between S_{11} and S_{22} , and achieved enhancement gain of about 2 dB. This paper presents an effective method to suppress the coupling between micro-strip antenna arrays by using the defect ground (DGS) structure. The defect structure is applied to the design of micro-strip antenna array. To demonstrate the effectiveness of this method, we analyze the coupling of two micro-strip antenna elements on a thick substrate with a large dielectric constant. Simulation and experimental results show that the introduction of the defective structure can effectively reduce the coupling between antenna elements, suppress the side lobe level, and improve the performance of the original antenna when the coupling between elements is significant. This structure has a good application prospect in mutual coupling suppression of micro-strip antenna array.

References

1. Fu S-H, Tong C-M, Li X-M, Shen K (2013) A novel small wide resistance live tape gap structure. *Des Appl* 882–885
2. Zhou D-L (2012) Study on coupling degree of a new type of directional coupler with defective ground structure. *Internet Things Technol* 3: 67–68
3. Li, Q., G-F, Zhang, Fu, Y.-Q., Yuan, N.-C.: Micro-strip antenna array loaded with a new type of electromagnetic tape gap structure. *J Natl Univ Def Technol* 32(4), 73–77 (2010)
4. Zhang Y, Zhang XY, Pan YM (2017) Compact single- and dual-band filtering patch antenna arrays using novel feeding scheme. *IEEE Trans Antennas Propag* 65(8)
5. Zhang Y, Zhang X-Y, Ye L-H, Pan Y-M (2016) Dual-band base station array using filtering antenna elements for mutual coupling suppression. *IEEE Trans Antennas Propag* 64(8): 1
6. Wei K, Li JY, Wang L, Xing ZJ, Xu R (2016) Mutual coupling reduction by novel fractal defected ground structure band-gap filter. *IEEE Trans Antennas Propag* 64(10): 1
7. Habashi A, Nourinia J, Ghobadi C (2011) Mutual coupling reduction between very closely spaced patch antennas using low-profile folded split-ring resonators (FSRRs). *IEEE Antennas Wirel Propagation Lett* 862–865
8. Chow, Y.L.: Mutual coupling between two micro-strip antennas. *J NanJing Univ Sci Technol (Nat Sci Ed)* 04, 1–10 (1982)
9. Wei K, Li JY, Wang L, Xing ZJ, Xu R (2016) Mutual coupling reduction by novel fractal defected ground structure band-gap filter. *IEEE Trans Antennas Propag* 4328–4335
10. Wei K, Li JY, Wang L, Xing ZJ, Xu R (2016) S-shaped periodic defected ground structures to reduce micro-strip antenna array mutual coupling. *Electron Lett* 1288–1290
11. Zhu F-G, Xu J-D, Xu Q (2009) Reduction of mutual coupling between closely-packed antenna elements using defected ground structure. *Electron Lett* 601–602
12. Mohamadzade B, Afsahi M (2017) Mutual coupling reduction and gain enhancement in patch array antenna using a planar compact electromagnetic band-gap structure. *IET Microw, Antenna Propag* 1719–1725
13. Farsi S, Aliakbarian H, Schreurs D, Nauwelaers B, Vandenbosch GA (2012) Mutual coupling reduction between planar antennas by using a simple micro-strip U-section. *IEEE Antennas Wirel Propag Lett* 1501–1503
14. Park C-H, Son H-W (2009) Mutual coupling reduction between closely spaced micro-strip antennas by means of H-shaped conducting wall. *Electron Lett* 1093–1094
15. Cheng YF, Ding X, Shao W, Wang BZ (2017) Reduction of mutual coupling between patch antennas using a polarization-conversion isolator. *IEEE Antennas Wirel Propag Lett* 1257–1260
16. Li, C., Xue, L., Feng, X.: Design of micro-strip antenna array based on mutual coupling reduction between very closely spaced patch antennas. *J Microw* 29(1), 74–77 (2013)

Chapter 6

Power Control in D2D Underlay Distributed Antenna Systems



Gongbin Qian, Ce Zhang, Chunlong He, Xingquan Li, and Chu Tian

Abstract A new scenario is considered that device-to-device (D2D) communication users underlay the spectrum resource of cellular user in distributed antenna systems (DAS) is discussed in this paper. We mainly focus on how to improve spectral efficiency (SE) and energy efficiency (EE) of the system. Under the maximum transmit power constraint per antenna unit, we propose two resource allocation algorithms to solve the optimal problems of maximum SE and EE. The first problem can be transformed into a difference of convex (DC) structure problem by function recombination, then the concave-convex procedure (CCCP) algorithm and the interior point method which are adopted to get the optimal solutions for the maximum SE. Subsequently, by using the Dinkelbach algorithm based on the parameter method, a power allocation algorithm for energy efficiency is developed to solve the maximum EE optimization problem. The optimal solutions are also obtained by the CCCP algorithm and the interior point method. Simulation results show that compared to co-located antenna systems (CAS) with D2D users, the SE and EE performances of the proposed system have a significant improvement.

Keywords Distributed antenna systems · Device-to-device · Spectral efficiency · Energy efficiency · Power allocation

6.1 Introduction

With the increasing demand for smartphones and fast mobile Internet services, the fifth generation (5G) of mobile networks is being researched to support large amounts of data traffic. One of the key performance indicators (KPIs) in future communication network design is the energy consumption, which means that spectral efficiency (SE) and energy efficiency (EE) are important factors in the 5G design. There are two techniques presented, they are: (i) Distributed antenna systems (DAS), because DAS

G. Qian · C. Zhang · C. He · X. Li (✉) · C. Tian
Shenzhen Key Laboratory of Advanced Machine Learning and Applications
Shenzhen University, Shenzhen 518060, China
e-mail: 2156130104@szu.edu.cn

© Springer Nature Singapore Pte Ltd. 2021
S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation,
Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_6

can reduce the communication distance between mobile phones and remote access units (RAUs), the DAS has many advantages to increase capacity, improve coverage and EE [1–3]; (ii) device-to-device (D2D) communication, that can underlay the spectral resource of cellular users to enable a user device communicating with another nearby user device directly without extra hop from base station. It increases the overall network SE and thus allows the network to admit more users [4, 5].

It is well known that power allocation will become an urgent problem in the future. In the field of DAS and D2D communication, there are a number of efficient approaches which have been presented to solve this problem [6–12]. For instance, for DAS, a power allocation approach to maximize SE has been provided for generalized DAS in [6]. The authors in [7] have proposed a power allocation approach to maximize the EE, which transforms the fractional form of non-convex problem into its equivalent subtractive form. For D2D communication, the authors considered maximizing sum-rate over signal to interference and noise ratio of the system in [10]. In order to keep the quality of service (QoS) of D2D users and cellular user equipments, a three-step approach has been presented to improve the total transmit rate of the system in [11].

The above methods all improve the performance of communication system. However, among the aforementioned power allocation approaches, there is no paper considering the scenario of coexistence of DAS and D2D communication. In this paper, to further improve the performance of system, a new scenario for D2D communication underlaid DAS is proposed. We mainly focus on how to improve the SE and EE of the system. We first convert the maximizing SE and EE objective functions to a DC problem by function reorganization, CCCP algorithm and the interior point method which are presented to get optimal solutions. In particular, the Dinkelbach algorithm based on the parameter method is utilized in EE power allocation algorithm, and we transform the fractional form of EE optimization into a subtractive form that is easier to solve. In order to confirm the reliability of the proposed algorithm, we also compare with co-located antenna systems (CAS) with D2D communication [13]; experiment results demonstrate that the proposed algorithm has a better performance. Unlike the existing approaches, the proposed one has a good performance in improving system efficiency. So it is a key technique for the future communication systems.

6.2 System Model and Problem Formulation

6.2.1 System Model

In this section, the model of D2D user underlaying the spectral resource of cellular user in DAS is established. We consider downlink transmission in a cellular network where UE and D2D pairs use the same frequency bands. The locations of N RAUs are uniformly located in the cell and connected to the central base station (e.t. RAU1) through optical fiber. In one cell, there are M cellular user equipments (UEs) and K D2D pairs, and they are both equipped with one single antenna. Each channel under-

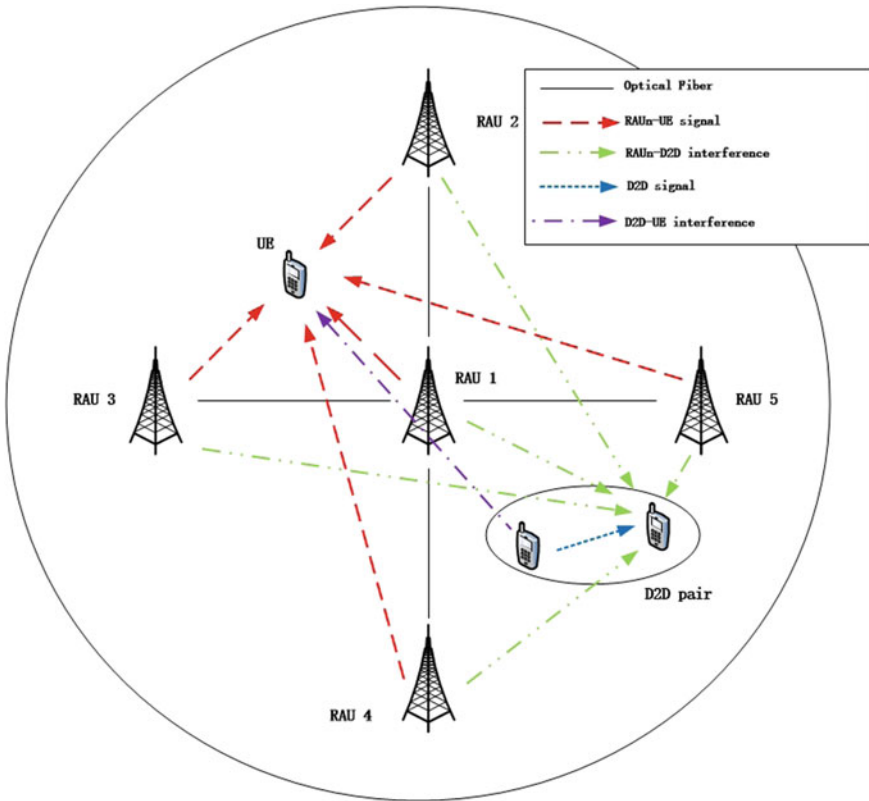


Fig. 6.1 DAS with D2D system model

goes independent and identically distribution (i.i.d.). We can define configuration specified in the system as (N, M, K) . For example, Fig. 6.1 is a $(5, 1, 1)$ system that is discussed in this paper, where $M = 1, N = 5, K = 1$. In addition, there are two special cases

1. $(N, M, 0)$ stands for the DAS with fully distributed antennas;
2. $(1, M, 0)$ represents the co-located antenna system (base station can equip with multiple antennas).

6.2.2 Channel Model

In this paper, $h_{n,c}$ denotes the channel which consists of a small- and large-scale fading, which can be written as [14, 15]

$$h_{n,c} = g_{n,c} w_{n,c}, \tag{6.1}$$

where

$$w_{n,c} = \sqrt{\frac{cs_{n,c}}{d_{n,c}}}, \quad (6.2)$$

$g_{n,c}$ represents the small-scale fading, $g_{n,c} \sim \mathcal{CN}(0, 1)$ and $w_{n,c}$ represent the large-scale fading factor, which has no relationship with $g_{n,c}$. c denotes the median of the mean path gain, $d_{n,c}$ denotes the communication distance between cellular user and RAU n , α and $s_{n,c}$ are constants.

6.2.3 Achievable Rate

We assume that the RAUs and UEs in the system can easily get the channel state information (CSI) and the total system bandwidth is 1 MHz.

The following parameters are used in the description of the system model

- p_d : Transmit power of the D2D transmitter.
- $p_{n,c}$: Transmit power from the n th RAU to the UE.
- P_{\max}^d : Maximum transmit power of the D2D transmitter.
- P_{\max}^n : Maximum transmit power of RAU n .
- $h_{n,c}$: The channel gain from the RAU n to cellular user.
- h_d : The channel gain from the D2D transmitter to D2D receiver.
- $h_{d,c}$: The channel gain from the D2D transmitter to UE.
- $h_{n,d}$: The channel gain from the D2D transmitter to RAU n .
- σ_c^2, σ_d^2 : The power of the white Gaussian noise (AWGN) of UE and D2D user, respectively.
- R_c : The transmission rate of UE.
- R_d : The transmission rate of D2D user.

The SE of UE and D2D user is expressed as follows

$$R_c = \log_2 \left(1 + \frac{\sum_{n=1}^N p_{n,c} |h_{n,c}|^2}{p_d |h_{d,c}|^2 + \sigma_c^2} \right), \quad (6.3)$$

$$R_d = \log_2 \left(1 + \frac{p_d |h_d|^2}{\sum_{n=1}^N p_{n,c} |h_{n,d}|^2 + \sigma^2} \right). \quad (6.4)$$

6.3 Objective Optimization Formulation

In the first part, the maximum SE optimization problem is discussed. Then, the EE optimization model is considered in the second part including the power consumption of circuit and optical fiber. Finally, an effective power allocation scheme is presented to maximizing the EE of system.

6.3.1 Maximum SE Optimization

Due to the D2D pair and UE use the same spectrum at the same time, there exists interference between them, which makes the problem becomes more complicated. It can be modeled as

$$\begin{aligned} \max_{\mathbf{P}} \quad & SE \triangleq R_c + R_d \\ \text{s.t.} \quad & p_{n,c} \in [0, P_{\max}^n] \quad \forall n \in \{1, 2, \dots, N\}, \\ & p_d \in [0, P_{\max}^d]. \end{aligned} \quad (6.5)$$

where $\mathbf{P} \triangleq [\mathbf{p}, p_d]$, $\mathbf{p} = \{p_{n,c}, \text{ for } n = 1, 2, \dots, N\}$.

Readjusting the expression of the objective function (6.5), we can find that the objective function has a special DC structure. We can exploit the similar methods based on DC structure to solve the optimization problem [16–18]. Let $f_{se}(\mathbf{P})$ represents the variable and objective functions in (6.5), respectively. So the (6.5) can be decoupled as

$$f_{se}(\mathbf{P}) = f_{cave}(\mathbf{P}) + f_{vex}(\mathbf{P}) \quad (6.6)$$

where

$$\begin{aligned} f_{cave}(\mathbf{P}) \triangleq & \log_2 \left(\sum_{n=1}^N p_{n,c} |h_{n,c}|^2 + p_d |h_{d,c}|^2 + \sigma_c^2 \right) \\ & + \log_2 \left(p_d |h_d|^2 + \sum_{n=1}^N p_{n,c} |h_{n,d}|^2 + \sigma_d^2 \right), \end{aligned} \quad (6.7)$$

$$\begin{aligned} f_{vex}(\mathbf{P}) \triangleq & -\log_2 \left(p_d |h_{d,c}|^2 + \sigma_c^2 \right) \\ & -\log_2 \left(\sum_{n=1}^N p_{n,c} |h_{n,d}|^2 + \sigma_d^2 \right). \end{aligned} \quad (6.8)$$

We can learn that $f_{cave}(\mathbf{P})$ and $f_{vex}(\mathbf{P})$ are strict convex and concave functions of \mathbf{P} , respectively. So the objective function in (6.6) is a function with DC structure.

Let \mathbf{S}_R represents the set of constraints of (6.5), Therefore, \mathbf{S}_R is a convex set. The optimizing SE problem can be transformed into an equivalence problem containing the objective function with DC structure [16].

$$\max_{\mathbf{P} \in \mathbf{S}_R} \{f_{cave}(\mathbf{P}) + f_{vex}(\mathbf{P})\}. \quad (6.9)$$

In [17, 18], the author further points out that when there is a partial derivative of the convex function part in the DC objective function, the DC algorithm can be simplified to the CCCP algorithm, and its core idea is to use Majorization-Minimization (MM) method [19], stepwise iteratively linearizing the convex function part of the DC objective function.

Table 6.1 Maximum SE power allocation algorithm

Algorithm 1 Maximum SE power allocation algorithm
1: Initialize $k = 0, \forall \mathbf{P}^{(0)} \in \mathbf{S}_R, \varepsilon > 0$.
2: Do
3: $\mathbf{P}^{(k+1)} = \arg \max_{\mathbf{P} \in \mathbf{S}_R} \{f_{cave}(\mathbf{P}) + \nabla f_{vex}(\mathbf{P}^{(k)}) * \mathbf{P}^T\}$
4: Use interior point method to solve convex optimization problem in the above formula: <ul style="list-style-type: none"> a: Exploiting logarithmic barrier function transforming the original problem into an unconstrained problem, b: Use the Quasi-Newton method to obtain the search direction, c: Obtain optimal step size with Backtraking linear search based on Armijo criteria.
5: $k=k+1$.
6: Until $\ \mathbf{P}^{(k+1)} - \mathbf{P}^{(k)}\ < \varepsilon$.
7: Return $\mathbf{P}^{(k+1)}$.

Due to the convex function part of $f_{vex}(\mathbf{P})$ in (6.9) has a partial derivative. Therefore, we can linearize $f_{vex}(\mathbf{P})$ according to the first-order Taylor expansion in each iteration to get the iteration equation as below

$$\begin{aligned}
\mathbf{P}^{(k+1)} &= \arg \max_{\mathbf{P} \in \mathbf{S}_R} \{f_{cave}(\mathbf{P}) + \nabla f_{vex}(\mathbf{P}^{(k)}) * \mathbf{P}^T\} \\
&= \arg \max_{\mathbf{P} \in \mathbf{S}_R} \{f_{cave}(\mathbf{P}) + \\
&\quad \left[\frac{|h_{d,c}|^2}{(p_d^k |h_{d,c}|^2 + \sigma_c^2) \ln 2}, \frac{\sum_{n=1}^N |h_{n,d}|^2}{(\sum_{n=1}^N p_{n,c} |h_{n,d}|^2 + \sigma_d^2) \ln 2} \right] * \mathbf{P}^T \},
\end{aligned} \tag{6.10}$$

where \mathbf{P}^T is the transposition of \mathbf{P} , $\nabla f_{vex}(\mathbf{P}^{(k)})$ represents the gradient of $f_{vex}(\mathbf{P})$ at $\mathbf{P}^{(k)} \triangleq [p^{(k)}, p_d^{(k)}]$, $\mathbf{p}^{(k)} = \{p_{n,c}^{(k)}, \text{ for } n = 1, 2, \dots, N\}$.

At this time, the objective function in (6.10) is convex, which can be solved by traditional methods such as interior point method. The specific algorithm is showed in Table 6.1.

The convergence of the CCCP algorithm can be guaranteed by the following two theorems [18, 20].

Theorem 1 *The optimization objective function (6.9) increases with the power sequence $\{\mathbf{P}^k\}$ generated by the convex optimization problem in (6.10) monotonically.*

Theorem 2 *The power sequence $\{\mathbf{P}^k\}$ generated by the convex optimization problem in (6.10) converges to its limit point \mathbf{P}^∞ when $\mathbf{S}_R \neq \Phi$, and at this point, the KKT condition in the original optimization problem (6.9) is satisfied.*

6.3.2 Maximum EE Optimization

(1) Power Consumption

The total power consumption P_{total} can be decoupled into two parts: the transmit power consumption of power amplifier at antennas (RAUs and D2D transmitter) and the extral circuit power consumption. The first part can be written as [21]

$$P_{trans} = \frac{\sum_{n=1}^N P_{n,c} + P_d}{\tau}, \quad (6.11)$$

where τ is a constant, representing the drain efficiency.

The second part is denoted as $P_{circuit}$, which is consisted of three parts. (i): the circuit power consumption P_b ; (ii): the basic power consumption P_u ; (iii): the wasted power of signals transmit through optical fiber P_o . So it can be modeled as

$$P_{circuit} = P_b + P_u + P_o, \quad (6.12)$$

The total power consumed by DAS with D2D communication, denoted as P_{total} , is given by:

$$\begin{aligned} P_{total} &= P_{trans} + P_{circuit} \\ &= \frac{\sum_{n=1}^N P_{n,c} + P_d}{\tau} + P_b + P_u + P_o. \end{aligned} \quad (6.13)$$

(2) EE Problem Formulation

We focus on optimizing the power allocation to maximize the system EE. It can be expressed as (unit: bits/J/Hz) [22, 23]

$$\begin{aligned} \max \quad EE &\triangleq \frac{R_c + R_d}{\frac{\sum_{n=1}^N P_{n,c} + P_d}{\tau} + P_b + P_u + P_o} \\ \text{s.t.} \quad p_{n,c} &\in [0, P_{\max}^n] \quad \forall n \in \{1, 2, \dots, N\}, \\ p_d &\in [0, P_{\max}^d]. \end{aligned} \quad (6.14)$$

(3) Maximize EE Optimization Model

Through the above analysis, the overall energy efficiency optimization problem of the user terminals can be expressed as

$$\max_{\mathbf{V} \in \mathbf{S}} EE \triangleq \frac{R_c + R_d}{\frac{\sum_{n=1}^N P_{n,c} + P_d}{\tau} + P_b + P_u + P_o}. \quad (6.15)$$

where \mathbf{V} and \mathbf{S} represent the optimization variables and constraint sets, respectively. According to [24], (6.15) is equivalent to the following problem

$$\max_{\mathbf{V} \in \mathbf{S}} \left\{ R_c + R_d - \lambda^* \left(\frac{\sum_{n=1}^N P_{n,c} + P_d}{\tau} + P_b + P_u + P_o \right) \right\} = 0, \quad (6.16)$$

Table 6.2 Dinkelbach algorithm

Algorithm 2 Dinkelbach algorithm
1: Initialize $k = 0, \varepsilon > 0$.
2: $\lambda(0) = EE _{\mathbf{V}=\mathbf{V}(0)}, \forall \mathbf{V}^{(0)} \in \mathbf{S}$.
3: Repeat
$\mathbf{V}^{(k+1)} =$
4: $\arg \max_{\mathbf{V} \in \mathbf{S}} \{R_c + R_d - \lambda^{(k)} (\frac{\sum_{n=1}^N p_{n,c} + p_d}{\tau} + P_b + P_u + P_o)\}$.
5: $\lambda^{(k+1)} = EE _{\mathbf{V}=\mathbf{V}^{(k+1)}}$
6: $k = k + 1$
7: Until $ ee(\lambda^{(k)}) < \varepsilon$,
$ee(\lambda^{(k)}) =$
$\max_{\mathbf{V} \in \mathbf{S}} \{R_c + R_d - \lambda^{(k)} (\frac{\sum_{n=1}^N p_{n,c} + p_d}{\tau} + P_b + P_u + P_o)\}$.
8: Return $\mathbf{V}^{(k+1)}$.

where $\lambda^* = \max_{\mathbf{V} \in \mathbf{S}} EE$ represents the maximum value of the optimization goal.

For the above conclusions, the [24] has a simple and constructive proof, which will not be repeated here. In addition, the [24] also gives an iterative algorithm based on the parameter method (Dinkelbach algorithm) to find the optimal solution $\mathbf{V}^* \triangleq \arg \max_{\mathbf{V} \in \mathbf{S}} EE$ of the optimization problem in (6.15). The specific process is shown in Table 6.2.

In the Dinkelbach algorithm, the most critical step is to solve the following sub-problems for a given parameter λ

$$ee(\lambda) \triangleq \max_{\mathbf{V} \in \mathbf{S}} \{R_c + R_d - \lambda \left(\frac{\sum_{n=1}^N p_{n,c} + p_d}{\tau} + P_b + P_u + P_o \right)\}, \quad (6.17)$$

In each iteration, if the solution of Eq. (6.17) can be obtained, then the iteration can continue until the optimal solution of the optimization problem in Eq. (6.15) is obtained. The convergence of the Dinkelbach algorithm can be ensured in each iteration, $\lambda^{k+1} \geq \lambda^k$ and $ee(\lambda^{k+1}) \leq ee(\lambda^k)$ ($k = 0, 1 \dots$), and the specific proof process is in [24].

Next, we will give the solution to the sub-problem for the energy efficiency optimization problem. By the parameter transformation in the Dinkelbach algorithm, the NFP optimization problem in (6.15) can be expressed as the following subproblem

$$\begin{aligned} & \max_{p_{n,c}, p_d} \left\{ R_c + R_d - \lambda \left(\frac{\sum_{n=1}^N p_{n,c} + p_d}{\tau} + P_b + P_u + P_o \right) \right\}, \\ & s.t. \quad p_{n,c} \in [0, P_{\max}^n] \quad \forall n \in \{1, 2, \dots, N\}, \\ & \quad \quad p_d \in [0, P_{\max}^d]. \end{aligned} \quad (6.18)$$

According to the discussion of the D.C. optimization problem in the previous section, the above problems can be expressed as

$$\begin{aligned} & \max_{p_{n,c}, p_d} \{f_{cave}(\mathbf{Q}) + f_{vex}(\mathbf{Q})\}, \\ & s.t. \quad p_{n,c} \in [0, P_{\max}^n] \quad \forall n \in \{1, 2, \dots, N\}, \\ & \quad \quad p_d \in [0, P_{\max}^d]. \end{aligned} \quad (6.19)$$

where $\mathbf{Q} = [p_{n,c}, p_d]$ represents the optimization variable,

$$\begin{aligned} f_{cave}(\mathbf{Q}) \triangleq & \log_2(\sum_{n=1}^N p_{n,c} |h_{n,c}|^2 + p_d |h_{d,c}|^2 + \sigma_c^2) \\ & + \log_2(p_d |h_d|^2 + \sum_{n=1}^N p_{n,c} |h_{n,d}|^2 + \sigma_d^2). \end{aligned} \quad (6.20)$$

$$\begin{aligned} f_{vex}(\mathbf{Q}) \triangleq & -\log_2(p_d |h_{d,c}|^2 + \sigma_c^2) \\ & -\log_2(\sum_{n=1}^N p_{n,c} |h_{n,d}|^2 + \sigma_d^2) \\ & -\lambda(\frac{\sum_{n=1}^N p_{n,c} + p_d}{\tau} + P_b + P_u + P_o). \end{aligned} \quad (6.21)$$

where $f_{cave}(\mathbf{Q})$ and $f_{vex}(\mathbf{Q})$ represent concave function part and convex function part of the objective function, respectively. \mathbf{S}_T is the set of constraints in (6.18). Since all constraints are linear inequalities, \mathbf{S}_T is a convex set. In addition, the convex function in (6.18) has a partial derivative, so the above D.C. problem can be transformed into the following sequential convex program problem by the CCCP algorithm

$$\begin{aligned} \mathbf{Q}^{(k+1)} &= \arg \max_{\mathbf{Q} \in \mathbf{S}_T} \{f_{cave}(\mathbf{Q}) + \nabla f_{vex}(\mathbf{Q}^{(k)}) * \mathbf{Q}^T\} \\ &= \arg \max_{\mathbf{Q} \in \mathbf{S}_T} \{f_{cave}(\mathbf{Q}) \\ & \quad + [\frac{|h_{d,c}|^2}{(p_d^{(k)} |h_{d,c}|^2 + \sigma_c^2) \ln 2} - \lambda, \frac{\sum_{n=1}^N |h_{n,d}|^2}{(\sum_{n=1}^N p_{n,c}^{(k)} |h_{n,d}|^2 + \sigma_d^2) \ln 2} - \lambda] * \mathbf{Q}^T\}. \end{aligned} \quad (6.22)$$

where \mathbf{Q}^T is the transposition of \mathbf{Q} , $\nabla f_{vex}(\mathbf{Q}^{(k)})$ represents the gradient of $f_{vex}(\mathbf{Q})$ at $\mathbf{Q}^{(k)} \triangleq [p^{(k)}, p_d^{(k)}]$, $\mathbf{p}^{(k)} = \{p_{n,c}^{(k)}, \text{ for } n = 1, 2, \dots, N\}$.

Because the objective function in equations (6.18) is a concave function. So we can exploit the traditional methods to obtain the optimal solutions. After transformation, the optimizing energy efficiency problem in (6.15) can be solved by a three-layer nested loop algorithm, which is concluded in Table 6.3.

6.4 Numerical Results

In the simulations, to simplify the computational complexity, we only consider a single-cell DAS with one UE and one D2D pair in the downlink transmission, both of which are uniformly located in the cell. The parameters values are showed in Table 6.4. The system is set as a circle of radius D . The layout of the RAUs is similar to [25].

Table 6.3 Maximum energy-efficient power allocation algorithm**Algorithm 2** Maximum energy efficient power allocation algorithm

-
- 1: **Initialization** $k = 0, \varepsilon > 0$
 - 2: $\lambda^{(0)} = EE|_{\mathbf{Q}=\mathbf{Q}^{(0)}}, \forall \mathbf{Q}^{(0)} \in \mathbf{S}_T$.
 - 3: **Do**
 - 4: $\mathbf{Q}^{(k+1)} =$
 $\arg \max_{\mathbf{Q} \in \mathbf{S}_T} \{R_c + R_d - \lambda^{(k)} (\frac{\sum_{n=1}^N p_{n,c} + p_d}{\tau} + P_b + P_u + P_o)\}$
 - 5: Use CCCP algorithm to solve DC optimization problem (19)
 - 6: Use interior point method to solve the problem:
 - a: Exploiting logarithmic barrier functions to transform the original problem into an unconstrained optimization problem,
 - b: Use the Quasi-Newton method to obtain the search direction,
 - c: Obtain the optimal step size with Backtraking linear search based on Armijo criteria.
 - 7: $\lambda^{(k+1)} = EE|_{R_D=\mathbf{Q}^{(k+1)}}$
 - 8: $k = k + 1$
 - 9: **Until** $|ee(\lambda^{(k)})| < \varepsilon$
 $ee(\lambda^{(k)}) =$
 $\max_{\mathbf{V} \in \mathbf{S}} \{R_c + R_d - \lambda^{(k)} (\frac{\sum_{n=1}^N p_{n,c} + p_d}{\tau} + P_b + P_u + P_o)\}$.
 - 10: **Return** $\mathbf{Q}^{(k+1)}$.
-

Table 6.4 Simulation parameters.

Parameters	Value
The cellular radius D	1000 m
The D2D distance L	20 m
The UE number M	1
The D2D pairs number K	1
The noise power σ_c^2	-114 dBm
The noise power σ_d^2	-114 dBm
The maximum transmit power of UE P_{\max}^c	30 dBm
The maximum transmit power of D2D P_{\max}^d	30 dBm
The circuit power consumption P_d	20 dBm
The basic power consumption P_u	30 dBm
The optical fiber transmission P_o	30 dBm
Path loss exponent α	3.8
Drain efficiency τ	38%

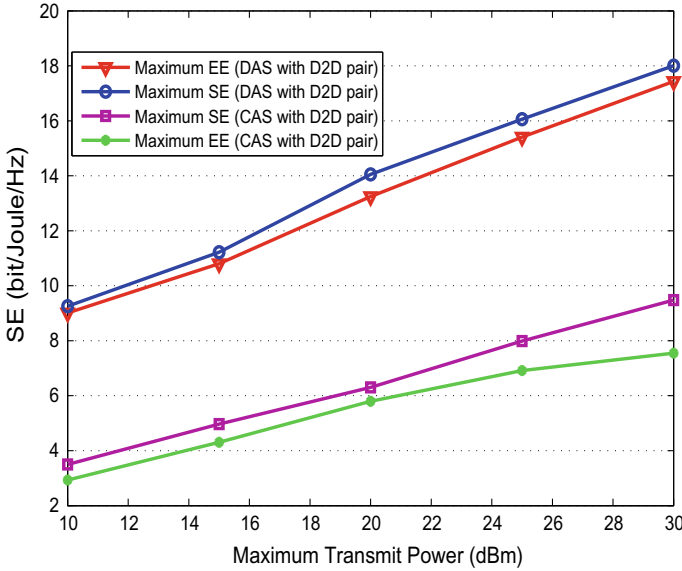


Fig. 6.2 SE versus maximum transmit power

In Fig. 6.2, P_{\max}^c changes from 5 to 30 dBm to show its effects on SE of the system. It shows that the SE increases with the increase of P_{\max}^c , and the performance of the power allocation methods used in DAS with D2D pair is better than used CAS with D2D pair in [13]. We also compare with two different optimization objectives of maximizing SE and EE. From Fig. 6.2, for maximizing SE in DAS with D2D communication, maximizing SE power allocation algorithm is better than the algorithm used to maximize EE. Compared to CAS with D2D communication, the maximum SE in DAS is approximately 89.9% higher than maximum SE in CAS when $P_{\max}^c = P_{\max}^d = 30$ dBm.

In Fig. 6.3, maximizing SE and maximizing EE algorithms are both used in increasing the EE of DAS with D2D communication. In this case, the maximizing EE algorithm is much better than the algorithm of maximizing SE power allocation. We also show the impact on the overall system performance after introducing DAS. Obviously, compared to CAS with D2D communication in [13], the EE has improved significantly in DAS with D2D communication. The EE of maximum EE in DAS is approximately 408.9% higher than maximum EE in CAS when $P_{\max}^c = P_{\max}^d = 30$ dBm.

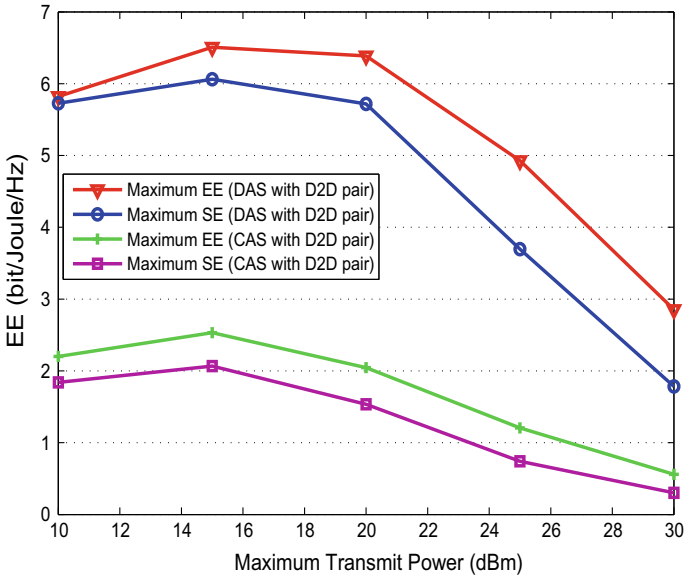


Fig. 6.3 EE versus maximum transmit power

6.5 Conclusion

We considered a coexistence scenario of DAS and D2D communication in this paper. CSI is assumed known at both receiver and transmitter side. We first presented a optimization problem with respect to the maximizing SE power allocation, and the original problem was transformed into a DC structure problem by function recombination. Then the CCCP process was exploited to solve the DC structure problem, in which the interior point method was used to get the optimal power allocation solution. Then maximizing EE of the system also considered in the following part. We proposed an algorithm to maximize EE by Dinkelbach algorithm based on parameter method. Simulation results indicated that the performance of the power allocation methods used in DAS with D2D user was better than used in CAS with D2D communication.

Acknowledgements This work was supported in part by the Natural Science Funding of Guangdong Province under Grant 2017A030313336, in part by the Shenzhen Overseas High-level Talents Innovation and Entrepreneurship under Grant KQJSCX201803280- 93835762, and in part by the Tencent "Rhinoceros Birds" Scientific Research Foundation for Young Teachers of Shenzhen University.

References

1. Heath, R., Peters, S., Wang, Y.: A current perspective on distributed antenna systems for the downlink of cellular systems. *IEEE Commun. Mag.* **51**(4), 161–167 (2013)
2. Park, E., Lee, S.R., Lee, I.: A current perspective on distributed antenna systems for the downlink of cellular systems. *IEEE Trans. Wireless Commun.* **11**(7), 2468–2477 (2012)
3. Yu, X., Tan, W., Wu, B., Li, Y.: Discrete-rate adaptive modulation with variable threshold for distributed antenna system in the presence of imperfect CSI. *China Commun.* **11**(13), 31–39 (2014)
4. Corson, M.S., Laroia, R., Li, J., Park, V., Richardson, T., Tsirtsis, G.: Toward proximity-aware internetworking. *IEEE Wireless Commun.* **17**, 6 (2010)
5. Lee, N., Lin, X., Andrews, J.G., Heath, R.W.: Power control for D2D underlaid cellular networks: modeling, algorithms, and analysis. *IEEE J. Sel. Areas Commun.* **33**(1), 1–13 (2015)
6. Chen, X., Xu, X., Tao, X.: Energy efficient power allocation in generalized distributed antenna system. *IEEE Commun. Lett.* **16**(7), 1022–1025 (2012)
7. He, C., Li, G.Y., Zheng, F., You, X.: Energy-efficient resource allocation in OFDM systems with distributed antennas. *IEEE Trans. Veh. Technol.* **63**(3), 1223–1231 (2014)
8. Kim, H., Lee, S., Song, C., Lee, K., Lee, I.: Optimal power allocation scheme for energy efficiency maximization in distributed antenna systems. *IEEE Trans. Commun.* **63**(2), 431–440 (2015)
9. He, C., Sheng, B., Zhu, P.: Energy-and spectral-efficiency tradeoff for distributed antenna systems with proportional fairness. *IEEE J. Sel. Areas Commun.* **31**(5), 894–902 (2013)
10. Yu, C., Doppler, K., Ribeiro, C., Tirkkonen, O.: Resource sharing optimization for device-to-device communication underlying cellular networks. *IEEE Trans. Wireless Commun.* **10**(8), 2752–2763 (2011)
11. Feng, D., Lu, L., Yuan, Y.: Device-to-device communications un-derlaying cellular networks. *IEEE Trans. Commun.* **61**(8), 3541–3551 (2013)
12. Wang, J., Zhu, D., Zhao, C., Li, J.C., Lei, M.: Resource sharing of underlying device-to-device and uplink cellular communications. *IEEE Commun. Lett.* **17**(6), 1148–1151 (2013)
13. Feng, D., Yu, G., Yuan, W.Y., Li, G.Y., Feng, G., Li, G.S.: Mode switching for device-to-device communications in cellular. *IEEE Signal Inf. Process* (2015)
14. You, X., Wang, D., Zhu, P., Sheng, B.: Cell edge performance of cellular mobile systems. *IEEE J. Sel. Areas Commun.* **29**(6), 1139–1150 (2011)
15. Wang, D., Wang, J., You, X., Wang, Y., Chen, M., Hou, X.: Spectral efficiency of distributed MIMO systems. *IEEE J. Sel. Areas Commun.* **31**(10), 2112–2127 (2013)
16. An, L.T.H., Tao, P.D.: The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Ann. Oper. Res.* **133**(1–4), 23C46 (2005)
17. Yuille, A.L., Rangarajan, A.: The concave-convex procedure (CC-CP). In: *Proceedings of Advances in Neural Information Processing Systems*, pp. 1033–1040, (2001)
18. Lanckriet, G.R., Sriperumbudur, B.K.: On the convergence of the concave-convex procedure. In: *Proceedings of Advances in Neural Information Processing Systems*, pp. 1759–1767 (2009)
19. Hunter, D.R., Lange, K.: A tutorial on mm algorithms. *Am. Statistician* **58**(1), 30–37 (2004)
20. Boyd, S., Vandenberg, L.: *Convex optimization*. Cambridge University Press (2004)
21. Arnold, O., Richter, F., Fettweis, G., Blume, O.: Power consumption modeling of different base station types in heterogeneous cellular networks. In: *Future Network and Mobile Summit*, pp. 1–8 (2010)
22. Miao, G., Himayat, N., Li, G.: Energy-efficient link adaptation in frequency-selective channels. *IEEE Trans. Commun.* **58**(2), 545–554 (2010)
23. Rui, Y., Zhang, Q., Deng, L., Cheng, P., Li, M.: Mode selection and power optimization for energy efficiency in uplink virtual mimo systems. *IEEE J. Sel. Areas Commun.* **31**(5), 926–936 (2013)
24. Dinkelbach, W.: On nonlinear fractional programming. *Manag. Sci.* **13**(7), 492–498 (1967)
25. He, C.: Comparison of three different optimization objectives for distributed antenna systems. *AEU-Int J Electron Commun* **70**(4), 442–448 (2016)

Chapter 7

Analysis of Transmission Efficiency of Magnetically Coupled Resonant Radio Energy



Xiaohu Yin and Yue Zhao

Abstract Magnetically coupled with resonant radio energy transmission technology can realize medium-distance power transmission between power supply equipment and electrical equipment. Working principle of resonant coupled radio energy transmission is introduced in this paper, and bilateral compensation method is used to analyze the two factors of large leakage inductance and low coupling coefficient, which affect the transmission efficiency of the system. In practical applications, to achieve maximum efficiency transmission of the system, it is necessary to determine the appropriate circuit operation mode, resonant frequency and coil parameters. According to the four different compensation topologies, the influence of the secondary quality factor and transmission distance of the wireless energy transmission system on the transmission efficiency of the system is analyzed in detail based on MATLAB simulation.

7.1 Introduction

In the recent years, wireless energy transmission technology has always been a hot issue in human research. Inductively coupled power transfer (ICPT) is a new type of power access method that achieves non-contact transmission of electrical energy in a small scale [1, 2], but the transmission distance is still limited to a small within the scale, transmission of 15 cm has been considered a considerable distance [3, 4]. In 2007, Professor Marin Soljagic of the Massachusetts Institute of Technology (MIT) and others made breakthroughs in radio energy transmission. They used the principle of electromagnetic resonance to realize the wireless transmission of medium-distance power, and a 60 W in 2 m distance. The bulb is lit and the transmission efficiency is around 40% [5].

Resonant coupled radio energy transmission utilizes electromagnetic near-field resonance coupling to efficiently transfer energy from one resonator to another in the form of a ‘tunnel’ without energy exchange with or little with non-resonant

X. Yin · Y. Zhao (✉)
Xi'an University of Science and Technology, Xi'an, Shanxi, China
e-mail: 1950247346@qq.com

objects. In theory, the energy that is not absorbed by the load will return to the transmitting end, so that it will not affect the efficiency [6]. The wireless energy transmission technology based on magnetic coupling resonance has the advantages of high transmission efficiency, large transmission distance, small radiation, non-magnetic objects and little harm to the human body [7], and it will certainly be a kind safe and reliable charging method.

However, at this stage, the resonant coupled power wireless transmission technology is still in its infancy, and many problems need to be solved, transmission efficiency, transmission power, transmission distance, electromagnetic compatibility, electromagnetic radiation pollution and biosafety issues. Especially, in the aspect of transmission efficiency analysis, MIT's analysis is only limited to physical analysis, and related theoretical and experimental research is still lacking [8, 9].

In this paper, the mathematical modeling and simulation analysis of the circuit topology of PSSS, PSSP, PPSS and PPSP are carried out, and the transmission characteristics of the resonant compensation mode are compared.

7.2 Resonant Coupled Power Wireless Transmissions Works

Magnetically coupled resonant radio energy transmission technology is a near-field strongly coupled radio energy transmission technology. The core is to use two inductive coils with the same operating frequency to achieve efficient power transmission when resonance occurs, and still at medium distance. It can carry out higher efficiency and higher power transmission of electric energy [10].

The resonant coupled power wireless transmission technology mainly utilizes two resonant circuits having the same frequency to realize wireless transmission of energy from the stationary power supply system to the power supply device through magnetic field coupling.

7.3 Circuit Design of Electromagnetic Coupling Resonant Radio Energy Transmission System

7.3.1 The Importance of Resonance Compensation

The wireless energy transmission system achieves non-contact transmission of energy through the resonant coupling of the separable transformer. The separable transformer of the wireless energy transmission system is a loosely coupled transformer. The gap between the primary and secondary is relatively large, resulting in a large magnetic leakage, which greatly reduces the transmission performance of the wireless transmission system. In order to reduce the voltage and current stress of

the transformer [11], the unnecessary power in the power transmission network is effectively compensated, so the resonance compensation technique is used.

In order to improve the efficiency of the system, at the beginning of the wireless energy transmission system, the secondary circuit is added with an appropriate equal amount of compensation capacitance and the leakage inductance of the primary and secondary coils to form a resonance compensation circuit [12]. According to the different topological structure of the resonant circuit, there are eight kinds of compensation methods. The primary coil and the secondary coil are denoted by P and S , respectively; the series and the parallel are, respectively, denoted by S and P .

According to the theoretical analysis, the single-side resonance compensation topologies PSS_, P_SS, PPS_, P_SS can only improve the working efficiency of the primary loop and the secondary loop unilaterally. In order to increase the efficiency of the entire radio energy transmission system, a bilateral compensation topology is required. According to the above analysis, the bilateral compensation topology can be divided into the following four types: PSSS, PSSP, PPSS and PPSP.

7.3.2 Circuit Design of the Resonant Coupling Part

The diagram of the system's structural frame is shown in Fig. 7.1.

Among them, C_1 and C_2 are series or parallel compensation capacitors of primary and secondary inductors L_1 and L_2 , respectively, and M is mutual inductance. When using bilateral compensation, there are two quality factors: primary and secondary quality factor. For the convenience of calculation, the internal resistance of the power

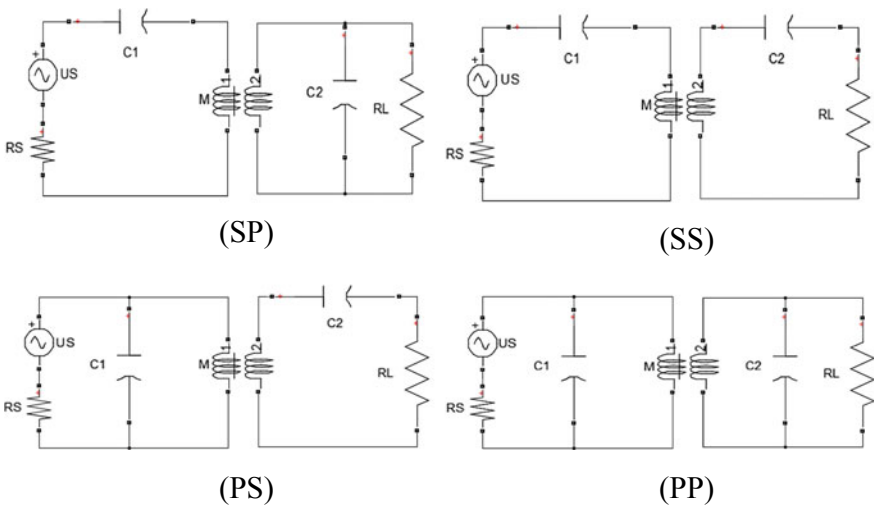


Fig. 7.1 Resonant coupling circuit design

supply. Quality factor as Formula (7.1):

$$Q = \frac{wL}{r} = \frac{1}{wCr} = \frac{1}{r} \sqrt{\frac{L}{C}} \quad (7.1)$$

7.3.3 Experimental Analysis

In practical applications, the wireless energy transmission system needs to provide energy to different loads, and the change in load causes a change in the secondary quality factor. Based on the MATLAB under the compensation topology, the influence of the secondary quality factor and the transmission distance of the two coils on the transmission efficiency are analyzed.

For the secondary loop, the secondary loop transmission capability is proportional to its quality factor. However, the improvement of the quality factor will also lead to an increase in the volt-ampere characteristics of the secondary circuit. Therefore, it is a very important parameter in the entire wireless energy transmission system. In practical applications, the value is generally between 2 and 10, and the primary circuit. The geometrical characteristics and primary current determine the value of the primary quality factor. The primary quality factor is affected by different design schemes, and the range of values is generally between 2 and 50.

- A. Analysis of the influence of secondary quality factor and transmission distance on voltage gain.

The voltage gain is the ratio of the output to the input voltage. The expression for the voltage gain of the wireless energy transmission system is derived below. Taking the PSSS resonance compensation topology as an example, the voltage gain expression is derived:

The ratio of the operating frequency to the magnitude of the resonant frequency can be expressed as Formula (7.2):

$$\omega_n = \frac{\omega}{\omega_0} \quad (7.2)$$

The coupling coefficient k is expressed as Formula (7.3):

$$k = \frac{M}{\sqrt{L_P L_S}}, \quad M = \frac{\mu_0 \pi N_1 N_2 r_1 r_2}{2D^3} \quad \mu_0 = 4\pi * (10^{-7}) \text{ N/A}^2 \quad (7.3)$$

The magnetic permeability in the vacuum is the radius of the coil, and the distance between the coils is the number of turns of the coil.

The reflected impedance when the secondary series is (4)–(6):

$$Z_r = \text{Re} - j\text{Im} \quad (7.4)$$

$$\text{Re} = \frac{\omega^4 C_s^2 M^2 R_L}{(1 - \omega^2 C_s L_s)^2 + \omega^2 C_s^2 R_L^2} \quad (7.5)$$

$$\text{Im} = \frac{\omega^3 C_s M^2 (1 - \omega^2 C_s L_s)}{(1 - \omega^2 C_s L_s)^2 + \omega^2 C_s^2 R_L^2} \quad (7.6)$$

The voltage gain of PSSS is (7):

$$G_v = \frac{nk\omega_n^2}{\sqrt{\omega_n^2(1 - \omega_n^2)^2 + (Q_s(1 - \omega_n^2)^2 - k^2 Q_s \omega_n^4)^2}} \quad (7.7)$$

It can be seen from the above formula that the voltage gain is related to n , w , D , r_1 , r_2 and the like. Similarly, Cocoa derives the voltage gain expressions for the other three compensation topologies.

B. Influence of secondary quality factor and transmission distance D on current gain.

The ratio of the output current to the input current is called the current gain. The expression of the current gain of the wireless power transmission system is derived based on the PSSS compensation structure as Formula (7.8):

$$i_2 = \frac{j\omega M i_1}{j\omega L_2 + \frac{1}{j\omega C_2} + R_L} \quad (7.8)$$

Therefore, the current gain in the PSSS compensation circuit is expressed in Formula (7.9):

$$G_i = \frac{n\omega_n^2 Q_s k}{\sqrt{Q_s^2 (1 - \omega_n^2)^2 + \omega_n^2}} \quad (7.9)$$

Similarly, the current gain expressions of the other three compensation topologies can be derived.

C. Influence of secondary quality factor and transmission distance D on system efficiency.

The purpose of introducing resonance compensation into the radio energy transmission system is to improve transmission efficiency. Transmission efficiency can be expressed as Formula (7.10):

$$\eta = \frac{P_o}{P_i} = \left| \frac{v_o i_o}{v_i i_i} \right| = G_v G_i \quad (7.10)$$

The transmission efficiency is equal to the ratio of the output power to the input power, that is, the product of the voltage gain and the current gain, as in the above equation.

$$N1 = N2 = 10, \quad r1 = r2 = 0.01,$$

$$\mu = 4\pi \times 10^{-7} L1 = L2 = 3.94 * 10^{-6}, \quad n = 1$$

From this, an expression about the relationship between efficiency and the secondary quality factor and the coupling coefficient can be obtained. Here, it represents the output voltage and current of the system, and represents the input voltage and current of the system. Substituting the above voltage gain and current gain expressions into the above equation yields an expression about the relationship between efficiency and secondary quality factor, and transmission distance. In the simulation below, let, so;

String-string compensation loop:

- ① Secondary quality factor affects transmission efficiency ($D = 0.05$ m).

The simulation diagram is shown in Fig. 7.2.

Can be seen from Fig. 7.2. When the system operating frequency is equal to the resonant frequency, different quality factors can achieve the highest efficiency. However, when $D = 0.5$ and 1, the curve has only one peak; when it is equal to 4, 8 and 10, the curve shows three peaks, indicating that when it becomes larger, it will cause frequency bifurcation.

- ② Transmission distance D affects transmission efficiency:

The simulation diagram is shown in Fig. 7.3.

Fig. 7.2 String-string compensation and relationship curve

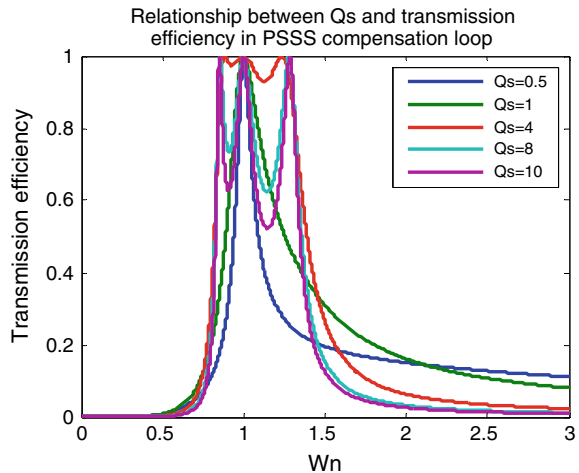
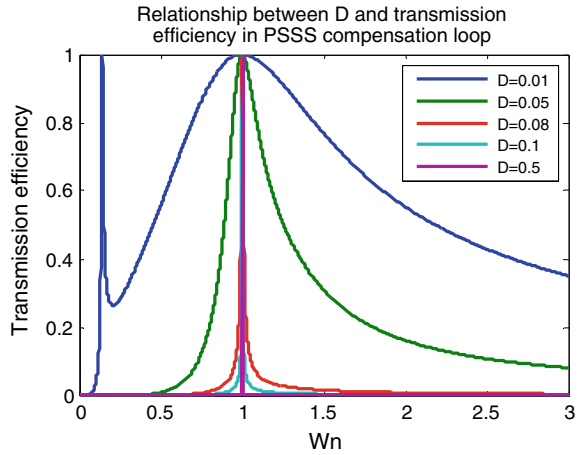


Fig. 7.3 String-string compensation D and the relationship curve



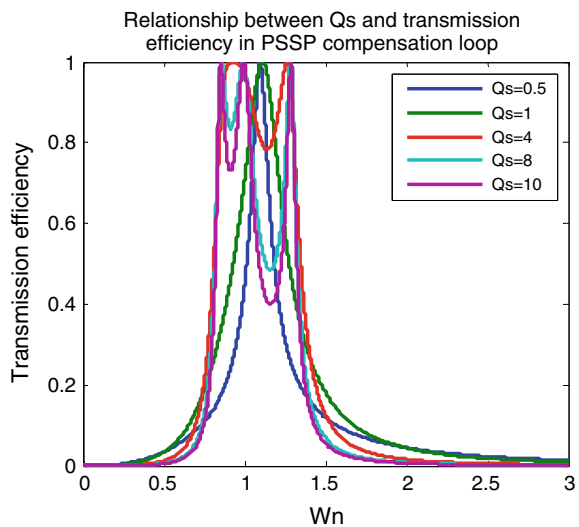
When the serial string compensation topology is used, when the secondary quality factor is determined, the curve around the apex is steepest when the distance is transmitted. As D decreases, the curve gradually becomes gentle, indicating that when D is small, the wireless energy transmission system can achieve higher efficiency and a wider range.

String and compensation loop

- ① Secondary quality factor affects transmission efficiency:

The simulation diagram is shown in Fig. 7.4.

Fig. 7.4 String and compensation curve



As the number continues to increase, the frequency bifurcation phenomenon becomes more apparent.

② Transmission distance D affects transmission efficiency:

The simulation diagram is shown in Fig. 7.5.

As the transmission distance D becomes smaller, frequency bifurcation is prone to occur.

Parallel compensation loop

① Secondary quality factor affects transmission efficiency:

The simulation diagram is shown in Fig. 7.6.

Fig. 7.5 Series and compensation for the relationship between D and

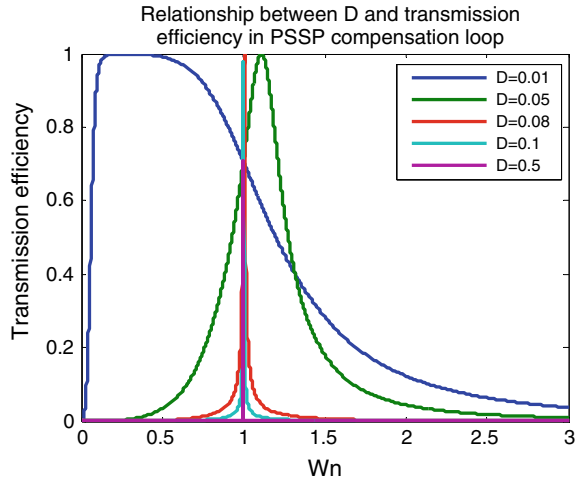


Fig. 7.6 Parallel compensation and relationship curve

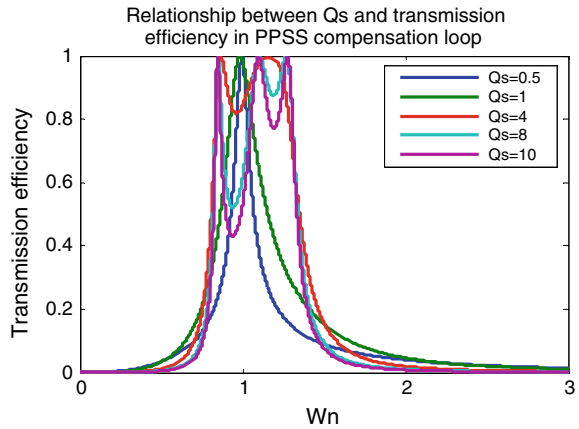
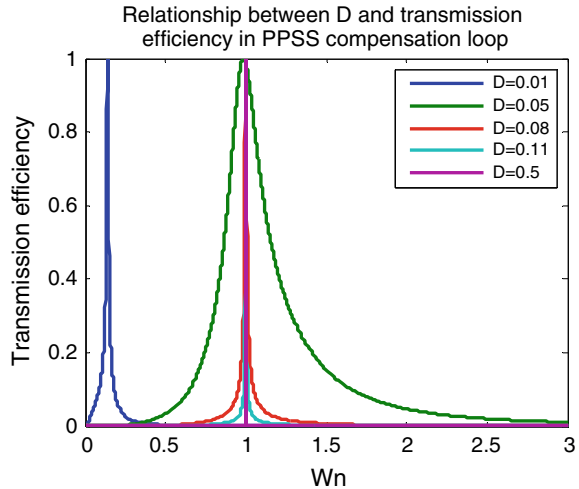


Fig. 7.7 Parallel string compensation D and the relationship curve



As the number continues to increase, the frequency bifurcation phenomenon becomes more apparent.

② Transmission distance D affects transmission efficiency:

The simulation diagram is shown in Fig. 7.7.

When the parallel-compensated topology is used, the frequency bifurcation phenomenon is not easy to occur.

And compensated in the loop

① Secondary quality factor affects transmission efficiency:

The simulation diagram is shown in Fig. 7.8.

When the structure is compensated, as the frequency increases, the frequency bifurcation phenomenon becomes more and more obvious.

② Transmission distance D affects transmission efficiency:

The simulation diagram is shown in Fig. 7.9.

When the sum is used and the structure is compensated, the distance is above 0.05 m, and the transmission efficiency reaches a maximum value under resonance. Regardless of the compensation topology used, an increase in the secondary quality factor leads to the occurrence of frequency bifurcation. When the secondary quality factor is determined to be 1, when the transmission distance $D = 0.5$, its curve around the apex is the steepest. As the transmission distance becomes smaller, the curve at the apex becomes more gradual, indicating that the transmission distance D is longer. When small, the radio energy transmission system can achieve a higher efficiency frequency range.

Fig. 7.8 and the relationship between compensation and

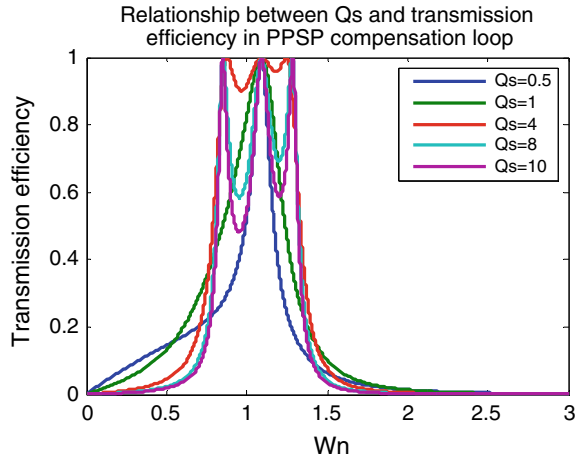
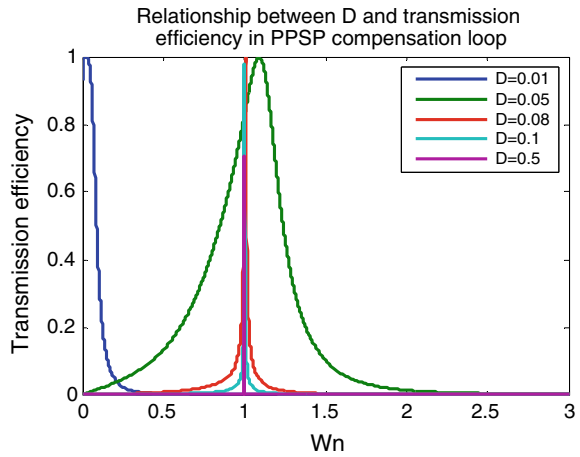


Fig. 7.9 and the relationship between compensation and



The optimum input frequency depends on the relative distance between the two coils. When the input frequency is equal to the natural frequency (resonant frequency), the transmission efficiency is maximized when the transmission distance is 0.05. The simulation analysis is basically consistent with the variation of the theoretical part.

7.4 Conclusion

Bilateral resonance compensation can effectively compensate for the shortcomings of low transmission efficiency of the system. In this paper, the bilateral resonant

compensation wireless power supply system is taken as the research object, and the influence of transmission distance and secondary quality factor on transmission efficiency under resonance is derived. The simulation proves that when the secondary quality factor is greater than 1, the frequency bifurcation phenomenon will occur. The larger the secondary quality factor, the more obvious the frequency bifurcation phenomenon; when the transmission distance is too small (near-distance transmission—the distance that the radio energy can transmit is smaller than the device) (No more than two times the size), no resonance compensation circuit is needed at this time. When the transmission distance is increasing, the frequency range in which the wireless power transmission system can achieve less efficiency is relatively narrow.

References

1. Yue, S., Chenyang, X., Xin, D., et al.: Analysis and optimization of mutual inductance coupling parameters of inductively coupled power transmission system. *Proc. CSEE* **30**(33), 44–50 (2010)
2. Xin, D., Yue, S.: A new power supply mode and related technology analysis of monorail driving. *J. Chongqing Univ.* **26**(1), 50–53 (2003)
3. Wang, G., Liu, W., Sivaprakasam, M., et al.: A dual band wireless power and data telemetry for retinal prosthesis. In: *Proceedings of the 28th IEEE EMBS Annual International Conference* (2006)
4. Sallan, J., Villa, J.L., Llombart, A., et al.: Optimal design of ICPT system applied to electric vehicle battery charge. *IEEE Trans. Ind. Electron.* **56**(6):2140–2149 (2009)
5. Andre, K., Aristeidis, K., Robert, M., et al.: Wireless power transfer via strongly coupled magnetic resonances. *Science* **317**(6), 83–84 (2007)
6. Kurs, A., Karalis, A., Moffatt, R., et al.: Wireless power transfer via strongly coupled magnetic resonances. *Science* **317**(5834), 83–86 (2007)
7. Hamam, R. E., Karalis, A., Joannoponlos, J.D., et al.: Coupled-mode theory for general free-space resonant scattering of waves. *Phys. Rev. A* **75**(5):1–5 (2007)
8. Karalis, A., Joannoponlos, J.D., Soljacic, M.: Efficient wireless non-radiative mid-range energy transfer. *Ann. Phys.* **323**(1), 34–48 (2008)
9. Fu, W., Zhang, B., Qiu, D., et al.: Maximum efficiency analysis and design of self-resonant coil coupled power wireless transmission. In: *Proceedings of the Chinese Society of Electrical Engineering*, vol. 29(18), pp. 21–26 (2009)
10. Xia, C.: Analysis and optimization of energy efficiency characteristics of inductively coupled power transmission system. Chongqing: Chongqing University 11 (2010)
11. Zhao, K., Chen, X.: New concept physics course. *Electromagnetics*, pp. 100–200. Higher Education Press, Beijing (2003)
12. Stielau, O.H., Covic, G.A.: Design of loosely coupled inductive power transfer system. In: *International Conference on Power System Technology, 2000. Proceeding. PowerCon2000*, pp. 85–90 (2000)

Chapter 8

Remote Monitoring of River Water Pollution Using Multiple Sensor System of WSNs and IoT



Evizal Abdul Kadir, Hitoshi Irie, Sri Listia Rosa, Bahruddin Saad, Sharul Kamal Abdul Rahim, and Mahmud Othman

Abstract Rivers are commonly found in the tropical regions because of high rain intensity. Many people and communities like to live along the riverside since decades ago. Rivers play a significant role in communities for transportation and daily activities. This research was aimed to design and develop a system with multiple sensors to monitor river water pollution because most of the community members use river water in their daily activities. In this design and development of system, wireless sensor networks (WSNs) were applied because of many advantages that can be enjoyed. Multiple sensor nodes were installed for the detection of water pollution parameters such as temperature, electrical conductivity (EC), water pH, and dissolved oxygen (DO). The system was designed to monitor river water pollution parameters and send the information to the data center (backend system). Arduino microcontroller was used to process and filter the data before sending to the backend system. Only valuable information was collected and kept in the database. Results show that the system was able to detect polluted water by showing the parameters of interest in a graph. Based on the analysis, it can be concluded that polluted water indicators were mostly contributed from residence waste and industries. Furthermore, WSN sensors will be deployed in some areas, and the results will be compared to each other. Furthermore, the Internet of things (IoT) technology will be used for data sharing and communication.

E. A. Kadir (✉) · S. L. Rosa

Department of Informatics Engineering, Universitas Islam Riau, Jl. Kaharuddin Nasution, Pekanbaru, Riau 28284, Indonesia

e-mail: evizal@eng.uir.ac.id

H. Irie

Center for Environmental Remote Sensing (CEReS), Chiba University, 1-33 Yayoicho, Inage Ward, Chiba 263-8522, Japan

B. Saad · M. Othman

Fundamental and Applied Sciences Department, Universiti Teknologi Petronas, Seri Iskandar, Perak 32610, Malaysia

S. K. A. Rahim

Wireless Communication Centre, Universiti Teknologi Malaysia, Johor Bahru, Johor 81310, Malaysia

8.1 Introduction

Water pollution is one of the issues that has been raised in some of the areas in Indonesia. This research focussed on Siak River located in Riau Province because of the issue of river pollution since a few years and until now no significant solution by the respective authorities had been taken. Riau Province is located in central Sumatera Island in Indonesia. This province has five long and deep rivers, and one of the rivers is the deepest in Indonesia. Along the Siak River, many companies are operating, the big company being pulp and paper beside other small companies. Sometimes, the companies emit pollution to the river, and this contaminates the river. Contamination of the river water may be contributed by various sources such as industrial waste, chemical spill, and community and residence waste. Flooding and other disaster might also pollute the river. This research aims to introduce a monitoring system which incorporates four basic sensing systems which include temperature, dissolved oxygen (DO), water pH, and electrical conductivity.

The conventional techniques to measure the quality of water using several methods had been discussed [1–4]. The methods collect river water and information regarding the water quality conducted in the laboratory including biological, chemical, and physical parameters of the water. It is time consuming and expensive to get full information along the river as many samples need to be sampled. Real-time water quality monitoring system using WSNs is popular in recent years because of the advantages of technology to collect data and information through the sensor node. The requirement for continuous data monitoring for water quality in a real-time system to establish trend and predict behavior from the history is discussed in [5–9].

A biological sensing system for water quality detection is a method to detect bacteria using a computer vision system in analysis, and chemical analysis of the pollutant is discussed [10–12]. The use of multi-sensors for the water pollutant detection system for the basic parameters is not limited to a few parameters. The use of locals for the collection of the data to support the monitoring system that has limited access is elaborated [13–16]. The image of river water capture by camera as a sample then send to the backend station for analysis, then compare to analysis used sensors. In the previous research, the maximum distance was 10 m for the image analysis [17, 18]. In deep waters or river, the use of robotic system for water quality detection has several advantages. Another scenario is in the ocean where mobile communication can assist in the detection of water quality. In others, research conducted uses the method for polluted water detection but ineffective way to the permanent solution system; furthermore, the case of the area is discussed [19–22].

In this research, multiple sensors are proposed to study water quality or polluted water. Multiple sensors are able to achieve better analysis of the samples compared to the detection using a single sensor. The basic parameters proposed are temperature, water pH, electrical conductivity, and DO. Output gained from this approach will contribute to valuable knowledge for the monitoring of water quality in rivers. The transfer of data between WSN nodes and a node sink for smooth data communication

is also a part of the objective in this work. Siak River, being one of the longest and deepest rivers in Riau Province, Indonesia, was chosen for the studies.

8.2 Multiple Sensor System of WSNs

The proposed design of multiple sensor systems for water quality monitoring using WSNs was based on a case study at Siak River in Riau Province, Indonesia. The river is very long (more than 200 km). Most of the community and rural residents live along the riverside and do their daily activities using river water. A preliminary survey on the river and geographical information was used to design the sensing system for the detection of pollutants in the river. Figure 8.1 shows the geographical location of Siak River in Riau Province, Indonesia. The river water was polluted from housing wastes as well as unhealthy community practices. Furthermore, in the rainy season the situation becomes worst because of the flooding and all the rubbish and wastes find their way into the river through the canals.

Based on an early survey, the river is highly polluted because the river is not only used for the residents living along the riverside but also for transportation where many vessels and wooden ship ply through the river carrying various materials including



Fig. 8.1 Geographical location of Siak River and the testing location

people using high-speed boats. Furthermore, many companies operate along the river because of easy transportation and water supply. Some of the companies spilled chemical wastes and other materials into the river. Figure 8.2 shows a scene of the polluted river water as indicated by the colored water sensors required for the detection of water contaminant in the river. This information will be used to decide the type of sensors that need to be installed to the system for detection such as temperature, DO, pH, and electrical conductivity as well as other parameters for the future.

Multiple sensor system was designed to incorporate four parameters for the establishment of river water pollutant index, and results of all the sensors provide information on how polluted the water was. A complete block diagram of WSN system is shown in Fig. 8.3. The Arduino Uni Microcontroller use for the data processing, while sensor node is the series of node in data hopping between sensor until to the end of node which is gateway to data center. In the last step of the system block diagram, an antenna was used to transmit the signal and information to the other sensor node and send the information to the data center.

The complete expected indicator of measurement and range of the results in the unit as well as the accuracy is shown in Table 8.1.



Fig. 8.2 Actual scene of Siak River in Riau, Indonesia

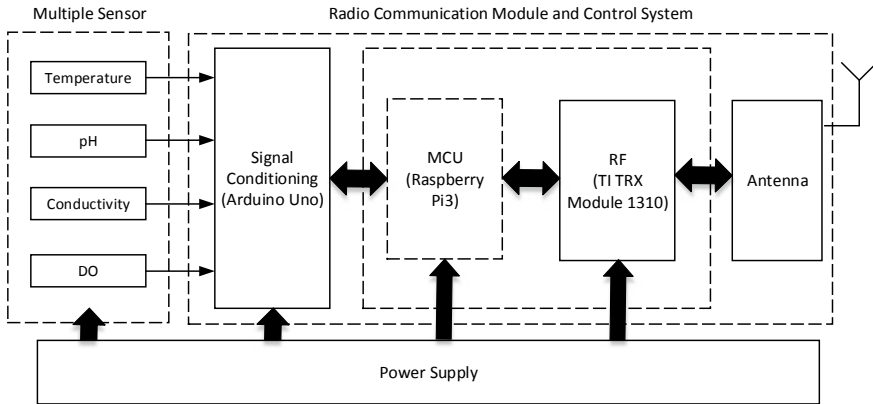


Fig. 8.3 Block diagram of multiple sensor system

Table 8.1 Design specification of the multiple sensor system

Parameter	Range	Accuracy	Method
pH	0–14	±0.1	Glass electrode
Temperature	0–16 °C	±0.5 °C	Thermistor
DO	0–20 mg/L	±0.5 mg/L	Polarography
Electrical conductivity	0–50	±0.5	Conductivity
			Measurement

8.3 System Design for Polluted Water Detection

In the proposed design of the multiple sensing systems for polluted water using four parameters as mentioned earlier, every sensor is contributed to the collection of data and performs the analyses in the system. To detect common polluted water in real time, the sensing system must be deployed to the river. Solar system is required to supply power to the system because of its remote location with no electricity available. In this case, a set of solar panel with backup battery was provided based on the capacity of sensing system that has been tested. In the monitoring of polluted river water, there are several external factors that need consideration such as environmental condition, weather, and temperature. Based on the earlier testing, some parameters increase abnormally with exponential curve. This is because some of the sensing surface was covered by foreign materials that registers high signal but fortunately this takes place for short time.

Design of multiple sensors uses several modes of sensing, and one of them is temperature sensor which contributed to thermistor with nonlinear parameter between temperature and internal resistance. The range of temperature sensor was from 0 to 40 °C. The scale of temperature was selected based on early measurement and average temperature of Siak River water as well as environmental conditions in

Pekanbaru city in Riau Province. In common use, the thermistor is valid for high temperature which is more than 300 °C; thus, low range of temperature is better in detection to avoid the nonlinearity. The resistance of sensor can be scaled using general formula as invented in Steinhart–Hart thermistor third-order approximation and can be written as in Eq. (1) [2]:

$$\frac{1}{T} = A + B \cdot \ln(R) + C \cdot (\ln(R))^3 \quad (1)$$

where T indicates the temperature of water detected in degree Kelvin and R is the measured resistance in Ohm. The parameters A , B , and C are standard constants that were obtained from the manufacturer. These parameters determine the accuracy reading of the sensor. When the sensor was powered, voltage was induced across the thermistor at the fed point and goes into the operational amplifier to gain and fine-tune the offset signal. Value of voltage output from the sensor is in analog which must be converted to digital to match with the WSN system which used Arduino microcontroller. The value of thermistor internal resistance is very much dependent on manufacturer. The internal resistance of sensor represent of temperature, for example 25° is similar to 20 kΩ then the characteristics can be written as in Eq. (2):

$$R_T = |R_0 \cdot e^{\beta \cdot \frac{1}{T} - \frac{1}{T_0}} \quad (2)$$

where R_T is the resistance of the thermistor at T and the temperature is in Kelvin. The value of T_0 is 298.15 °K (or 25 °C), and the value of beta is based on manufacturer's datasheet and specification. Equation 3 was used to calculate the temperature based on manufacturer datasheet as a comparison to the actual value detected during testing. The results of temperature based on the analysis using Eq. (3) are required for the calibration of the temperature as detected by the sensor.

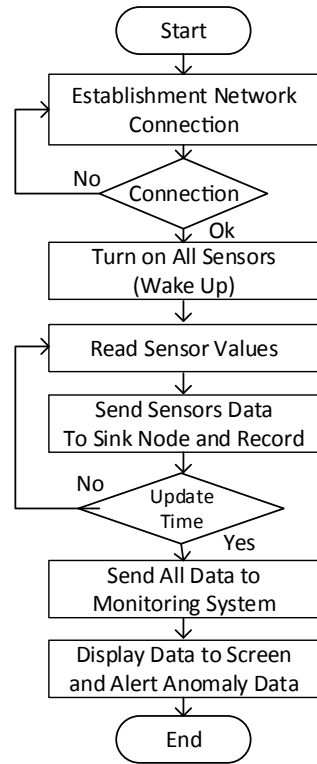
$$T = \frac{\beta}{\ln \frac{R}{r_\infty}} \quad (3)$$

where

$$r_\infty = R_0 \cdot e^{-\beta/T_0} \quad (4)$$

The flow of process in the system is started with establishment of connection from WSN node to sink in order to confirm that all the nodes are connected. Water sensor may in sleep mode (inactive) for while to save power supply, and will on (wake up) when request to collect data or to sense the water parameter. Figure 8.4 shows the flowchart of the process in the sensing system for water pollution.

Fig. 8.4 A flowchart of the process in WSN system



8.3.1 Multiple Sensor System

The multiple sensor system has been done on a prototype as well as tested in the laboratory. Sensors for the detection of river water pollutants as shown in the specifications in Table 8.1 were fabricated to connect to the microcontroller. Figure 8.5 shows the fabricated system tested on a mini scale. Results show that the system was able to read all the water parameters and are shown in the LCD. Next step is to test the prototype after improving the casing to be taken to the riverside.

The testing was done in the laboratory for long period of time to check its long-term performance. Results show that the readings for the various parameters give accurate results when compared to the manual calibration or conventional way. The use of multiple systems for detection is good because it provides various indicators of polluted water to be analyzed to ensure that the final result for determination of polluted water is more accurate. Furthermore, introducing an intelligent system on microcontroller programming assists the accuracy of the decision on the results.

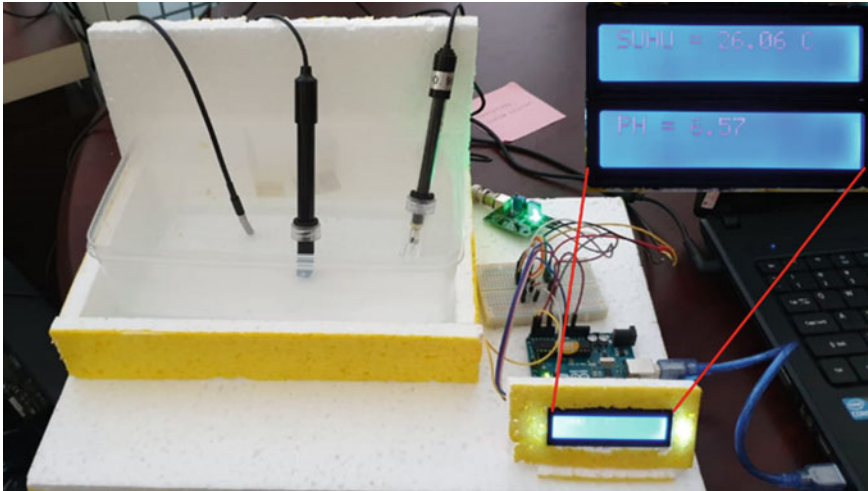


Fig. 8.5 Prototype of multiple sensors for water pollutant detection

8.3.2 WSN Sensing Communication

Communication to the backend system is one of the requirements to pull data to the server and monitoring system. In actual conditions, the sensor system was located on-site at the riverside. Several sensors were connected to each other for data sharing, and a system was used as a gateway for communication to the backend system. Based on the survey, the distance of site location can be more than 30 km to the backend system. Figure 8.6 shows a design of data transfer to the remote monitoring, and every sensor node represented water pollutant sensing system that has their individual sink for data collection and was kept on-site by local host before sending to the monitoring system. The proposed multiple sensing for water pollutant applies 4G as network for communicating from sink node to backend. The database or data center allows faster

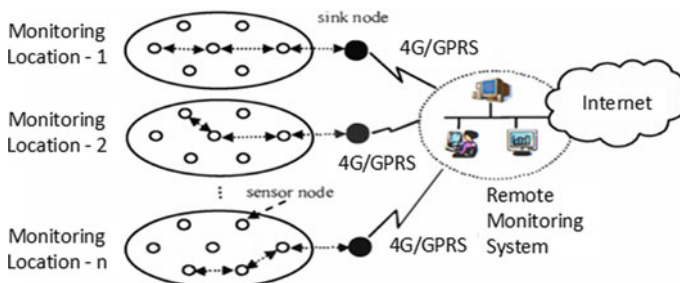


Fig. 8.6 WSN node communication to sink node and monitoring system

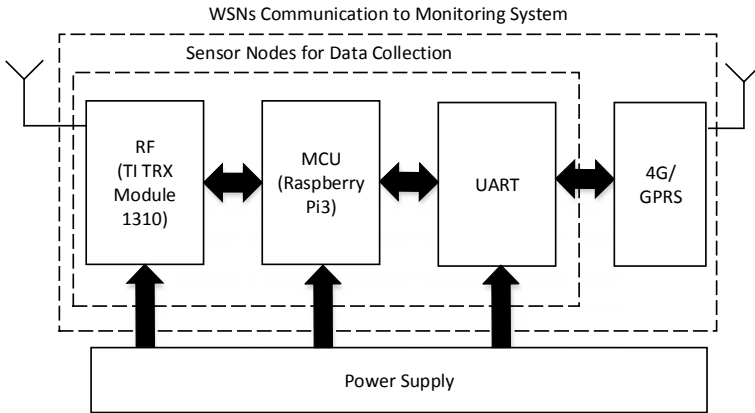


Fig. 8.7 Proposed diagram of WSN sink node communication to sensors

transfer of data as well as real-time monitoring, which in most of area are currently covered by network in latest technology such as 4G.

The system design with real-time monitoring system thus was able to detect data from the sensors which must be transferred immediately with short delay. A block diagram for communication to the backend is shown in Fig. 8.7. Fourth-Generation (4G) technology or General Package Radio Service (GPRS) was used; then, river water pollution data were collected in an interval of time to minimize dumb and useless data that can be waste in local memory. Universal asynchronous receiver/transmitter (UART) unit was used as interface between MCU and the 4G communication unit.

The design of multiple sensor system consists of four parameters which are common indicator in polluted water, but the sensor node for WSNs was able to serve up to fifty nodes or location in 10 of a sink node. The distance of sensing system node from a location to others according to site survey must not be more than 500 m because longer distance results in low accuracy because of the fast river water flow, especially during rainy season. The system was supplied with independent power system from the solar panel because some locations are very far away from the electrical utility. Figure 8.8 shows an actual scan of Siak River located in the capital of Riau Province. A set of sensor under testing on-site was used to get the readings for analysis and calibration to be compared to the actual value. Many activities on the river create the pollution to the river and affect the sensor reading. Needless to say, the polluted river water can be very hazardous to the community when consumed or used for daily activities.



Fig. 8.8 A set of sensor under testing on-site

8.3.3 *IoT System and Communication*

Internet of things (IoT) is a technology that was used as support in this case with integration to WSN system, the design scenario of sensing system for effective communication and data transfer in order to achieve good response of sensors in the location. Combination between WSN system and IoT as proposed in this design enables optimum data collection from every sensor to be kept in buffer memory in sink node of WSNs. Figure 8.9 shows a network architecture of IoT and WSN integration.

After the sensors collect and store signals in memory in sink node, the next process is to provide useful data information for IoT application. In addition, integration of WSNs and IoT can be as such properties.

The placement of the sensor nodes of IoT and sink nodes of WSNs is in static locations. Location of sensor node is fix (coordinate x and y) for all node and has same distance between of each node, then data will passing by hopping all the way of node Eq. (5).

$$\text{dist} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (5)$$

where $(x_1; y_1)$ is the location of the first node and $(x_2; y_2)$ is the location of the second node. The distance is calculated based on effective communication of WSN node to transfer the data.

The power of transmission for each sensor node is set as required; although the distance is the same, every sink node has a different path loss and environmental effect.

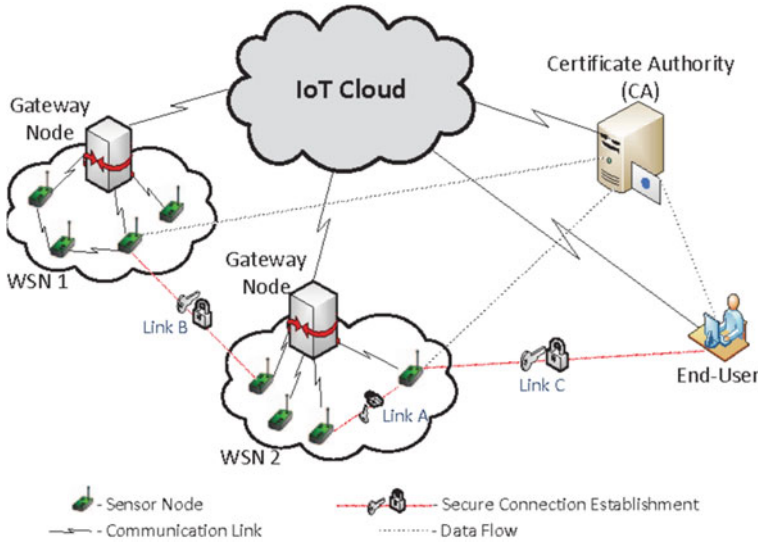


Fig. 8.9 An architecture of WSNs and IoT system integration

8.4 Simulation Results and Discussion

The results from the in-house testing system that measures temperature, water pH, electrical conductivity, and DO are compared to the conventional system which manually measures the quality of the tested water. Preliminary testing is very important to make sure that the readings of the sensing system are accurate compared to the actual conditions. Some of the results from the sensors were compared to other datasheet and literature in reference [2]. Good agreement of results from the two methods was found (Fig. 8.10). The deviation between the sensing system and the readings of manual thermometer is very minimum (0.071–1 °C).

Electrical conductivity is one of the indicators used to obtain information on polluted water. Small error on results based on testing for the electrical conductivity is expected to get high accuracy for the determination of polluted water. Normally, error for this measurement is not more than 15%, similar to other indicators of polluted water. Good agreement of results in the comparison between readings of simulated electrical conductivity and the actual signal conditioning was found (Fig. 8.11).

Another common indicator to measure water quality is water pH, sensing system for detection of water pH was designed to integrate other sensors, and results were analyzed using the same microcontroller. Referring to the table, the specification of pH within the range from 0 to 14 is within 0.1 accuracy. Figure 8.12 shows the results of measurement of water pH compared to the theoretical values.

According to the initial testing in the laboratory, all the sensors were able to function well and were able to detect water parameters as displayed on the LCD. Further action is required to install and to do testing at the actual site as the proposed system.

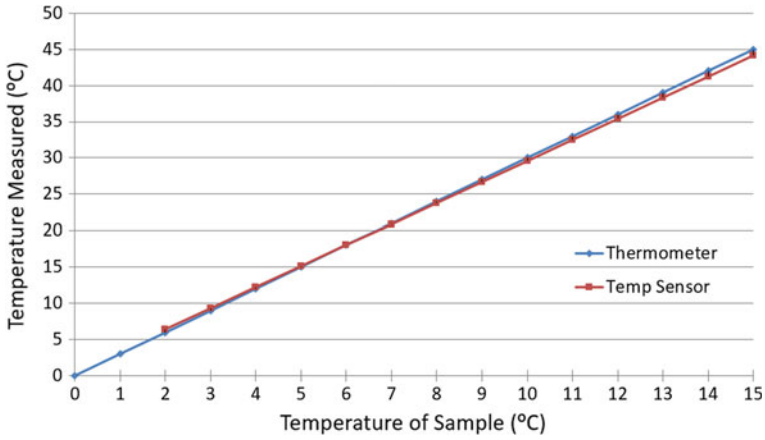


Fig. 8.10 Comparison of temperature obtained from the proposed system to that obtained using conventional method using thermometer

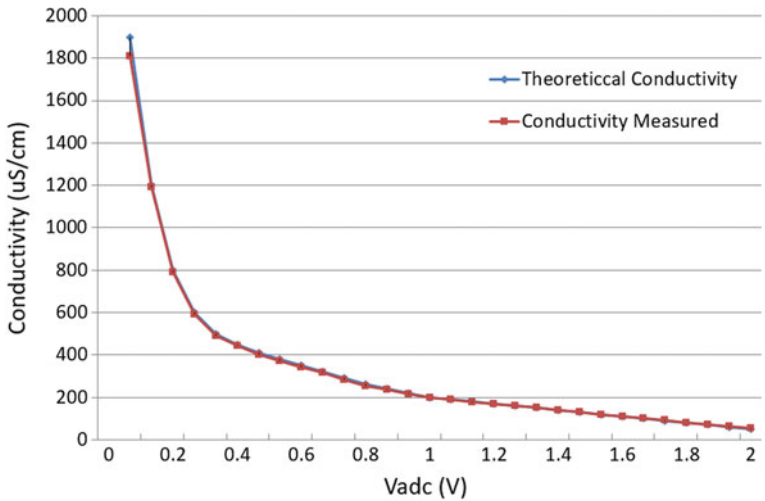


Fig. 8.11 Test results of the electrical conductivity of the sensor node

The results expected to achieve high accuracy based on actual contain polluted water. Figure 8.13 shows the results of water flow meter between manual and sensing systems.

According to the initial testing in the laboratory for all the parameters of the water, good agreement of results between the conventional measurement unit and the sensing systems was found. Thus, the proposed system has a significant impact on the community and further step was continued to test on-site for real environmental conditions. The proposed system applied both WSN and IoT technologies

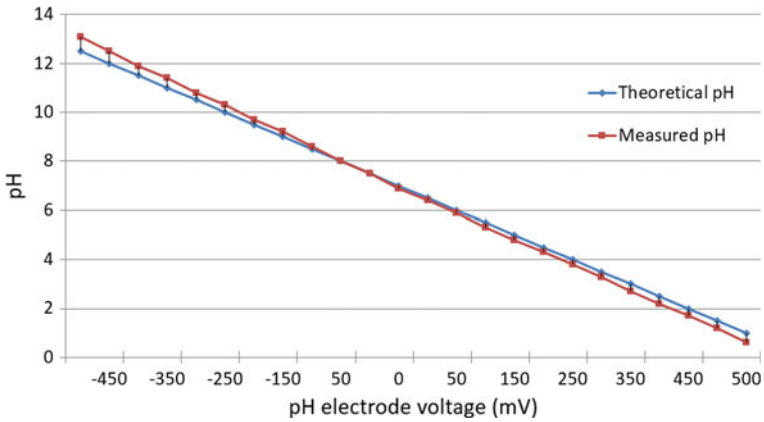


Fig. 8.12 Water pH test results versus theoretical values

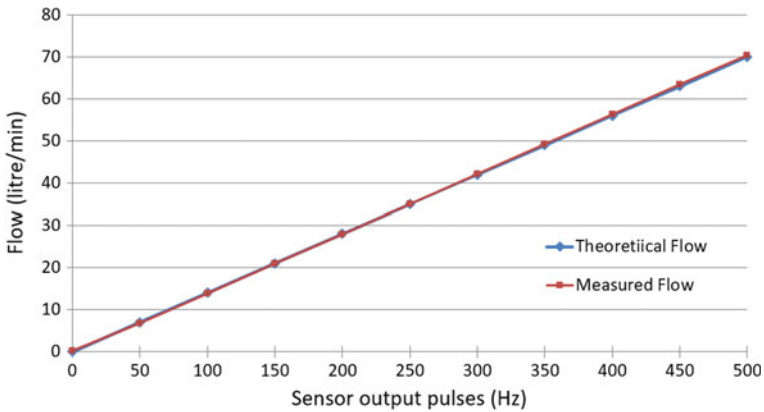


Fig. 8.13 Water flow meter test versus theoretical analysis

for monitoring. Distance between sensing node and the sampling site is one of the considerations to achieve good and representative readings of the sensing system.

8.5 Conclusion

A system was proposed for the assessment of polluted water using multiple sensors; the fabricated unit was tested in the laboratory as well as initial testing on-site. Results show good agreement between the proposed system and that of conventional measurement. Four main indicators in sensing system such as water pH, temperature, DO, and electrical conductivity were measured to assess the quality of river

water. The proposed system applied intelligent system as well in programming the microcontroller to achieve high accuracy in the final decision based on the detected values. For further development is for water level that benefit for community for flooding alert. Finally, to make sensing system smart, intelligent algorithm should be applied into microcontroller programming because of various types of material and chemicals in the water.

Acknowledgements We appreciate the Ministry of Research, Technology, and Higher Education (KEMENRISTEKDIKTI) of Indonesia for funding this research project and Universitas Islam Riau, Indonesia, to support the facilities as well as Chiba University, Japan; Universiti Teknologi Malaysia; and Universiti Teknologi PETRONAS, Malaysia.

References

1. Guo, Y., et al.: An inversion-based fusion method for inland water remote monitoring. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **9**(12), 5599–5611 (2016)
2. Cloete, N.A., Malekian, R., Nair, L.: Design of smart sensors for real-time water quality monitoring. *IEEE Access* **4**, 3975–3990 (2016)
3. Wang, Y., Rajib, S.M.S.M., Collins, C., Grieve, B.: Low-cost turbidity sensor for low-power wireless monitoring of fresh-water courses. *IEEE Sens. J.* **18**(11), 4689–4696 (2018)
4. Woutersen, M., et al.: Development and validation of an on-line water toxicity sensor with immobilized luminescent bacteria for on-line surface water monitoring. *MDPI Sens.* **17**(2682), 1–14 (2017)
5. Li, L.Y., Jaafar, H., Ramli, N.H.: Preliminary study of water quality monitoring based on WSN technology. In: 2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA), pp. 1–7 (2018)
6. Kadir, E.A., Rosa, S.L., Yulianti, A.: Application of WSNs for detection land and forest fire in Riau Province Indonesia. In: 2018 International Conference on Electrical Engineering and Computer Science (ICECOS), pp. 25–28 (2018)
7. Lockridge, G., Dzwonkowski, B., Nelson, R., Powers, S.: Development of a low-cost arduino-based sonde for coastal applications. *MDPI Sens.* **16**(528), 1–16 (2016)
8. Islam, T., Lee, Y.K.: A two-stage localization scheme with partition handling for data tagging in underwater acoustic sensor networks. *MDPI Sens.* **19**(2135), 1–27 (2019)
9. Kadir, E.A., Siswanto, A., Rosa, S.L., Syukur, A., Irie, H., Othman, M.: Smart sensor node of WSNs for river water pollution monitoring system. In: 2019 International Conference on Advanced Communication Technologies and Networking (CommNet), pp. 1–5 (2019)
10. Yuan, F., Huang, Y., Chen, X., Cheng, E.: A biological sensor system using computer vision for water quality monitoring. *IEEE Access* **6**, 61535–61546 (2018)
11. Lin, W.-C., Brondino, K., Monroe, C.W., Burns, M.A.: Multifunctional water sensors for pH, ORP, and conductivity using only microfabricated platinum electrodes. *MDPI Sens.* **17**(1655), 1–9 (2017)
12. Lu, Y., Macias, D., Dean, Z.S., Kreger, N.R., Wong, P.K.: A UAV-mounted whole cell biosensor system for environmental monitoring applications. *IEEE Trans. NanoBiosci.* **14**(8):811–817 (2015)
13. Lambrou, T.P., Anastasiou, C.C., Panayiotou, C.G., Polycarpou, M.M.: A low-cost sensor network for real-time monitoring and contamination detection in drinking water distribution systems. *IEEE Sens. J.* **14**(8), 2765–2772 (2014)
14. Tian, J., Wang, Y.: A novel water pollution monitoring approach based on 3s technique. In: 2010 International Conference on E-Health Networking Digital Ecosystems and Technologies (EDT), vol. 1, pp. 288–290 (2010)

15. Grossi, M., Lazzarini, R., Lanzoni, M., Pompei, A., Matteuzzi, D., Riccò, B.: A portable sensor with disposable electrodes for water bacterial quality assessment. *IEEE Sens. J.* **13**(5), 1775–1782 (2013)
16. Kadir, E.A., Efendi, A., Rosa, S.L.: Application of LoRa WAN sensor and IoT for environmental monitoring in Riau Province Indonesia. In: 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2018), Malang. IEEE (2018)
17. Doña, C., Sánchez, J.M., Caselles, V., Domínguez, J.A., Camacho, A.: Empirical relationships for monitoring water quality of lakes and reservoirs through multispectral images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **7**(5), 1632–1641 (2014)
18. Olatinwo, S., Joubert, T.-H.: Optimizing the energy and throughput of a water-quality monitoring system. *Sensors* **18**(4), 1198 (2018)
19. Teixidó, P., Gómez-Galán, J., Gómez-Bravo, F., Sánchez-Rodríguez, T., Alcina, J., Aponte, J.: Low-power low-cost wireless flood sensor for smart home systems. *Sensors* **18**(11), 3817 (2018)
20. Luna, F.D.V.B., de la Rosa Aguilar, E., Naranjo, J.S., Jagüey, J.G.: Robotic system for automation of water quality monitoring and feeding in aquaculture shadehouse. *IEEE Trans. Syst. Man Cybern. Syst.* **47**(7), 1575–1589 (2017)
21. Wu, Z., Liu, J., Yu, J., Fang, H.: Development of a novel robotic dolphin and its application to water quality monitoring. *IEEE/ASME Trans. Mechatron.* **22**(5), 2130–2140 (2017)
22. O'Connor, E., Smeaton, A.F., O'Connor, N.E., Regan, F.: A neural network approach to smarter sensor networks for water quality monitoring. *MDPI Sens.* **12**, 4605–4632 (2012)

Part II
Signal and Image Processing

Chapter 9

Dimension Detection of Varistor Based on Random Hough Transform



Wei Chen and Xueying Yang

Abstract An improved method based on edge detection and random Hough transform is proposed for industrial detection of varistors. Firstly, using the improved edge detection to detect the contour, the protruding part can be smoothed effectively, making it more fit to the circular feature without changing the size. Then, the improved random Hough transform is used to accurately locate the circle, the properties of the circle are determined by using three points, three non-collinear pixel points are extracted from the edge point, the circle satisfying the conditions is found, and then the fourth point is randomly taken out. If the difference between the distance and radius from the point to the center of the circle is within the threshold 1, it may be true circle. Then, if the difference between the center and radius of the detected circle is greater than the threshold 2, the detected circle is determined to be true circle. Otherwise, the duplicate circle is deleted.

9.1 Introduction

Product testing is an important link to ensure product quality. Traditional manual testing is easy to be influenced by subjective factors. Long-term testing can lead to fatigue of workers and lead to missed or wrong inspection, which leads to low precision of testing results. And in a large number of product testing, sampling inspection is usually adopted, which cannot guarantee the quality of all products. Varistors are widely used in industry. Because the shape of varistors is round, image recognition can be used to solve this problem. The common circular detection methods include standard Hough transform (SHT) [1] and Hough transform based on gradient transform [2]. But the method needs to introduce a three-dimensional voter, the calculation amount is large, and a large amount of occupied calculation memory is occupied, so the operation speed is relatively slow, and the timeliness of industrial detection

W. Chen · X. Yang (✉)

College of Communication and Information Engineering, Xi'an University of Science and Technology, 710000 Xi'an, China
e-mail: Young11201211@163.com

© Springer Nature Singapore Pte Ltd. 2021

S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation, Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_9

117

is not met. In view of this defect of standard Hough transform, Xu et al. [3] proposed random Hough transform (RHT). By randomly sampling three non-collinear edge points, the position parameters of circle can be accumulated, but it is easy to detect repeated circle by this method. In order to solve this problem, an improved algorithm based on random Hough transform is proposed in this paper. The limit of center coordinate and radius is added. If the two center distance and radius distance are less than the threshold, then one of them is kicked out. Then, the accuracy of detection is greatly improved.

9.2 Improvement of a Canny Edge Algorithm

The conventional Canny edge detection is processed by using a Gaussian filter at the time of filtering, while the interference of the noise can be suppressed to a certain extent, but the useful edge cannot be extracted well for the image with the “Burr” [4]. In this case, an improved Canny edge detection is proposed.

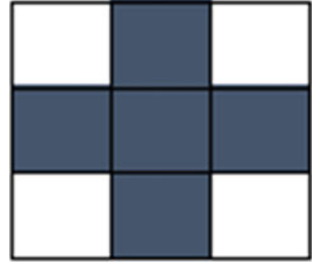
9.2.1 Mean Filter and Closed Operation

Mean filtering is also called linear smoothing filter. The principle of mean filtering is to calculate the average value of the neighborhood of any point, and the average value is obtained from the sum of the neighborhood divided by the area of the neighborhood [5]. The filtering effect of the mean filtering method is related to the template size used. Usually, the larger the radius of the template, the more blurred the filtered image is and the slower it runs. In this paper, the window selected by many experiments is order 5. Meanwhile, under the condition that the edge information is preserved as far as possible without changing the size of the image, the closed operation is introduced. Closed operation is the first expansion of the image and then corrosion [6].

$$I \cdot S = (I \oplus S) \otimes S \quad (9.1)$$

where I is the image and S is the structural element. Common structural elements are rectangular structure, elliptical structure and cross-shaped structure. The image to be detected in this paper is circular, and through practice, the structure is more similar to the cross-shaped structural elements. Figure 9.1 is a cross-shaped structural element. $I \oplus S$ denotes the expansion operation on the image, taking the maximum value within the cross neighborhood of each position as the output grayscale value of that position. $I \otimes S$ indicates that the image is eroded. That is, take the minimum value of the value in the cross neighborhood of each position as the output value of that position.

Fig. 9.1 Cross-shaped structure



9.2.2 Edge Detection

The edge of the image refers to the location where the grayscale value changes dramatically. Sometimes the image undefined content can be understood only by considering the edge of the image [7]. In this paper, the third-order sobel operator is used to calculate the gradient amplitude and direction.

$$\text{sobel}_x = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} * \begin{bmatrix} 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \quad (9.2)$$

$$\text{sobel}_y = \begin{bmatrix} 1 & 2 & 1 \end{bmatrix} * \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (9.3)$$

where sobel_x denotes the convolutional kernel in the horizontal direction and is used to detect the vertical edge. sobel_y denotes the convolutional kernel in the vertical direction and is used to detect the horizontal edge. Image matrix I and sobel_x , sobel_y convolution to get dx and dy . So you can get to the gradient G and direction θ of the pixel.

$$G = \sqrt{dx^2 + dy^2} \quad (9.4)$$

$$\theta = \arctan 2(dy, dx) \quad (9.5)$$

In order to improve the accuracy of edge location, the gradient assignment is refined by interpolation on the basis of Canny algorithm [8]. Interpolation pixels in the neighborhood. Decompose the perimeter of the pixel into 8 pixel neighborhoods, as shown in Fig. 9.2. A total of 9 pixels centered on $A(i, j)$ are calculated and interpolation in the gradient direction. If the amplitude of the center point $A(i, j)$ is less than the interpolation of its adjacent two gradient directions, then it should be a non-edge point and assign $A(i, j)$ to 0; if $A(i, j)$ is greater than or equal to two adjacent interpolation in the gradient direction, then $A(i, j)$ is used as the primary edge point.

Fig. 9.2 Location of pixel points

$A(i-1,j-1)$	$A(i,j-1)$	$A(i+1,j-1)$
$A(i-1,j)$	$A(i,j)$	$A(i+1,j)$
$A(i-1,j+1)$	$A(i,j+1)$	$A(i+1,j+1)$

Taking $A(i, j)$ as an example, the gradient angle at $A(i, j)$ is first found, then the straight line in the gradient direction is drawn according to the gradient direction, and finally in the neighborhood centered on $A(i, j)$, the neighborhood in the gradient direction is generally located. Next, $A(i, j)$ is compared with the value in the gradient direction, if its value is larger than the neighborhood value in the gradient direction, it can be regarded as the maximum value, $A(i, j)$ is selected as the edge point; if its value is not all larger than the value of the neighborhood in the gradient direction, it can be regarded as a non-maximum value, which needs to be suppressed so that $A(i, j) = 0$ [9].

The false edge can be effectively reduced by threshold of the processed image. It sets two threshold values and deals with them according to the following rules. The main results are as follows:

1. Those points whose edge strength is greater than the high threshold are used to determine the edge points.
2. Those points whose edge strength is smaller than the low threshold are immediately eliminated.
3. Those points whose edge strength is between the low threshold and the high threshold can only be accepted as the edge points when they can be connected to the determined edge points according to a certain path. The edge strength of all the points that make up this path is greater than the low threshold.

9.3 Improved Circle Detection Algorithm Based on Random Hough Transform

On the basis of Canny edge detection, the circular device is located. The traditional Hough transform (CHT) detection principle is as follows:

The circle with (a, b) as the center of the circle and r as the radius is transformed into the three-dimensional space with a, b, r as the parameter, and the following equation is obtained: $(x_i - a)^2 + (x_i - b)^2 = r^2$. Each pixel on the edge of the same hole corresponds to a unique conical surface in the parameter space, and these cones intersect at one point. The point circle parameter can be obtained by cumulative voting, that is, the position parameter of the hole.

However, this method needs to establish a three-dimensional voting machine, the calculation is large, the operation speed is slow, and it is not suitable for industrial detection. Random Hough transform is the random selection of three non-collinear

points in the image space to map to a point in the parameter space, which is a multi-to-one mapping; thus, avoiding the huge amount of computation of the traditional Hough transform from one to multiple mapping. In order to reduce the memory requirement, the dynamic linked list structure is adopted to accumulate the parameter allocation units obtained by more than one mapping. Compared with the traditional Hough transform, the memory requirement is reduced and the parameter space of RHT is infinite. The parameter precision is arbitrarily high and so on. However, when dealing with complex images, a large number of invalid units are still introduced into random sampling, resulting in a large number of invalid accumulations. In this paper, an improved RHT is proposed for circle detection.

Let D be the set of edge points, select four edge pixels randomly, and define a distance criterion to find the possible circle. When we find a possible circle, we use constraints to determine whether it is a true circle or not. The principles of the algorithm are as follows:

In a rectangular coordinate system, the general equation of a circle is expressed as follows:

$$(x - a)^2 + (y - b)^2 = r^2 \quad (9.6)$$

It can also be expressed as:

$$2xa + 2yb + d = x^2 + y^2 \quad (9.7)$$

Among them: $d = r^2 - a^2 - b^2$. Select three pixels $v_i(x_i, y_i)$, $i = 1, 2, 3$. If these three points are not collinear, they may be a circle with a center (a_{123}, b_{123}) and a radius of r_{123} . Obtained by Eq. (9.5):

$$\begin{cases} 2x_1 a_{123} + 2y_1 b_{123} + d_{123} = x_1^2 + y_1^2 \\ 2x_2 a_{123} + 2y_2 b_{123} + d_{123} = x_2^2 + y_2^2 \\ 2x_3 a_{123} + 2y_3 b_{123} + d_{123} = x_3^2 + y_3^2 \end{cases} \quad (9.8)$$

Among them, $d_{123} = r_{123}^2 - a_{123}^2 - b_{123}^2$. The above three equations are expressed in matrix form:

$$\begin{pmatrix} 2x_1 & 2y_1 & 1 \\ 2x_2 & 2y_2 & 1 \\ 2x_3 & 2y_3 & 1 \end{pmatrix} \begin{pmatrix} a_{123} \\ b_{123} \\ d_{123} \end{pmatrix} = \begin{pmatrix} x_1^2 + y_1^2 \\ x_2^2 + y_2^2 \\ x_3^2 + y_3^2 \end{pmatrix} \quad (9.9)$$

When v_1, v_2, v_3 not collinear, that is, $(x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1) \neq 0$. Solve it:

$$a = \frac{\begin{vmatrix} x_2^2 + y_2^2 - (x_1^2 + y_1^2) & 2(y_2 - y_1) \\ x_3^2 + y_3^2 - (x_1^2 + y_1^2) & 2(y_3 - y_1) \end{vmatrix}}{4((x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1))} \quad (9.10)$$

$$b = \frac{\left| \begin{array}{l} 2(x_2 - x_1) x_2^2 + y_2^2 - (x_1^2 + y_1^2) \\ 2(x_3 - x_1) x_3^2 + y_3^2 - (x_1^2 + y_1^2) \end{array} \right|}{4((x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1))} \quad (9.11)$$

$$r_{123} = \sqrt{(x_i - a_{123})^2 + (y_i - b_{123})^2} \quad (9.12)$$

Is taken $v_4(x_4, y_4)$ as the fourth pixel point, and the distance from the pixel v_4 to the circle c_{123} is d_{4-123} . That is:

$$d_{4 \rightarrow 123} = \left| \sqrt{(x_4 - a_{123})^2 + (y_4 - b_{123})^2} - r_{123} \right| \quad (9.13)$$

We set a threshold t ($t = 1$), when $d_{4-123} \leq t$, the point is considered to be on the circle. Random Hough transform sometimes detects multiple adjacent circles near the detected target, as shown in Fig. 9.1. In order to solve this problem, we introduce the limit of center coordinate distance and radius coordinate distance. The distance between the center of the circle, the longitudinal coordinates, and the radius of the detected circle is calculated in pairs, that is:

$$m = |a_{123} - a_{1'2'3'}| \quad (9.14)$$

$$n = |b_{123} - b_{1'2'3'}| \quad (9.15)$$

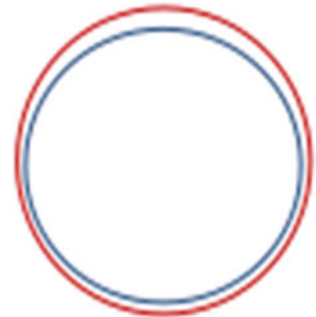
$$l = |r_{123} - r_{1'2'3'}| \quad (9.16)$$

If the values of the above three are all less than the set threshold z ($z = 15$), then delete one of the circles, leaving a better detection effect of the circle (Fig. 9.3).

The algorithm steps are as follows:

1. Constructing edge pixels l set V , $V_i = (x_i, y_i)$. Initialize the failure counter f to be 0. Set the six thresholds of $T_f, T_{\min}, T_a, T_r, T_b, T_c$. If is the maximum number of failures allowed, T_{\min} represents the minimum number of remaining pixels in V ; T_a represents the minimum distance between any two selected pixels; T_d

Fig. 9.3 Two adjacent circles



represents the distance from the fourth pixel point to the center of the circle being detected; T_r represents the ratio of the detected edge point set to the theoretical $2\pi r$ point set; T_b represents the distance threshold of the detected close circular horizontal and vertical coordinates; T_c represents the radius distance threshold. $|V|$ represents the number of pixels remaining in V .

2. If $f = T_f$ or $|V| < T_{\min}$, then stop detecting. Otherwise, 4 pixels ($v_i, i = 1, 2, 3, 4$) are selected randomly. At the same time, $v = v - \{v_i\}$.
3. Select three edge pixels, solve the parameters of the circle and satisfy the constraints. If fit, turn Eq. (9.4). Otherwise, it is put back into collection V , and $f = f + 1$. Then, turn Eq. (9.2).
4. Initialize the counter $C = 0$, and for each v_i in V , check whether d_{i-ijk} is less than the given distance threshold T_d . If satisfied, the value of the counter is added by 1, and the v_i is taken out of V . Traverse all the pixels in V , let $c = d$, the number of pixels satisfying the condition is np .
5. If $np \geq 2\pi r T_r$ and $m, n \leq T_b, l \leq T_c$, turn Eq. (9.6). Otherwise, the detected circle is considered to be a false circle. Put np pixels back into V and turn Eq. (9.2).
6. Think that the possible circle detected is true circle, set $f = 0$, turn Eq. (9.2).

9.4 Dimensional Inspection of Varistor

In order to verify the effectiveness of the algorithm, this paper uses OpenCV3.4 and Python3.5 language to simulate. In the experiment, a 2 million-pixel industrial camera was used to capture a varistor image. Figure 9.4 is the traditional Canny algorithm detection, Fig. 9.5 is improved Canny algorithm detection, we can clearly see that the improved algorithm not only smoothed the contour of varistor image, but also eliminated the noise well, and made the edge clearer. And retained the real edge, more in line with the industrial testing requirements of varistors.

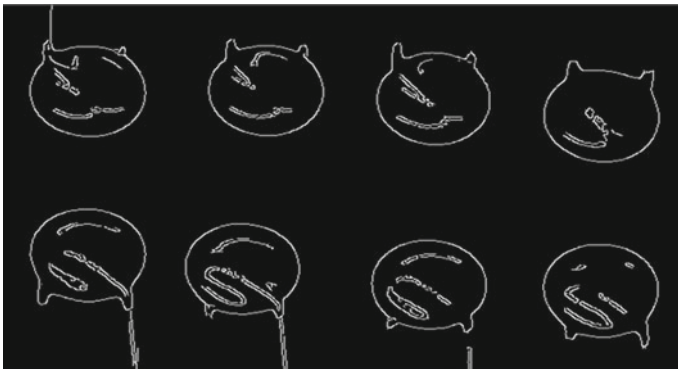


Fig. 9.4 Traditional Canny edge detection

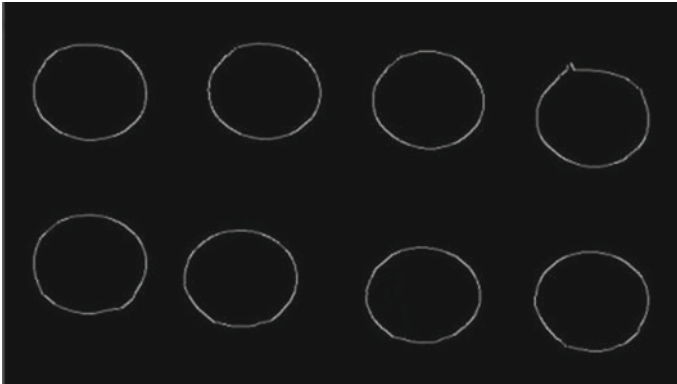


Fig. 9.5 Improved Canny edge detection

Traditional RHT detection circles are prone to false circles or omissions, that is, multiple circles are detected in the same circle or some circles are not detected, as shown in Fig. 9.6. In view of this situation, an improved RHT is proposed to increase the limit of center distance and radius distance, which effectively solves the problem. The average detection time is kept at 0.9 s, and the time standard of industrial detection is met at the same time. The test result is shown in Fig. 9.7.

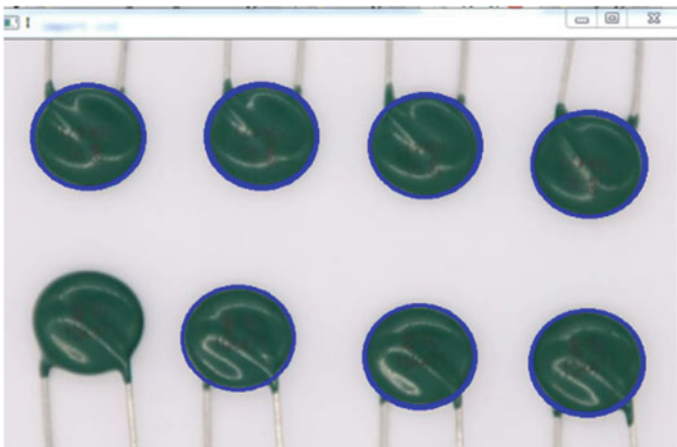


Fig. 9.6 Traditional RHT detection images

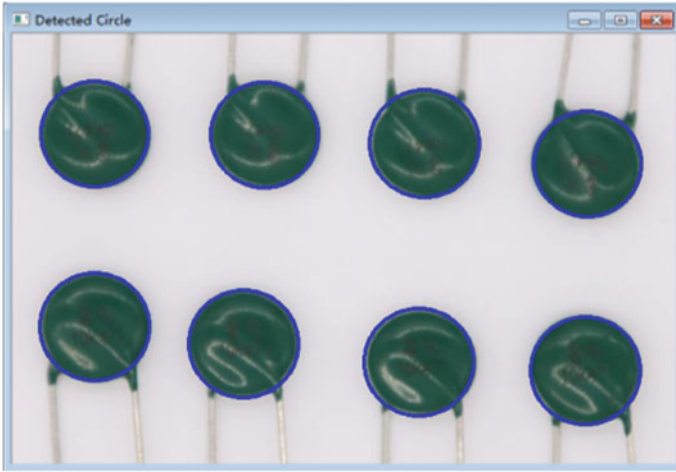


Fig. 9.7 RHT detection image after improvement

9.5 Conclusion

Varistor is a common electronic component, for its automatic detection, and is a certain market demand. The time and space complexity of the standard Hough transform is relatively high, so the detection time is relatively long that it is not suitable for industrial detection. Before the detection, the mean filter is used to de-noising the image, and the expansion and corrosion of the image edge are smoothed to make it more consistent with the circular features. As can be seen from Fig. 9.6, the traditional Hough transform detects a plurality of duplicate circles in the first row and omits the first varistor in the second line. Obviously, this effect does not meet the test requirements. The research in this paper shows that the improved Hough transform can avoid the occurrence of this situation, and the new detection method satisfies the detection time and the detection accuracy.

References

1. Dongfeng, R., Qiubing, W., Fujun, S.: A fast and effective algorithm based on improved hough transform. *J. Indian Soc. Remote. Sens.* 44(3), (2016)
2. Shang, F., Liu, J., Zhang, X., Tian, D.: An improved circle detection method based on right triangles inscribed in a circle. *WRI World Congress on Computer Science and Information Engineering* (2009)
3. Kalviainen, H., Oja, E., Xu, L.: Randomized Hough transform applied to translational and rotational motion analysis. *Proceedings, 1992 Conference A: Computer Vision and Applications, 11th IAPR International Conference on Pattern Recognition, vol. I* (1992)
4. Zhao, X. M., Wang, W. X., Wang, L. P.: Parameter optimal determination for canny edge detection. *Imaging Sci. J.* 59(6), (2011)

5. Rosal, D., Setiadi, M. I., Jumanto, J.: An enhanced LSB-image steganography using the hybrid Canny-Sobel edge detection. *Cybern. Inf. Technol.* **18**(2), (2018)
6. Yiming, Z., Jun, W.: Research on iris recognition algorithm based on hough transform. *IOP Conf. Ser. Mater. Sci. Eng.* **439**(3), (2018)
7. Parida, N., Bhoi, N.: 2-D Gabor filter based transition region extraction and morphological operation for image segmentation. *Comput. Electr. Eng.* **62**, (2017)
8. Meng, Y., Zhang, Z., Yin, H., Ma, T.: Automatic detection of particle size distribution by image analysis based on local adaptive canny edge detection and modified circular Hough transform. *Micron* (Oxford, England: 1993) **106**, (2018)
9. Duan, L., Wang, W., Zhang, X.: Improved Hough transform for circular inspection. *Comput. Integr. Manuf. Syst.* **19**(09): 2148–2152, (2013)

Chapter 10

Phase Retrieval Method Based on Transport of Intensity Equation with Microscope Single Field of View



Hong Cheng, Rui Wang, Fen Zhang, Wenxia Bao, and Quanbing Zhang

Abstract In microscopic imaging, transport of intensity equation (TIE) is an effective phase retrieval method. In order to effectively calculate the lost phase from the intensity information, a phase retrieval method based on the combination of traditional microscope monocular movement and TIE is proposed. In the eyepiece interface, a C adapter ring is used to connect the CCD detection device, and then, the C adapter ring is moved along the optical axis to obtain multiple single field of view intensity images. After registration, the phase of the sample is calculated by combining TIE. This method utilizes the microscope eyepiece interface, which can change the defocus distance conveniently and quickly. Real experiments, respectively, test the phase retrieval ability of the method for different complexity case. Finally, the correctness and effectiveness of the algorithm are verified by experimental results.

10.1 Introduction

In the field of optical microscopy, there are a variety of observed samples that are phase objects. The amplitude transmission distribution of such objects is uniform, but the spatial distribution of refractive index or thickness is not uniform. This causes the change in the amplitude portion of the wave to be weak, but the phase change is very significant, which means that the phase distribution highlights the differences of refractive index and thickness of microscopic structures. However, the human eye or other digital detector can only obtain the amplitude changes caused by the object but cannot detect the phase changes, so it is necessary to obtain phase information in microscopic imaging.

Numerous techniques have been developed for phase measurement in the field of microscopic imaging, such as Zernike phase contrast microscopic [1] and differential interference contrast (DIC) microscopic imaging [2] which can be qualitatively observed, but these methods are impossible to quantitatively recover the

H. Cheng (✉) · R. Wang · F. Zhang · W. Bao · Q. Zhang
Anhui University, Hefei, China
e-mail: chenghong@ahu.edu.cn

phase of the object. For quantitative phase imaging methods [3], there are interferometric approaches by introducing the principles of interferometry and holography into microscopy, and on a different note, there is a phase retrieval algorithm based on transport of intensity equation (TIE) method [4–8]. Compared with the former, TIE is a very important non-interferometric method, which only needs to capture the intensity images of multiple planes along the optical axis without explicit manipulation of object and reference beam, then it can directly obtain the phase information through calculation. TIE-based approaches are also applicable under some partially coherent illumination, so more and more scholars apply TIE to microscopic imaging field [9–11]. In 1998, Barty et al. [12] described the quantitative phase contrast imaging of unstained cheek cells with optical fiber by TIE, which was the first application of TIE in quantitative optical phase contrast imaging. In recent years, Zuo et al. of Nanjing University of Science and Technology have proposed multimodal computational microscopy based on TIE [13], which used an electrically tunable lens module to connect with the camera port of the microscope. Tian et al. [14] from Jiangnan University proposed a dual-view TIE method based on traditional microscope equipment by placing a 3 mm C-mount brass spacer rings between the C interface of the CCD and the binocular tube to capture the over- and under-focus images. , it is a pity that the method cannot change the defocus distance conveniently, and the detailed correction process to solve the problem of image deviation is not given.

In this work, we propose a new phase retrieval method based on the combination of monocular movement and TIE in traditional microscope. This method ingeniously utilizes the eyepiece interface of a traditional microscope, but the position between the images will be misaligned. Therefore, the geometric position correction of intensity image is realized by registration. The phase results of phase objects, such as cells, can be obtained in a simple experimental device, which provides an auxiliary function for morphological observation of phase objects and will be helpful for its 3D reconstruction in the future.

10.2 Transport of Intensity Equation Under Partially Coherent Optical Field

With the beam of monochromatic optical waves propagating along the z -axis and under paraxial approximation conditions, the relationship between intensity and phase can be expressed as [6]

$$\frac{\partial I(x, y, z)}{\partial z} = -\frac{\lambda}{2\pi} \nabla_{\mathbf{x}} \cdot (I(x, y, z) \nabla_{\mathbf{x}} \varphi(x, y, z)) \quad (10.1)$$

where λ is the wavelength, $I(x, y, z)$ is the intensity, $\varphi(x, y, z)$ is the phase, $\partial I(x, y, z)/\partial z$ denote the change of the intensity along the z -axial direction, \mathbf{x} denotes the 2D coordinate vector (x, y) , and $\nabla_{\mathbf{x}}$ is the 2D gradient operator over \mathbf{x} . Eq. (10.1)

is known as the traditional TIE, which is derived under the condition of completely coherent optical field.

When the optical field to be measured is a quasi-monochromatic partially coherent light, TIE has been proven to be connected to the Wigner distribution function (WDF) [15]. For the case of partially coherent optical field in space, the generalized TIE can be well expressed as

$$\frac{\partial I(\mathbf{x})}{\partial z} = -\lambda \nabla_{\mathbf{x}} \cdot \int \mathbf{u} W_{\omega}(\mathbf{x}, \mathbf{u}) d\mathbf{u} \quad (10.2)$$

where \mathbf{u} is the spatial frequency coordinate corresponding \mathbf{x} , the 2D coordinate vector (x, y), and $W_{\omega}(\mathbf{x}, \mathbf{u})$ represents the WDF of monochromatic component (characterized by the optical frequency $\omega = c/\lambda$, where c is the speed of light and λ is the wavelength) in the partially coherent optical field.

When monochromatic optical waves are also completely coherent in space, the phase gradient under partially coherent case is related to the first-order conditional spatial frequency moment of the WDF as follows

$$\frac{\int \mathbf{u} W(\mathbf{x}, \mathbf{u}) d\mathbf{u}}{\int W(\mathbf{x}, \mathbf{u}) d\mathbf{u}} = \frac{1}{2\pi} \nabla_{\mathbf{x}} \varphi(\mathbf{x}) \quad (10.3)$$

Equation (10.3) can be substituted into Eq. (10.2) to obtain

$$\frac{\partial I(\mathbf{x})}{\partial z} = -\frac{\lambda}{2\pi} \nabla_{\mathbf{x}} \cdot (I(\mathbf{x}) \nabla_{\mathbf{x}} \varphi(\mathbf{x})) \quad (10.4)$$

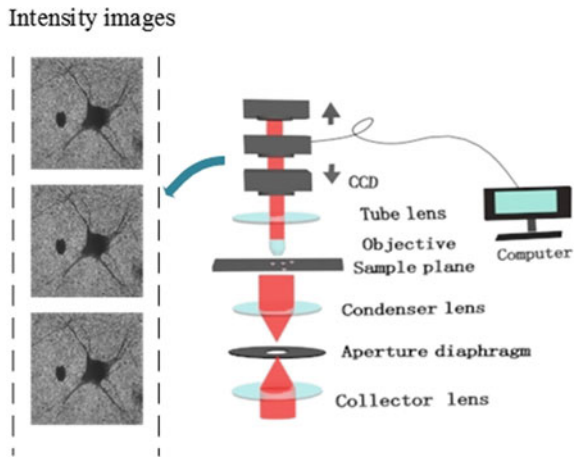
It can be seen that the forms of Eqs. (10.4) and (10.1) are the same. Therefore, phase also can be quantitatively retrieved by TIE in partially coherent optical field.

10.3 Phase Retrieval in the Microscopic Single Field of View System

10.3.1 Theory

The schematic diagram of the optical experiment system in this paper is shown in Fig. 10.1. Because the intensity of the main light source of the microscope is symmetrically distributed on the optical axis and the circular aperture diaphragm of the illumination is strictly aligned on the optical axis, the traditional Kohler illumination we use in the microscopic system satisfied the null frequency moment condition [15]. It is indicated that the traditional TIE phase retrieval method can be applicable to microscopes with partially coherent illumination.

Fig. 10.1 Schematic diagram of experimental system

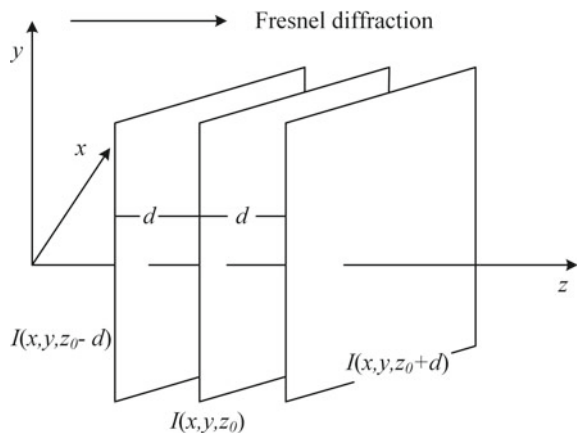


In this paper, in order to obtain the intensity images of different defocus distances, we use the C adapter ring set at the eyepiece tube interface to connect the CCD detection device, use the precision translation stage to obtain the focused image, and then, carry out the quantitative movement, so that CCD can capture the intensity images of different defocus distances.

It can be concluded from Eq. (10.4) and Fig. 10.1 that the intensity partial derivative $\partial I/\partial z$ of the corresponding z value needs to be accurately estimated when the phase is solved by TIE. As the light intensity variation diagram is shown in Fig. 10.2, the spacing between the over-focus diagram and the focus diagram, the focus diagram and the under-focus diagram are all d .

We first write the Taylor expansion of the intensity $I(x, y, z_0 + d)$

Fig. 10.2 Light intensity variation diagram



$$I(x, y, z_0 + d) = I(x, y, z_0) + \frac{\partial I}{\partial z}d + \frac{\partial^2 I}{\partial z^2} \frac{d^2}{2!} + O(d^3) \quad (10.5)$$

If we simply approximate the value of $\partial I/\partial z$ by $[I(x, y, z_0 + d) - I(x, y, z_0)]/d$, the error is large enough to reach the first order of the distance $O(d)$, and then Taylor's expansion of the intensity $I(x, y, z_0 - d)$ to get

$$I(x, y, z_0 - d) = I(x, y, z_0) + \frac{\partial I}{\partial z}(-d) + \frac{\partial^2 I}{\partial z^2} \frac{(-d)^2}{2!} + O((-d)^3) \quad (10.6)$$

The result of Eq. (10.5) minus Eq. (10.6) is

$$I(x, y, z_0 + d) - I(x, y, z_0 - d) = \frac{\partial I}{\partial z}(2d) + O(d^2) \quad (10.7)$$

So, the intensity change $\partial I/\partial z$ can be approximately replaced by a finite difference of the intensity distributions in plane perpendicular to the propagation axis z at distance of $z_0 + d$ and $z_0 - d$. Meanwhile, the error is reduced to the second order of the distance $O(d^2)$ [16]. Thus, Eq. (10.4) can be rewritten as

$$\frac{I(x, y, z_0 + d) - I(x, y, z_0 - d)}{2d} = -\frac{\lambda}{2\pi} \nabla_{\mathbf{x}} \cdot (I(\mathbf{x}) \nabla_{\mathbf{x}} \varphi(\mathbf{x})) \quad (10.8)$$

Although the defocus image can be obtained by quantitatively moving the C adapter ring, dislocation (rotation and translation) is inevitably introduced, which affects the clarity of the phase recovery results. Therefore, we solve this problem by using the image registration algorithm based on Harris operator [17–19]. The in-focus image is used as the reference image, and the defocus image is used as the image to be registered.

Firstly, Harris feature points are detected and extracted from the reference image and the defocus image. Then, the feature points between the two images are matched, and the degree of normalized cross-correlation [20] between the feature points on the two matching images is compared. Finally, the spatial transformation model \mathbf{T}' between images is solved according to the matching points. θ' is the rotation angle, and t'_x, t'_y are the translation parameter.

$$\mathbf{T}' = \begin{bmatrix} \cos \theta' & -\sin \theta' & t'_x \\ \sin \theta' & \cos \theta' & t'_y \\ 0 & 0 & 1 \end{bmatrix} \quad (10.9)$$

The defocus image is geometrically transformed according to the transformation model \mathbf{T}' , so that the misaligned defocus image and the in-focus image are placed in the same spatial position. $I_{rg}(x, y, z_0 - d)$ and $I_{rg}(x, y, z_0 + d)$ are the under- and over-focus images after registration, respectively.

$$\begin{cases} I_{rq}(x, y, z_0 - d) = \mathbf{T}'(I(x, y, z_0 - d)) \\ I_{rg}(x, y, z_0 + d) = \mathbf{T}'(I(x, y, z_0 + d)) \end{cases} \quad (10.10)$$

10.3.2 Experiment and Result Analysis

To validate the implemented method, a conventional microscope (Olympus CX23) is used, and the experimental setup is shown in Fig. 10.3. The partially coherent light source is a built-in LED, using a green filter with a center wavelength of 532 nm and FWHM of 22 nm to ensure the partially coherent light in monochrome. A motor neuron cell smear is used as sample, and a 10x infinity corrected objective is used for imaging. Therefore, the beam derived from LED light was filtered by a green filter and then passed through the sample to be measured after aperture diaphragm and the condenser lens. After the Fourier transform of the objective lens, the infinite parallel beam enters the tube lens, and the tube lens is used for imaging. Finally, the CCD connected to the eyepiece tube receives and records the intensity information.

As shown in Fig. 10.3, the inside of the red circle is a C adapter ring which is used to adapt the connection between the CCD and the eyepiece tube interface. The defocus distance can be changed by moving the adapter ring, and the range can be adjusted from 0 to 20 mm.

Fig. 10.3 Device diagram of real experimental



In this experiment, the defocus distance is set as 3 mm. The three intensity images are captured by CCD according to the aforementioned approach, as shown in Fig. 10.4a–c. Reference intensity image is shown in Fig. 10.4e. Figure 10.4d, f show the over-focus image and the under-focus image after registration, respectively. Moreover, it can be clearly observed from Fig. 10.5a that there is a misalignment between Fig. 10.4a, c, which causes a ghost of the phase result shown in Fig. 10.5c. Figure 10.5b shows the intensity difference distribution in Eq. (10.8). We can also be clearly observed from Fig. 10.5d that the registered over-focus image has a good overlap with the registered under-focus image, and the misalignment position relationship shown in Fig. 10.5a is well improved. Then, the intensity difference distribution shown in Fig. 10.5e also demonstrates the corrective ability of the registration. Figure 10.5f shows the phase recovery result of the corrected over-focus image and the corrected under-focus image. The phase quality of the cells is improved, the phase misalignment is reduced, and the clarity is improved. Therefore, the mentioned method can have a well-defined phase retrieval result for the cells and greatly change the defocus distance.

In order to further illustrate the applicability of the proposed method, a more complex object is tested. A common optical resolution plate is used as sample. Due to difference in experimental materials, we use a 4x infinity corrected objective to capture optimal images, and the defocus distance is 2 mm. As with the above steps, Fig. 10.6a–c are the three intensity images captured in the experiment. Figure 10.6d shows the intensity distribution comparisons along the lines in Fig. 10.6a–c. Figure 10.6e is the

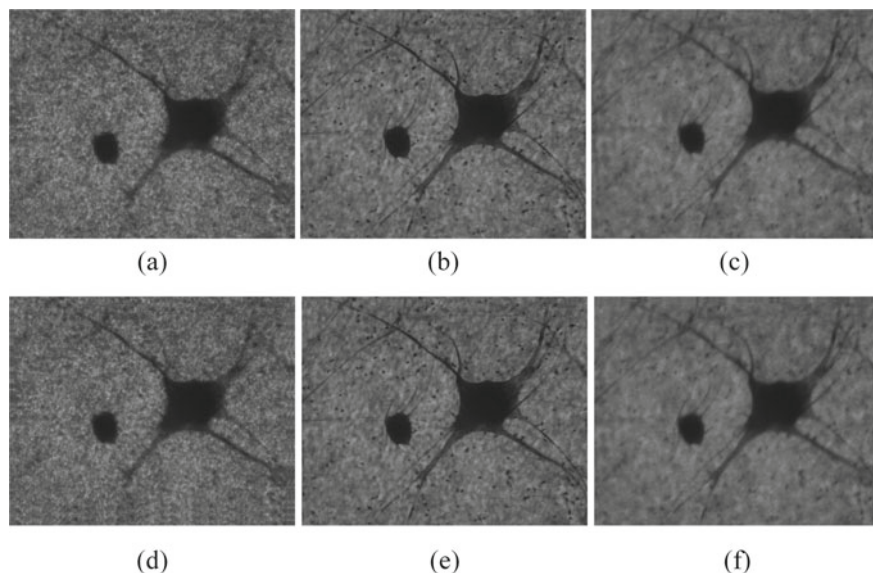


Fig. 10.4 Comparison diagram of intensity image. **a** Unregistered over-focus intensity image; **b** in-focus intensity image; **c** unregistered under-focus intensity image; **d** registered over-focus intensity image; **e** reference intensity image; **f** registered under-focus intensity image

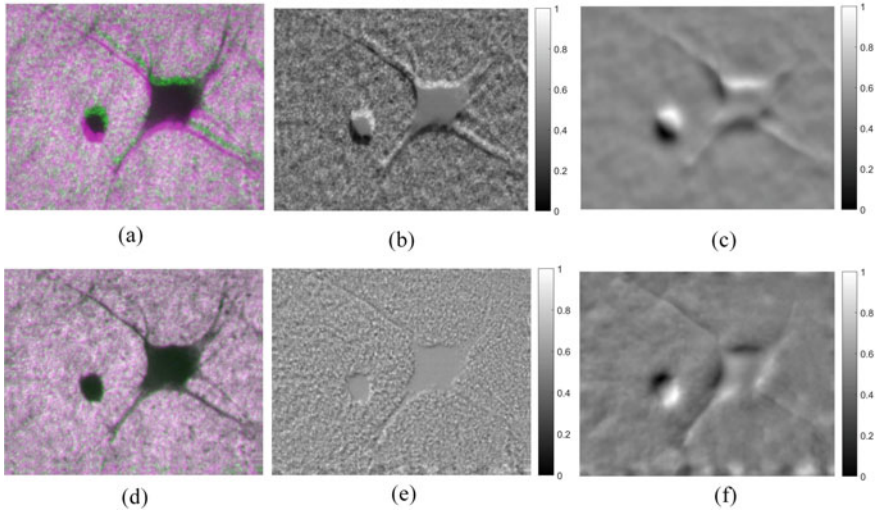


Fig. 10.5 Real experimental results of motor neuron cells. **a** Unregistered over-focus image and unregistered under-focus image misalignment; **b** unregistered intensity difference image; **c** unregistered phase via TIE; **d** registered over-focus image and registered under-focus position; **e** registered intensity difference image; **f** registered phase via TIE

geometric positional relationship of the lines in the image after registration. The ability of image correction can be observed quantitatively from Fig. 10.6d, e. Figure 10.6f shows the poor phase retrieval result before registration, with a severe ghost and a blurred edge. Figure 10.6g is the phase retrieval result after registration, as observed in Fig. 10.6, it can be observed that our proposed approach can also successfully achieve an accurate retrieval phase in complex case.

10.4 Conclusion

In this paper, the existing interface of the microscope is fully considered, and the optical characteristics of the microscope are combined with the C adapter ring to make a reasonable movement change of the defocus distance. Two sets of experiments verify the convenience of changing the defocus distance, and errors in the phase recovery result caused by movement can be compensated by registration.

The National Natural Science Foundation of China (Nos. 61301296, 61501001, 61605002, 61401001), Natural Science Foundation of Anhui Province (No. 1608085QF161), Natural Science Project of Anhui Higher Education Institutions of China (Nos. KJ2016A029, KJ2015A114).

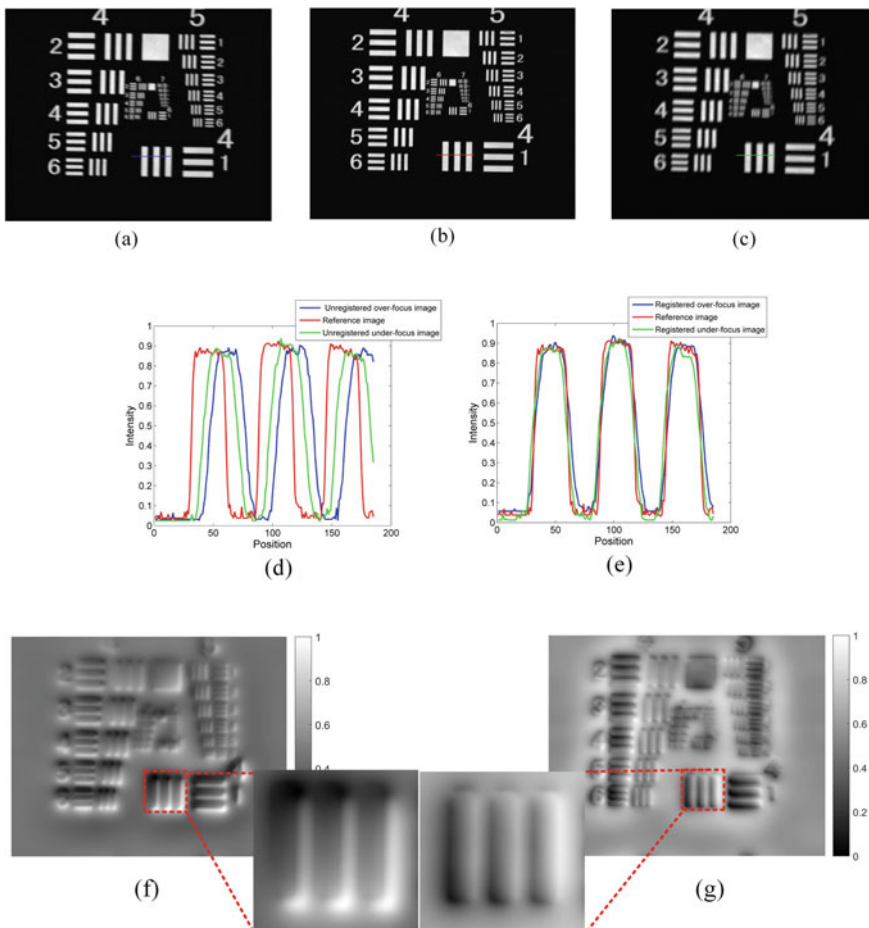


Fig. 10.6 Real experimental results of optical resolution plate. **a**, **b**, and **c** Captured over-focus image, in-focus image, and under-focus image; **d** intensity comparisons along the lines in **(a)**, **(b)**, and **(c)**; **e** comparison of intensity image lines after registration; **f** and **g** phase comparison before and after registration

References

1. Falch, K.V., Lyubomirsky, M., Casari, D., et al.: Zernike phase contrast in high-energy x-ray transmission microscopy based on refractive optics. *Ultramicroscopy* **184**, 267–273 (2018)
2. Lee, J., Rozell, C. J.: Precision cell boundary tracking on DIC microscopy video for patch clamping. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 1048–1052 (2017)
3. Park, Y.K., Depeursinge, C., Popescu, G.: Quantitative phase imaging in biomedicine. *Nat. Photonics* **12**(10), 578 (2018)
4. Schwarz, C.J., Kuznetsova, Y., Brueck, S.R.J.: Imaging interferometric microscopy. *Opt. Lett.* **28**(16), 1424–1426 (2003)

5. Pastorek, L., Venit, T., Hozák, P.: Holography microscopy as an artifact-free alternative to phase-contrast. *Histochem. Cell Biol.* **149**(2), 179–186 (2018)
6. Teague, M.R.: Deterministic phase retrieval: A Green's function solution. *JOSA* **73**(11), 1434–1441 (1983)
7. Zuo, C., Chen, Q., Sun, J.-S., et al.: Non-interferometric phase retrieval and quantitative phase microscopy based on transport of intensity equation: A review. *Chin. J. Lasers* **43**, 219–249 (2016)
8. Hu, C.-F., Popescu, G.: Quantitative phase imaging (QPI) in neuroscience. *IEEE J. Sel. Top. Quantum Electron.* **25**(1), 1–9 (2018)
9. Ishizuka, A., Mitsuishi, K., Ishizuka, K.: Direct observation of curvature of the wave surface in transmission electron microscope using transport intensity equation. *Ultramicroscopy* **194**, 7–14 (2018)
10. Hu, Y., Chen, Q., Feng, S.-J., et al.: A new microscopic telecentric stereo vision system—Calibration, rectification, and three-dimensional reconstruction. *Opt. Lasers Eng.* **113**, 14–22 (2019)
11. Cheng, H., Gao, Y.-L., Xu, S.-S., Deng, H.-L., Wei, S.: Non-interference phase retrieval algorithm with two wavelength illumination. *Acta Photonica Sin.* **47**(4), 132–140 (2018)
12. Barty, A., Nugent, K.A., Paganin, D.M., et al.: Quantitative optical phase microscopy. *Opt. Lett.* **23**(11), 817–819 (1998)
13. Li, J.-J., Chen, Q., Sun, J.-S., et al.: Multimodal computational microscopy based on transport of intensity equation. *J. Biomed. Opt.* **21**(12), 126003 (2016)
14. Tian, X.-L., Yu, W., Meng, X., et al.: Real-time quantitative phase imaging based on transport of intensity equation with dual simultaneously recorded field of view. *Opt. Lett.* **41**(7), 1427–1430 (2016)
15. Zuo, C., Chen, Q., Tian, L., et al.: Transport of intensity phase retrieval and computational imaging for partially coherent fields: The phase space perspective. *Opt. Lasers Eng.* **71**, 20–32 (2015)
16. Ishizuka, K., Allman, B.E.: Phase measurement in electron microscopy using the transport of intensity equation. *Microsc. Today* **13**(3), 22–25 (2005)
17. Zhang, J.-S., Zhang, H.-M., Luo, Y.-T., et al.: Image registration method for improved Harris corner detection. *Laser Infrared* **47**(2), 230–233 (2017)
18. Haggui, O., Tadonki, C., Lacassagne, L., et al.: Harris corner detection on a NUMA manycore. *Futur. Gener. Comput. Syst.* **88**, 442–452 (2018)
19. Liu, J.-P., Yang, Z., Huo, H., et al.: Checkerboard image processing under uneven illumination for robust Harris corner detection in camera calibration. In: Tenth International Conference on Digital Image Processing (ICDIP 2018), vol. 10806, (International Society for Optics and Photonics, 2018), pp. 1080664
20. Wang, C.-Y., Chen, J.-B., Chen, J.-S., et al.: Unmanned aerial vehicle oblique image registration using an ASIFT-based matching method. *J. Appl. Remote Sens.* **12**(2), 025002 (2018)

Chapter 11

Graphic QR Code with the Second Hidden QR Code by Codeword Rearrangement



Yi-Wei Juan, Tzren-Ru Chou, Chun-Shien Lu, and Hsi-Chun Wang

Abstract The development of science and technology has brought about tremendous changes in both the virtual and real worlds. The quick response code (QR code), a kind of two-dimensional barcode, has become an important interface between the two, playing the roles of advertising, marketing, transaction payment, and information authentication. The traditional QR code only has black-and-white information dots. It is difficult to distinguish the content from the appearance and also is easy to be copied. Therefore, for the protection of information and convenient interpretation, how to enhance the security and beauty of the QR code has also been studied by researchers all over the world. Therefore, this study proposes a double-encrypted graphic QR code to modify the previous halftoning technology, to enhance its aesthetics and to perform recognition analysis. The digital halftoning technique is used to hide the second QR code in the cover QR code, and the graphic QR code of the 41×41 modules (the sixth version) is obtained. The fourth version QR code has 33×33 modules. The codeword data of the fourth version QR code is hidden around the graphic QR's codeword, which did not affect the central area of the QR code, and its image quality was enhanced. Each module in the graphic QR code is divided into 3×3 sub-modules, and the explicit 41×41 QR code information is embedded in the middle of the 3×3 sub-modules. The 33×33 QR codeword information, according to the pseudo-random number generated by the key, is hidden in one of the other eight sub-modules, and the remaining seven sub-modules were used to display the gradation of the cover image. To study the recognition, the graphic QR code is outputted at four different sizes and then all of them are scanned for data recognition. The research results show that the modified double-encrypted QR code has better image quality, and the larger size of the QR code has better recognition rate. In addition to the double hidden information, the graphic QR code generated by this research could achieve anti-counterfeiting function and be aesthetically pleasing. In

Y.-W. Juan · T.-R. Chou · H.-C. Wang (✉)
Department of Graphic Arts and Communications, National Taiwan Normal University, Taipei,
Taiwan, Republic of China
e-mail: hsiwang@ntnu.edu.tw

C.-S. Lu
Institute of Information Science, Academia Sinica, Taipei, Taiwan, Republic of China

the future, double-encrypted graphic QR code could also be used to various image value-added applications that require halftone images.

11.1 Introduction

With the rapid development of modern technology and the Internet, various information exchanges are happening all the time. Two-dimensional barcodes play an important role in it. QR code is the most commonly used two-dimensional barcode. However, the traditional QR code is not beautiful enough, and it also has information security issue. Scholars have done researches in the beautification of QR code to make it both aesthetically pleasing and retaining the information readability. Currently, beautification of QR code is divided into two categories, one is for continuous tone color or grayscale graphic QR code [1–3], and the other is halftone-based graphic QR code [4]. In the study of continuous tone color graphic QR code, the QR code's information point will interfere with the visual quality of the image. Therefore, some scholars have studied the characteristics of the QR code error correction mechanism and the codeword for mask processing. It makes the image of the key areas to have better image quality [5, 6]. The aesthetics of halftone graphic QR code is worse than that of continuous tone color graphic QR code, but there are many applications where halftone graphic QR code has to be used (e.g., printing or laser perforation). The halftone QR code still has its unique research value.

In terms of security, there is currently a double-encrypted QR code such as Security QR code (SQRC) which is used to increase the security of the QR code. A study about double hidden double-encrypted QR code was proposed by Kuan et al. [7] and some researches used infrared to hide the message in the QR code [8, 9], but there was no research which applied the codeword characteristics on halftone graphic image. Thus, this study combined the QR code's codeword feature with both digital halftoning technique and double-encrypted QR code to develop a modified double-encrypted graphic QR code.

11.2 Related Works

In this study, the two-dimensional barcode and the cover image are combined to form a two-dimensional graphic barcode, and the pseudo-random random number is used to hide other two-dimensional barcode information into the two-dimensional graphic barcode. Decoding with a mobile phone can solve the information of the explicit QR code. Using the correct key, the implicit QR code information can be obtained. It can improve the anti-counterfeiting function and information security of the QR code. This section will be divided into two parts, first introducing the research of QR code and related beautification methods, and secondly, introducing the multiple information hiding techniques related to QR code.

Barcode was originally invented in 1949 by Woodland and Silver in the USA to manage food-related equipment. The data is stored in black-and-white lines or squares. When using the decoder, lines or squares can be converted into stored information in 0 s and 1 s. The common barcodes are mainly divided into two categories: “One-dimensional barcode” and “Two-dimensional barcode.” One-dimensional barcode can only store data in one direction, so the data capacity is limited. Two-dimensional barcode can record information in both horizontal and vertical directions. In addition to large data storage, it can also be quickly recognized by machines. Among different types of two-dimensional barcode, QR code, invented by Denso Wave Inc. in Japan in 1994 [10], is the most popular bar code specification. Quick response code (QR code, Fig. 11.1a) also has an error correction mechanism, so that the QR code can still be read by the machine under the attacks (damages, unexpected scratches, or stains) within error correction capability. Now, many companies have used the QR code error correction mechanism to put trademarks or images into the QR code. Even if the QR code itself loses some information points, it still can be read by a normal reader (Fig. 11.1b).

The QR code has 40 versions, from version 1 to version 40. Version means different QR code sizes, such as 21×21 , 25×25 , 29×29 , ..., 177×177 . The larger the version, the more information can be stored. The relationship between modules (M) and version (V) can be formulated as $M = 17 + 4V$. The QR code’s error correction mechanism has 4 levels, namely L (7%), M (15%), Q (25%), and H (30%). In order to obtain double encryption, the level H is chosen to study the decoding performance. The QR code storage information and error correction mechanism are encoded by codeword. Each codeword is composed of data of 8 bits. One bit of data (0 or 1) is a

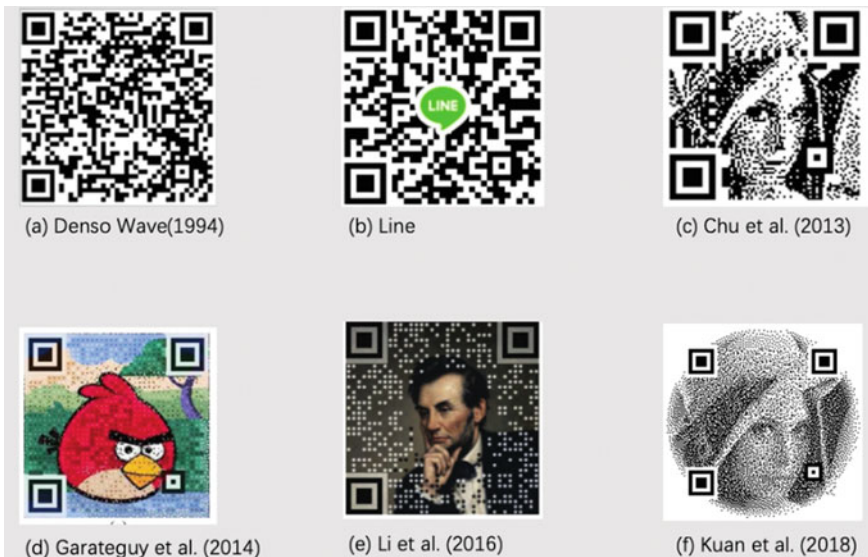


Fig. 11.1 Related graphic QR code researches

module in black-and-white dot, and the error correction mechanism for codeword is based on Reed–Solomon’s Code. Each version of the QR code has a different number of codewords. The larger version has more codewords, and the error correction level represents the number of codewords that can be misjudged. The H-level used in this study means that even 30% of the number of codeword are damaged or misjudged but it still can recover the original stored information. The QR code used in this study (41×41 and 33×33) has 172 and 100 codewords, respectively.

The research on QR code beautification is shown in (Fig. 11.1c–f). Chu et al. proposed the method of graphic QR code, which uses halftone techniques to incorporate QR code and cover image [4]. Each QR code module is divided into 3×3 sub-modules, and the original QR code information is hidden in the middle of the 3×3 sub-modules, and the remaining eight sub-modules are filled with black-and-white dots (Fig. 11.1c). Those halftone dots make the image gradation so that it can have both the original image information and the QR code information. Garateguy et al. proposed a full-frame color graphic QR code algorithm [1]. By inputting the color image, QR code, and mask, overall optimization and the interpolation method are used to combine the cover image and QR code information (Fig. 11.1d). Li et al. proposed a method of masking codeword and performing pixel analysis on the background image to make the image’s region of interest not affected by the original QR code information [5]. And then the pixels are analyzed to select the black-and-white points where the QR information would not be implanted, so that the user can generate the colorful graphic QR code by in an aesthetically pleasing manner (Fig. 11.1e). Kuan et al. [7] also proposed a double-encrypted QR code using halftone technique, which contains both explicit and implicit QR code with the same size. It can improve the aesthetics, anti-counterfeiting feature, and security of the QR code (Fig. 11.1f).

Regarding the multiple message hiding of QR code, Security QR code (SQRC) is a QR code encryption application proposed by Denso Wave Inc. The information in SQRC can be divided into two parts: public and private. When this type of encrypted QR code is decoded by a general scanner (QR code Reader), the public data can be seen. Only the specific SQRC scanner or software can interpret the private/encrypted data with a secret key, similar to a set of passwords. Japanese scholar Nobuyuki proposed a method of double-coding QR code using ordinary black ink and simulated black formed by mixing C (cyan), M (magenta), and Y (yellow) inks [8]. The hidden information can be obtained by the optical properties of the ink under the infrared detection. By using the characteristics of different inks, even if the same QR code is viewed under visible light, infrared illumination may produce different coded images. Therefore, the explicit and implicit data in double-encoding two-dimensional bar codes can be detected in visible light and infrared light. Wang et al. proposed a way to use information hiding and error diffusion techniques to hide two QR code information in the graphic QR code [9]. In order to avoid the interference between explicit and implicit QR code during detection, the black ink component of the explicit QR code is replaced by C (cyan), M (magenta), and Y (yellow) inks. The implicit QR code information is hidden in the K (black) ink of CMYK mode by infrared detection to enhance its security feature.

11.3 Methods

The double-encrypted QR code of this study used the halftone technique to hide the QR code codeword information of the fourth version of 33×33 modules into the graphic QR code of the sixth version of 41×41 modules. The method of generating double-encrypted QR code is based on the algorithm proposed by Kuan et al. [7] and we modify it to obtain better image quality. The methods of this study would be described separately in terms of image quality, different output size, and recognition error analysis. The double encryption method we used was generated by halftoning technique and error diffusion methods. The difference between the method by Kuan et al. and the proposed modified method of double-encrypted QR code is shown in Fig. 11.2.

It could be seen by the eye that the modified double-encrypted graphic QR code is better than Kuan's method to match the original image, and the using of the peak-signal noise ratio (PSNR) to evaluate the image quality will be carried out. In this study (shown in Fig. 11.3), we selected a color image to make the double-encrypted QR code. Referring to the proposed method by Kuan et al., we chose a 41×41 QR code as the explicit QR code of the double-encrypted QR code. Then, we selected 33×33 QR code to be processed and used it to be the double-encrypted QR code. The fourth version of the 33×33 QR code contains information in the codeword. The 33×33 QR has 100 codewords. The 172 codewords of the explicit 41×41 QR code were encoded according to the method of Fig. 11.4 to form a modified double-encrypted QR code, but the central region of interest area did not have the QR 33×33 codeword data, which had less influence on the cover image display. We selected four images (Fig. 11.5) to create the modified double-encrypted QR code, comparing the image quality of the original double-encrypted QR code of the four different images to the modified double-encrypted QR code of this study. The output sizes of four QR codes (Figs. 11.6 and 11.7) are 2.8×2.8 cm (sample #1),

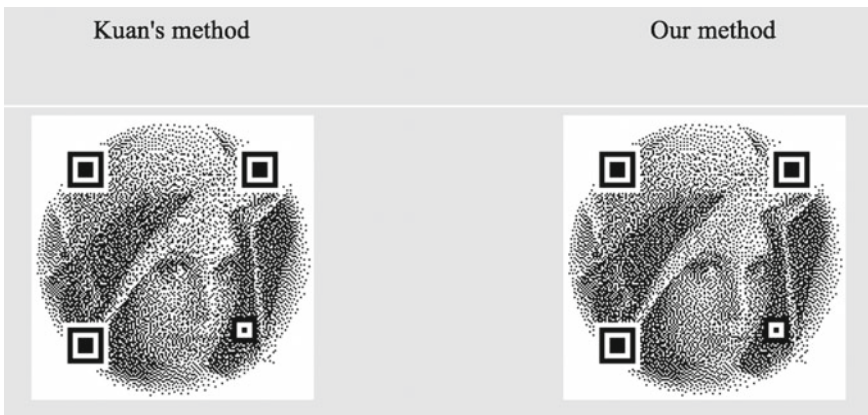


Fig. 11.2 Kuan et al. and our research about double-encrypted QR codes

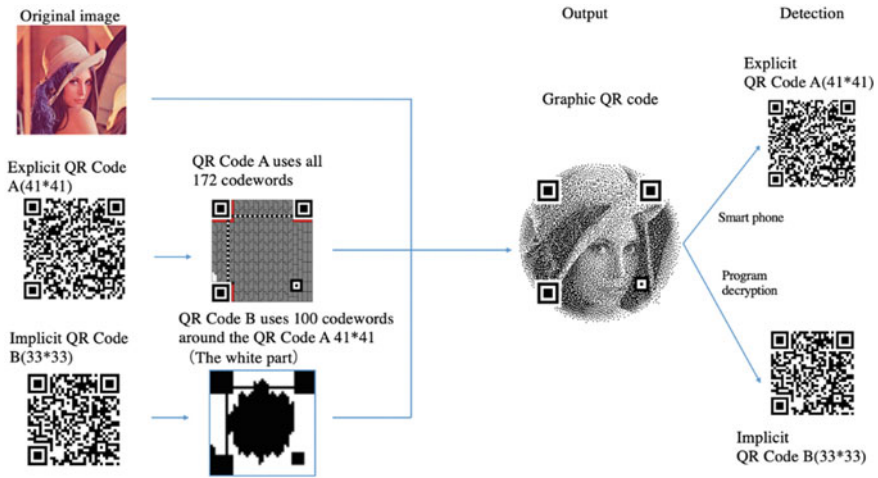
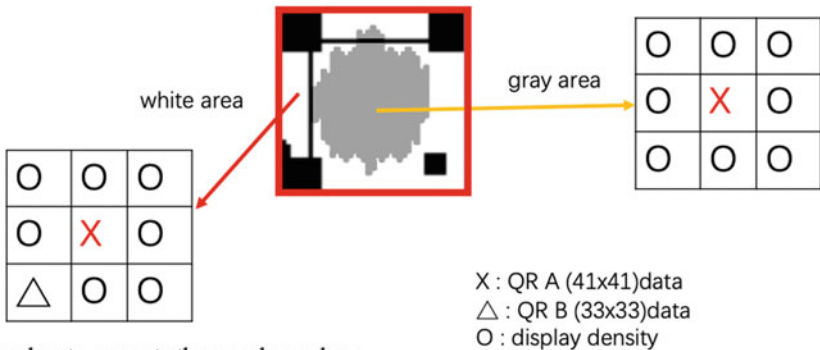


Fig. 11.3 A schematic diagram of research process



Using a key to generate the pseudo-random number and to select one of the eight position to hind QR B (33x33) data

Fig. 11.4 Data of the QR A and the QR B in the proposed graphical QR code

2.4 × 2.4 cm (#2), 1.8 × 1.8 cm (#3), and 1.4 × 1.4 cm (#4), receptively. Then, after scanning at 600 dpi, we compared the data of sub-module and codeword in the output double-encrypted QR code and the original digital data in image files to evaluate the recognition rate of sub-modules and codewords during the decoding of the printed graphic QR code.











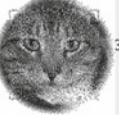






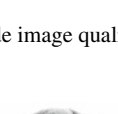
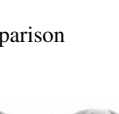

Image quality assessment		PSNR		PSNR		PSNR		PSNR
Original grayscale								
FM(Frequency Modulation)		34.18dB		34.67dB		34.38dB		34.85dB
Explicit 41*41 QR Implicit 33*33 QR (our method)		31.80dB		31.52dB		31.97dB		30.39dB
Explicit 41*41 QR Implicit 41*41 QR (Kuan's method)		31.66dB		31.35dB		31.96dB		30.13dB

Fig. 11.5 Double-encrypted QR code image quality comparison

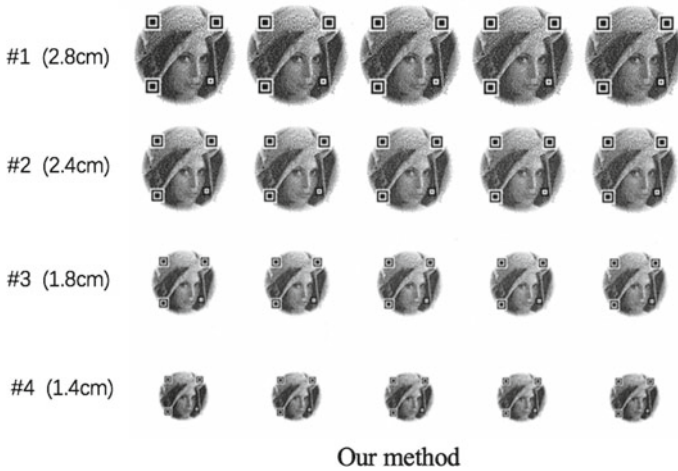


Fig. 11.6 Different output sizes by our method

11.4 Results and Discussions

First, the image quality comparison was performed for the double-encrypted QR code of different cover images, and we compared the data proposed by Kuan et al. It could be seen that in the image with more details, the performance difference between the new double-encrypted QR code and that of the previous one is not very obvious. In general, our proposed method can have higher image quality, and image

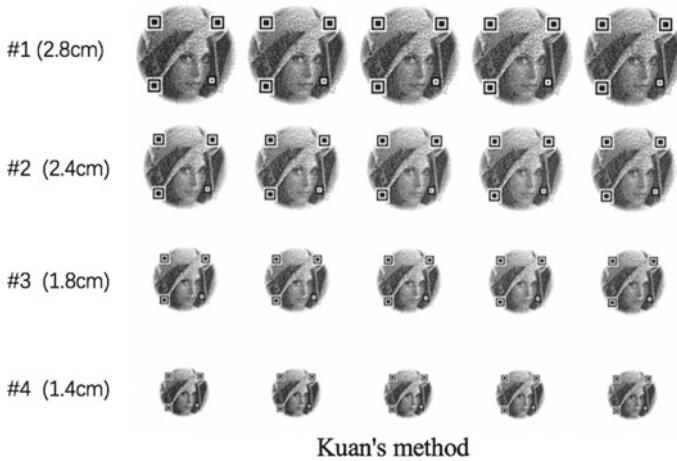


Fig. 11.7 Different output sizes by Kuan's method

with less detail had a bigger difference between the graphic QR codes of our method and the method of Kuan et al.

Methods of both the proposed and Kuan et al. were used to generate digital images which were outputted in four sizes for further recognition error analysis. Each size had been printed 5 copies for reproducible tests (Figs. 11.8 and 11.9). The error rate of both sub-modules and codewords were calculated. The physical output sizes are numbered #1 (2.8 cm), #2 (2.4 cm), #3 (1.8 cm), and #4 (1.4 cm). The bit (sub-module) error rate for all samples of both methods was kept within 20%. However, in the codeword, only the error rate of samples #1 and #2 remained within 30%. The codeword error rate of Kuan's sample #3 was more than 30% in the both explicit and implicit QR codes, which are very marginal for successful reading. The codeword error rate of sample #3 in the modified method was kept within the range of 30% for the explicit QR code, but the implicit QR code also had the error rate more than 30%, which may not be able to be recognized. The error rate of the double-encrypted QR code of sample #4 (1.4 cm) in both methods was out of the acceptable range, so there is no chance for the QR code to be interpreted.

Then, we analyzed the wrong recognition to understand the characteristics of error information. The errors can be classified into two cases, one is "black dots misjudged into white dots" (false white) and the other is "white dots misjudged into black dots" (false black) (Figs. 11.10 and 11.11). As QR code recognition error occurs, the situation of white dots being judged as black dots happened more than the black dots being judged as white dots. It probably was caused by the dot gain during the printing process.

The findings of this study could be used as a reference for generating printed double-encrypted QR code in the future. Using codeword rearrangement could improve image quality for double-encrypted QR code, and the information retention was also better than the Kuan's method. The image quality in the region of

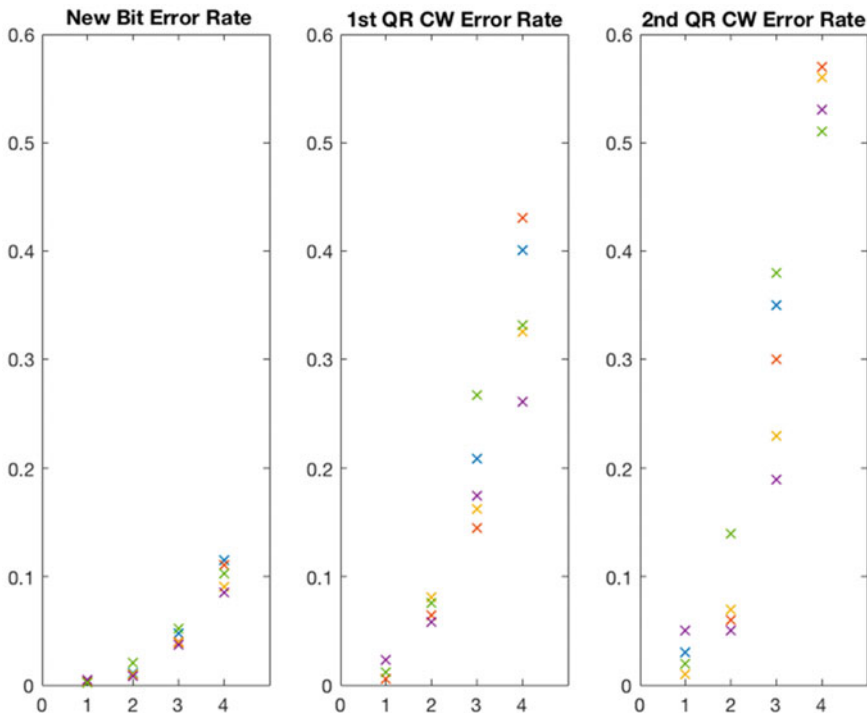


Fig. 11.8 Bit and codeword (CW) error rate for the modified method at different sizes

interest area was also higher by the proposed method. And it is not necessary to sacrifice the error correction capability to support visual image quality like the methods of [5, 6]. According to the data of image quality analysis, it could be found that for an image with more details, the image quality difference of both methods is smaller. The error positions of the data point after output were mostly false black. The finding could be helpful as a reference to the design of graphic QR code. Since double-encrypted image is an important factor of affecting image quality and human eye observation, therefore, if it is possible to select the original image, it is better to select an image with less details and the region of interest area could be in the central part. The codeword error rate for samples #1 and #2 were within 30%, and it provided the user flexibility to choose proper size of printed graphic QR code. Thus, the generated double-encrypted QR code could maintain a better recognition rate and better image quality improvement. Also, it will have a visually pleasing look and increases the security of the double-encrypted QR code.

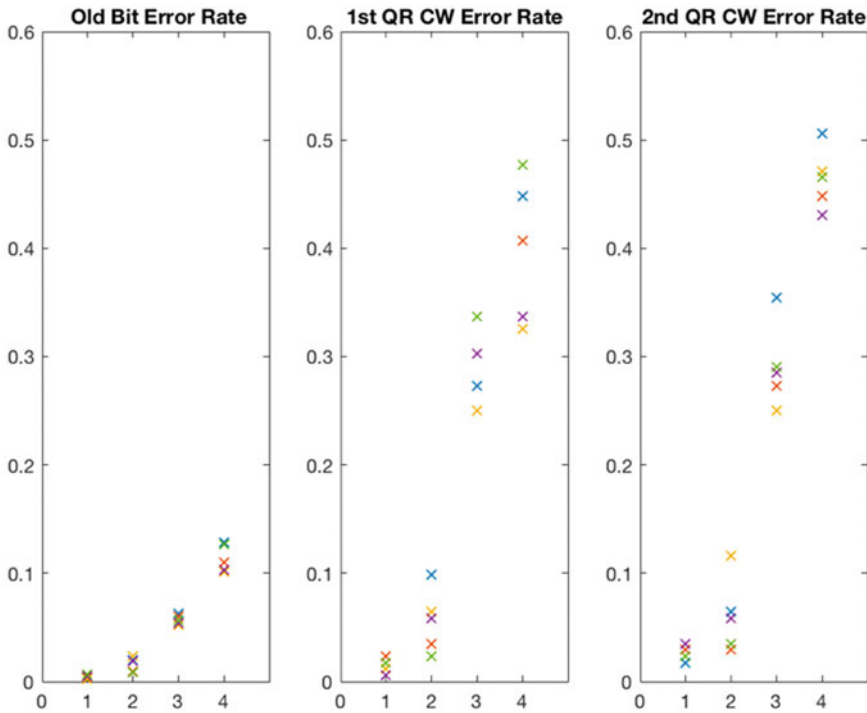


Fig. 11.9 Bit and codeword (CW) error rate for Kuan's method at different sizes

11.5 Conclusion

In this paper, the following conclusions could be drawn. First, in our modified method, the image quality of the double-encrypted QR code is significantly improved, especially in the case of low-spatial frequency components of the cover image. Both of the image quality measurement and the visual appearance are obviously improved. Second, in the QR code error rate analysis, the larger output size would lead better recognition; however, there is no significant difference between the previous and new methods. Third, in the case where the information point is misjudged, most of the errors occur in white dots being judged as black dots (false black), and the number of false white is less.

In terms of future suggestions, more versions of the double-encrypted QR code can be studied. In addition, researchers can also improve the algorithm for making double-encrypted QR codes to reduce the possible errors in the recognition. The use of the deep learning method is suggested to perform training analysis on the decoder to increase the recognition ability to decode the double-encrypted graphic QR code.

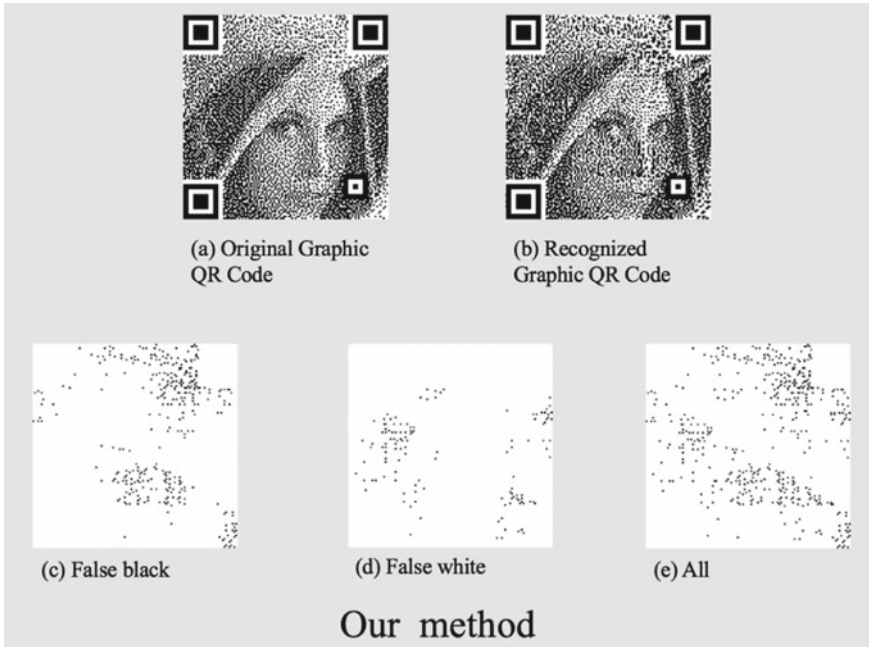


Fig. 11.10 Recognition error classification for our method

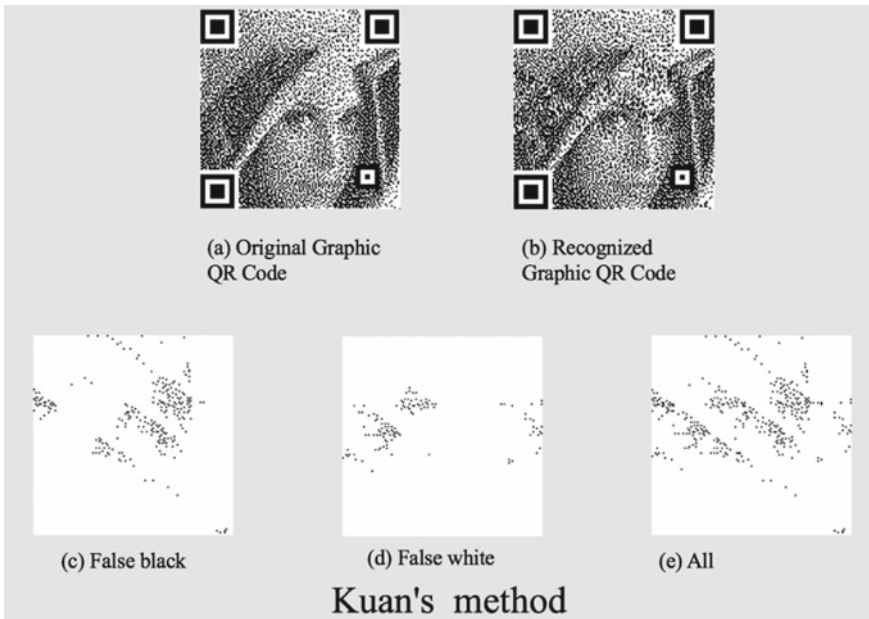


Fig. 11.11 Recognition error classification for Kuan's method

References

1. Garateguy, G.J., Arce, G.R., Lau, D.L., Villarreal, O.P.: QR images: optimized image embedding in QR codes. *IEEE Trans. Image Process.* **23**(7), 2842–2853 (2014)
2. Lin, S.S., Hu, M.C., Lee, C.H., Lee, T.Y.: Efficient QR code beautification with high quality visual content. *IEEE Trans Multimed* **17**(9), 1515–1524 (2015)
3. Lin, Y.H., Chang, Y.P., Wu, J.L.: Appearance-based QR code beautifier. *IEEE Trans Multimed* **15**(8), 2198–2207 (2013)
4. Chu HK, Chang CS, Lee RR, Mitra NJ (2013) Halftone QR codes. *ACM Trans Graph* 32(6): 217: 1–8
5. Li, L., Qiu, J., Lu, J., Chang, C.-C.: An aesthetic QR code solution based on error correction mechanism. *J. Syst. Softw.* **116**, 85–94 (2016)
6. Lin L, Wu S, Liu S, Jiang B (2017) Interactive QR code beautification with full background image embedding. In: Second international workshop on pattern recognition, pp 10443, 1044317. <https://www.doi.org/10.1117/12.2280282>
7. Kuan PC, Sun CT, Wang YM, Wang HC, Lu CS (2018) Visible graphic QR code with embedded invisible QR code to enhance anti-counterfeiting features. *Int J Electr, Electron Data Commun (IJEEDC)* 6(12): 17–22
8. Nobuyuki T (2017) Counterfeit detection by smartphone using double-encoded two-dimensional code. *Innov Mob Internet Serv Ubiquitous Comput*, 455–466. https://www.doi.org/10.1007/978-3-319-61542-4_43
9. Wang YM, Sun CT, Kuan PC, Wang HC, Lu CS (2018) Secured graphic QR code with infrared watermark. In: International conference on applied system innovation (ICASI'18), Chiba, Japan
10. Denso Wave Inc (2007) Secured QR code [Online]. Available: <https://www.denso-wave.com/en/system/qr/product/sqrc.html>

Chapter 12

A Novel Position-Shift Method of Double-Phase Fresnel Hologram for Synthesizing a Complex Fresnel Hologram



Chuan Shen, QinQin Zhu, QingQing Hong, Cheng Zhang, and Sui Wei

Abstract An ideal desired spatial light modulator that is capable of complex amplitude modulation will be one of the ultimate tools for holographic display. In this paper, an analytical method to overlap double-phase Fresnel holograms displayed on the phase-only spatial light modulators for synthesizing a complex Fresnel hologram is proposed. Two $4f$ configurations with the phase-only spatial light modulator inserted in the Fourier plane are employed. Different programmable blazed phase gratings can be easily implemented by encoding their phase functions onto the spatial light modulators. We theoretically analyze the implementation of the proposed system based on Fourier optics. The effects introduced by the discrete pixelated nature of the spatial light modulator are assessed. A $4f$ configuration of holographic display system based on two liquid crystals on silicons is designed to indicate the free position-shift of the hologram. Optical reconstructions yield satisfactory results.

12.1 Introduction

With the development of optoelectric technology, digital holographic display enjoyed a wide research. Many researchers set up varieties of holographic display systems to achieve the reconstruction of computer-generated holograms (CGHs) based on different spatial light modulators (SLMs), but there are still many challenges for utilizing digital media such as SLMs to realize holographic video display. The desired SLMs are supposed to have a smaller pixel pitch, a higher resolution, and the ability of complex arithmetic operations. In this paper, we concern the issues about controlling both the amplitude and phase of the wavefront. Unfortunately, the wavefront is

C. Shen (✉) · Q. Zhu · Q. Hong · C. Zhang · S. Wei
Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei 230601, China
e-mail: shenchuan@ahu.edu.cn

C. Shen
Key Laboratory of Modern Imaging and Display Technology of Anhui Province, Anhui University, Hefei 230601, China

complex-valued, wherever the state-of-the-art SLMs cannot modulate the amplitude and phase of the incident light simultaneously.

Even though the modulation capabilities of easily achievable digital micro-mirror device (DMD) or liquid crystal on silicon (LCoS) are limited, different approaches of complex amplitude modulation have been proposed. These approaches involve employing two SLMs for amplitude modulation or phase modulation [1], or employing the two different SLMs that are working in real-only configuration and in imaginary-only mode separately [2]; however, most of them require the pixel-to-pixel registration of the SLMs to achieve the complex modulation. Another consideration for saving number of SLMs is the single SLM configuration [3], where the holograms can be displayed side-by-side at the corresponding part of the SLM with a limited spatial resolution. Note that the diffraction efficiency is one of the critical parameters of the holographic display system, so encoding the complex amplitude information onto phase-only filters based on phase-only SLMs has already been reported in several literatures [4–6].

Another alternative approach to encode complex information is the double-phase hologram (DPH) [7, 8]. DPH method proposes to decompose an arbitrary complex number into the sum of two phase-only values with constant magnitude. Nowadays, the programmable property of the phase-only SLM enables it to represent the complex hologram cell by only two masks or pixels [9]. In another relative development, Liu et al. achieved the overlap of the holograms at the output plane of the $4f$ configuration system by using a sinusoidal grating [10]. Zhang et al. proposed a three-dimensional (3D) near-eye systems based on complex amplitude modulation [11]. As the techniques discussed above, the grating, one of diffractive optical elements (DOEs), plays an important role in the system. Benefiting from these methods, the purpose of this paper is not to use static gratings but to employ SLMs displayed with phase functions for providing a similar role, where we only focus on analyzing the implementation for the position-shift of hologram.

In this paper, an analytical method to display a complex Fresnel hologram by the phase-only SLM is proposed. Two $4f$ optical systems are employed to combine double-phase Fresnel holograms (DPFHs), and the dynamic DOEs (two different blazed phase gratings) generated by phase-only SLM are inserted in the Fourier plane. Our purpose is to control the position-shift of the DPFHs and then implement to overlap them for synthesizing a complex Fresnel hologram. Firstly, the principle of the proposed method that combines two phase holograms is introduced. Secondly, the numerical simulation is carried out to investigate the synthesis of a complex Fresnel hologram. Finally, a holographic display system based on two LCoSs is built to verify the position-shift of the hologram.

12.2 Principle of Complex Fresnel Hologram Synthesis

Considering the hologram reconstruction, 3D images can be understood as a result from Fresnel diffraction. The Fresnel diffraction or transform is directly related to

the paraxial or Fresnel approximation of the formalism describing light propagation between the optical field in the hologram plane and the field in the image plane. As shown in Fig. 12.1, phase holograms are displayed on the SLM plane. The field at distance z_c from the SLM can be expressed as,

$$U(\xi, \eta, z_c) = \frac{e^{ikz_c}}{i\lambda z_c} \exp\left[\frac{i\pi}{\lambda z_c}(\xi^2 + \eta^2)\right] \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left\{ U(x, y, 0) \exp\left[\frac{ik}{2z_c}(x^2 + y^2)\right] \right\} \exp\left[\frac{-i2\pi}{\lambda z_c}(\xi x + \eta y)\right] dx dy \quad (12.1)$$

where λ is the wavelength of the incident light, $k = 2\pi/\lambda$ is the wave number, and $U(x, y, 0)$ denotes the transmittance function of the SLM.

An equivalent expression for Eq. (12.1) can be given by,

$$U(\xi, \eta, z_c) = \frac{e^{ikz_c}}{i\lambda z_c} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U(x, y, 0) \exp\left\{\frac{ik}{2z_c}[(\xi - x)^2 + (\eta - y)^2]\right\} dx dy \quad (12.2)$$

It becomes a convolution integral, and the Fresnel impulse response is defined by,

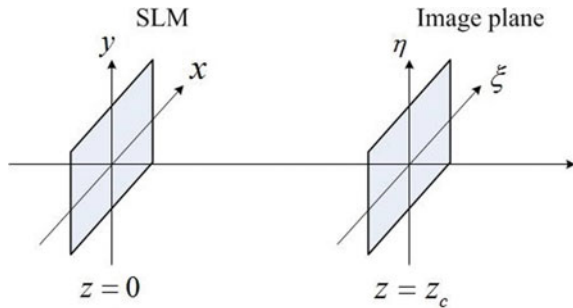
$$h(x, y) = \exp\left[\frac{i\pi}{\lambda z}(x^2 + y^2)\right] \quad (12.3)$$

So fast Fourier transform algorithms could be employed to compute the propagation [12], which is given in the following relation,

$$U(\xi, \eta, z_c) = \frac{e^{ikz_c}}{i\lambda z_c} \mathfrak{S}^{-1}\{\mathfrak{S}\{U(x, y, 0)\}\mathfrak{S}\{h(x, y)\}\} \quad (12.4)$$

where \mathfrak{S} denotes the Fourier transform and \mathfrak{S}^{-1} denotes the inverse Fourier transform.

Fig. 12.1 Schematic diagram of holographic reconstruction



It is a pity that the available SLM devices could not modulate both the phase and amplitude of the wavefront independently at the same time. The DPH encoding method makes it possible to display a complex hologram using a single phase-only SLM, as we know, any complex amplitude transmittance $t(x, y) = A \exp(i\Phi)$ can be decomposed into the sum of two vectors with constant magnitude of $1/2$ and different phase quantities, ϕ_1, ϕ_2 , the transmitted intensity remains constant while the phase shift is varied, i.e.,

$$t(x, y) = A \exp(i\Phi) = \frac{1}{2} \exp[i\phi_1(x, y)] + \frac{1}{2} \exp[i\phi_2(x, y)] \quad (12.5)$$

where the phases that displayed on neighboring pixel of the phase SLM are $\phi_1(x, y) = \Phi + \arccos A$, and $\phi_2(x, y) = \Phi - \arccos A$, respectively.

In solving the problem of complex modulation, we prefer to combine these DPFHs together. The system configuration that we used to realize this operation is shown in Fig. 12.2 (To make the analysis as simple as possible, we assume the phase-only SLMs selected work in the transmission mode.) The major idea of our method is as follow. Firstly, DPFHs, which could express an arbitrary complex optical field, are computed. Then, different from DPH method, the SLM1 is divided into two sub-parts along x axis: one half (Green color) is added the phase distribution $\phi_1(x, y)$ while the other right half (Yellow color) is added the other phase pattern $\phi_2(x, y)$. (Note that it also allows one to use two SLMs for displaying these holograms separately.) Furthermore, two $4f$ configurations with phase-only SLMs inserted in the Fourier planes are

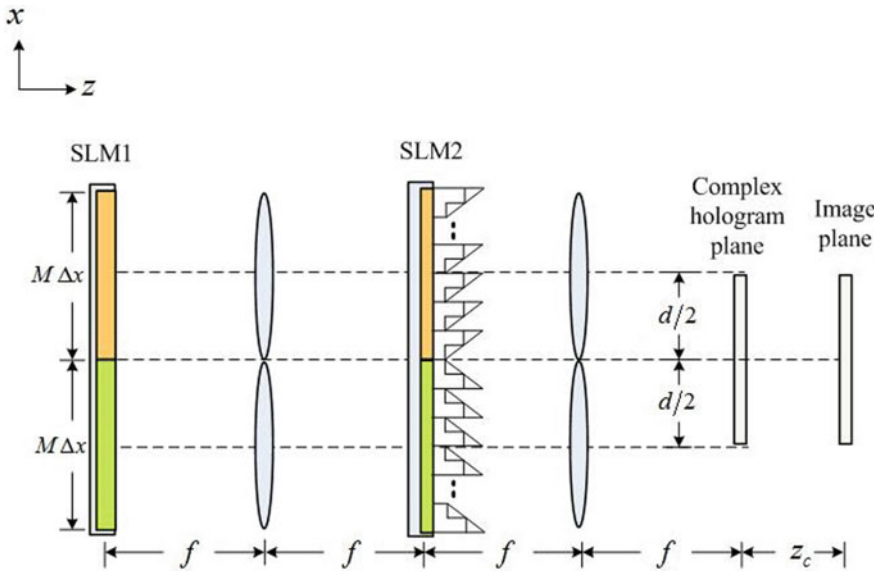


Fig. 12.2 Scheme of the proposed display system for synthesizing a complex Fresnel hologram

employed, and the SLMs of the proposed architecture comprise two different blazed phase gratings independently for the purpose of combining the DPFHs. Finally, the reconstruction in the image plane can be produced from the propagation of a complex Fresnel hologram.

In the above method, the key problem is to achieve the position-shift of hologram. As far as the SLM1 is divided into two sub-parts along x axis, the following formulation derived in the 1D case is straightforward. The amplitude transmittances of these DPFHs can be shown as,

$$\begin{aligned} t_1(x) &= \exp[\phi_1(x)] \\ t_2(x) &= \exp[\phi_2(x)] \end{aligned} \quad (12.6)$$

The Fourier transform of these functions is given by,

$$\begin{aligned} T_1(u) &= \mathfrak{F}\{t_1(x)\} \\ T_2(u) &= \mathfrak{F}\{t_2(x)\} \end{aligned} \quad (12.7)$$

Assume that two ideal blazed phase gratings are placed in the back focal plane. And the amplitude transmittances of the gratings within a single period can be represented by,

$$\begin{aligned} G_1(u) &= e^{i2\pi u/L}, 0 \leq u < L \\ G_2(u) &= e^{-i2\pi u/L}, 0 \leq u < L \end{aligned} \quad (12.8)$$

So the complex amplitude at the output plane and the modified spectrum of the input amplitude transmittance follows the relations,

$$\begin{aligned} \mathfrak{F}\left\{t_1\left(x + \frac{\lambda f}{L}\right)\right\} &= T_1(u)G_1(u) = \mathfrak{F}\{t_1(x)\}e^{i2\pi u/L} \\ \mathfrak{F}\left\{t_2\left(x - \frac{\lambda f}{L}\right)\right\} &= T_2(u)G_2(u) = \mathfrak{F}\{t_2(x)\}e^{-i2\pi u/L} \end{aligned} \quad (12.9)$$

Note that both the final coordinate systems could be reversed at the same time.

Corresponding to the position of original DPFHs, the resulting two phase Fresnel holograms appear at the focal plane of the second lens which are shifted with $\pm(\lambda f/L)$, respectively, along the x axis. Assuming, d being the distance between the centers of DPFHs, if the condition $d/2 = (\lambda f/L)$ is followed, the DPFHs are combined as a complex Fresnel hologram successfully. That makes phase-only SLM ideal for manipulating both the phase and amplitude of the optical wavefront.

One of the biggest advantages of the electronic-based processing systems is their flexibility due to their programmability. The SLM makes it possible to implement the programmable DOE whose characteristics can be changed dynamically [13]. Benefits from this, it serves a method we can employ to generate the blazed phase

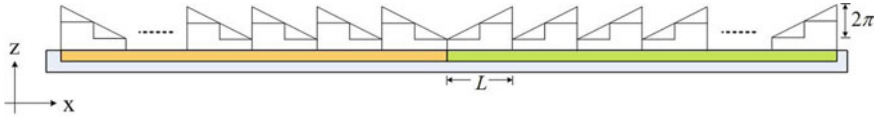


Fig. 12.3 Quantization of phase levels for a blazed phase grating

grating using phase-only SLM. As shown in Fig. 12.3, consider now the quantized grating (two-level), we have an amplitude transmittance over one period of,

$$G_1(u) = \begin{cases} e^{i0} = 1, & 0 \leq u < L/2 \\ e^{i\pi} = -1, & L/2 \leq u < L \end{cases} \quad (12.10)$$

The phase-only SLM applies a freedom way to generate different phase grating; however, it will also have an effect on the characteristics of the blazed phase grating. After some good approximation, the diffraction efficiency of first diffraction order we interested in given by

$$\eta_1 = \text{sinc}^2\left(\frac{1}{N}\right) \quad (12.11)$$

where N is denoted as phase levels. For SLMs, it means making a smaller pixel, because for a finite size of SLM period, scaling down the pixel would scaling up the phase levels. Since the smallest feature of the easily achievable DMD, LCD or LCoS technology is around 10 microns; if the SLM has been chosen, the minimum feature of the programmable phase grating is twice as much as the pixel pitch of the SLM. We generate blazed phase grating on the SLM assuming a spatial period of different pixels, which denies different values of L ; in this way, each phase hologram could be shifted with different distances along the x axis, and then we will give some results in the latter optical experiments to verify it.

12.3 Simulations

In this section, the numerical simulation is performed in order to investigate the synthesis of the DPFHs. The wavelength is $\lambda = 532$ nm, and the pixel pitch of the SLM1 is $\Lambda_1 = 12$ μm , which also determined the maximum resolution of the hologram. To use fast Fourier transform algorithms, we determine each Fresnel phase hologram has a 512×512 pixel resolution, and both of the DPFHs could be displayed side-by-side on the SLM1 along the horizontal direction. The pixel pitch of the SLM2 is $\Lambda_2 = 6.4$ μm . We set the focal lengths of lens1 and lens2 is $f_1 = f_2 = 300$ mm. And the spatial discretization by the SLM pixels will be further discussed in the following experiments.

In this paper, the Gerchberg–Saxton(GS) algorithm [14] could be used to calculate the Fresnel phase hologram. It should be noted that one could get a “conventional Fresnel phase-only hologram (CFPH),” keeping just the phase information after the iteration is over. However, we would like to encode arbitrary complex fields accurately, so the amplitude information does not be dropped. In this letter, the initial amplitude and phase functions of complex Fresnel hologram calculated by the GS algorithm are specially chosen with the purpose to obtain the DPFHs. As shown in Fig. 12.4a, a picture “GS” is the desired object pattern. Let $z_c = 200$ mm be the reconstruction distance and the iteration number we used is $k_i = 20$.

As shown in Fig. 12.4b, in order to observe the position-shift of each DPFH, only one of the DPFHs is displayed onto the right part of SLM1. Now, as far as the distance between the centers of each DPFHs is $d = 512 \times \Lambda_1 = 6144 \mu\text{m}$; finally, the grating period of the desired phase grating is calculated to be $L = 2\lambda f / d \approx 52 \mu\text{m}$, which using 8 pixels per period in the SLM2. Figure 12.4c shows the blazed phase grating loaded to the SLM2, and Fig. 12.4d shows the corresponding pattern obtained in the complex hologram plane. We can see that there is a spatial shift of the hologram along the x -axis. In the same way, the hologram which displayed onto the left part of SLM1 could be shifted and the combination of DPFHs can be achieved in the complex hologram plane.

For the purpose of comparison, the numerical reconstructions corresponding to the CFPH and complex Fresnel hologram obtained in the image plane are also shown in Fig. 12.4e, f, respectively. A quantitative estimation of the quality of the holographic

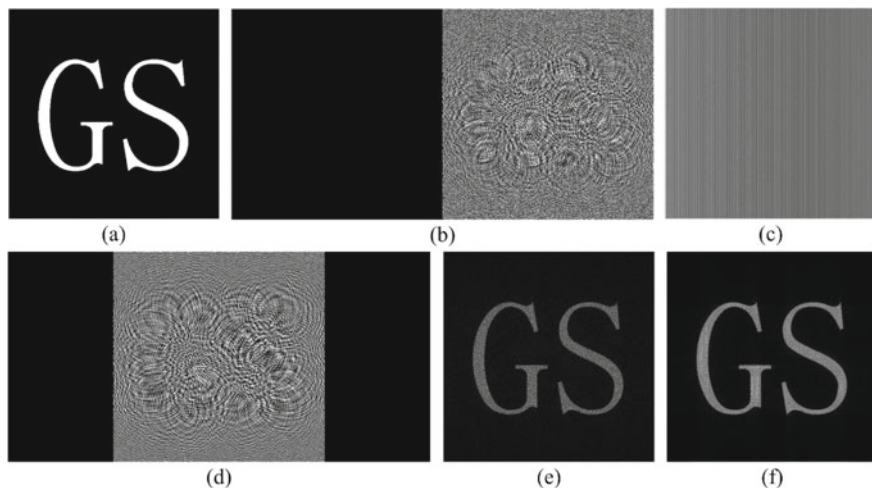


Fig. 12.4 **a** Object pattern, **b** only one phase hologram is displayed on the right part, **c** the blazed phase grating, **d** the corresponding pattern obtained in the complex hologram plane, **e** numerical holographic reconstruction of CFPH, **f** numerical holographic reconstruction of complex Fresnel hologram

reconstruction shall be given by the absolute root-mean-square (RMS) error, which is defined by,

$$\text{RMS} = \sqrt{\frac{\sum_{x,y} [|U_{\text{rec}}(x, y)| - |U_{\text{obj}}(x, y)|]^2}{M \times N}} \quad (12.12)$$

where $|U_{\text{rec}}(x, y)|$ and $|U_{\text{obj}}(x, y)|$ denote the amplitude of holographic reconstruction and the original object, with $M \times N$ as the pixel numbers of the image. The calculated RMS errors corresponding to Fig. 12.4e, f are 0.1505 and 0.1377, respectively. It points out that compared with the CFPH, the complex Fresnel hologram generates a more accurate reconstruction of the desired object.

12.4 Experimental Demonstration and Discussions

In our experiment, two LCoSs are employed to design a holographic display system. LCoS is the marriage of flat-panel display and silicon chip technology, so the resulting product could benefit from both sides [15]. In this paper, we focus on the position-shift of phase hologram, so let us consider the simple case of one $4f$ configuration; the detailed setup of the system is shown in Fig. 12.5, because the LCoS operates in reflection mode rather than transmission mode; two beam splitters naturally should be used to achieve the $4f$ configuration shown in Fig. 12.2. However, the efficiency is reduced to a half after the light pass through the beam splitter once. Finally, for the sake of high efficiency, we choose to use near-perpendicular incidence rather than perpendicular incidence, which make no use of two beam splitters. We assume this has a minor effect on behavior of our system.

The LCoS1 used in our work is MD1280 (Three-Five systems), which has 1280×1024 pixel resolution and a small pixel pitch of $12 \mu\text{m}$. It is illuminated by an

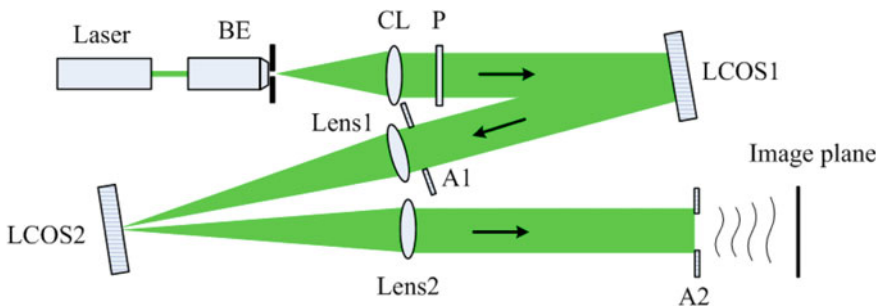


Fig. 12.5 Experimental setup based on two LCoSs. BE: beam expander, CL: collimating lens, P: polarizer, A1: square aperture, A2: square aperture

expanded collimated laser beam (diode-pumped solid-state green laser), and the illumination beam is linearly polarized. The LCoS2 we used is JD8554 (Jasper Display Corp.), which is capable of phase modulation. This reflective LCoS has 1920×1080 pixel resolution and a small pixel pitch of $6.4 \mu\text{m}$. The focal lengths of lens1 and lens2 are $f_1 = 300 \text{ mm}$ and $f_2 = 300 \text{ mm}$, respectively. The image plane is at the distance $z_c = 200 \text{ mm}$ from the back focal plane of lens2.

To provide a baseline position of hologram for comparison, we first assume a basic $4f$ setup, if a mirror instead of the LCoS2 inserted in the setup, then the light passes through the setup and we obtain the same exact hologram in the output plane as we load onto LCoS1. The pixelated structure of the LCoS causes multi-order diffractions. In this paper, we only focus on the reconstruction image placed in the center of the diffraction pattern. To avoid the aliasing effect between multi-order diffractions and the shifted hologram, aperture A1 is proposed to filter out the multi-order diffractions except the expected hologram. The final hologram obtained in the output plane of the $4f$ system is shown in Fig. 12.6a when the basic $4f$ configuration is used. As discussed in Sect. 2, the pixel pitch of LCoS2 is smaller than LCoS1 and that leads to the limited maximum shifted distance of the hologram is $\lambda f_2 / 2\Lambda_2 \approx 12.47 \text{ mm}$ from the original position. To investigate the dependence of performance on the different periods, we configure a series of blazed phase gratings using different pixels per period generated by phase-only LCoS2. N_g is the number of pixels per period we use, and the shifted holograms in the output plane are shown in Fig. 12.6b–d when $N_g = 8, 4, 2$. We have chosen the incident light diffracted into the -1 diffraction order as the target hologram. The measured shift distance of holograms is 3.1 mm , 6.2 mm , and 12.5 mm , respectively, and these results have a good agreement with the calculation results described in Sect. 3. It shows the capability of the method to encode blazed phase grating on a phase-only SLM. As it can be seen that there are some multi-order beams caused by the blazed phase grating with different periods, for the simulation in Sect. 3 when $N_g = 8$, as shown in Fig. 12.6b, our approach fails because the aliases of these multi-order beams are presented, a general solution to this problem is to increase the distance d by a loss of some spatial resolution for displaying the phase hologram. In this paper, to avoid the overlap of adjacent diffraction orders, it requires we get the following condition to be fulfilled,

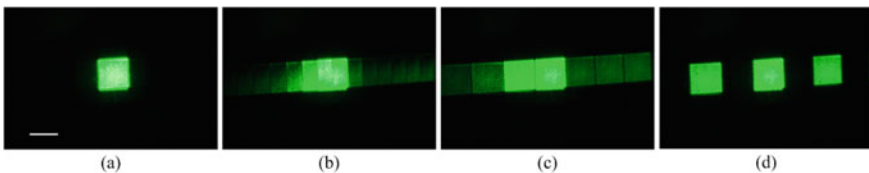


Fig. 12.6 Holograms obtained in the back focal plane of lens2, scale bar is 6.0 mm , **a** basic $4f$ setup, **b** 8 pixels per period, **c** 4 pixels per period, **d** 2 pixels per period.

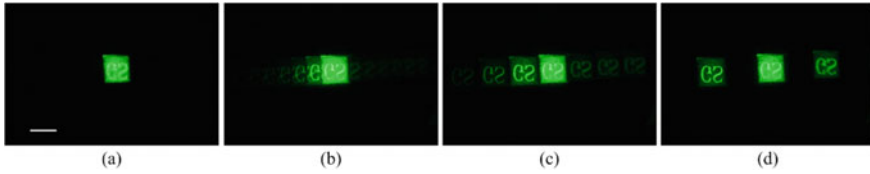


Fig. 12.7 Holographic reconstructions in the image plane, scale bar is 6.0 mm, **a** basic $4f$ setup, **b** 8 pixels per period, **c** 4 pixels per period, **d** 2 pixels per period

$$\frac{M\Delta_1}{2} \leq \frac{\lambda f_2}{N_g \Delta_2} \quad (12.13)$$

It is worth to mention that the small pitch of SLM2 is not only needed to diffract light to a large offset diffraction angle but it also can avoid the overlapping from higher orders into the first diffraction order used.

Here, by changing the period of the blazed phase grating, we also get the corresponding holographic reconstructions in the image plane, as shown in Fig. 12.7a–d. As already explained in Sect. 2, all the holographic reconstructions as well as the coordinate systems in the output plane of the $4f$ system are reversed. It can be noted that the convolution approaches exhibit the same magnification between the holograms and the reconstructions. For practical issue, we place an aperture A_2 at the back focal plane of lens2 to block the undesired orders except the -1 diffraction order caused by blazed phase grating; resulting in only the shifted hologram appears in the region of output plane. The effects of the aperture2 are shown in Fig. 12.8a–d.

Experimental reconstructions show that the method we proposed could achieve to implement the position-shift of hologram. A smaller pixel pitch and more pixels of SLM for displaying the blazed phase grating are effective enough to improve the accuracy for the shift of the hologram. As a notable result, it is possible to achieve a spatial shift of the RGB holograms along the coordinate axis to synthesis a color hologram for color holographic display.



Fig. 12.8 Holographic reconstructions in the image plane without undesired orders, scale bar is 6.0 mm, **a** basic $4f$ setup, **b** 8 pixels per period, **c** 4 pixels per period, **d** 2 pixels per period

12.5 Conclusions

In this paper, we have proposed an analytical method to display a complex Fresnel hologram using the phase-only SLM. Based on DPFHs encoding method, we prefer to employ two $4f$ configurations with the phase-only SLM inserted in the Fourier plane, and the effectiveness of the method is demonstrated by encoding different phase functions of the blazed phase gratings onto the SLMs. The use of two LCoSs also shows the basic design of one $4f$ configuration. Optical experiments verify that the position-shift of hologram has been implemented, and it provides a useful way for similar applications.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 61605002, 61501001, 61301296), Natural Science Foundation of Anhui Province (Grant Numbers 1508085MF121, 1608085QF161), Natural Science Project of Anhui Higher Education Institutions of China (KJ2016A029). The authors would like to thank the Jasper Display Corporation for providing the JD8554 LCoS.

References

1. Tudela, R., Martin-Badosa, E., Labastida, I., Vallmitjana, S., Juvells, I., Carnicer, A.: Full complex Fresnel holograms displayed on liquid crystal devices. *J. Opt. A: Pure Appl. Opt.* **5**(5), 189–194 (2003)
2. Becker, M.F., Wu, S.Y., Liang, J.: Encoding complex values using two DLP(R) spatial light modulators. In: Proceedings of SPIE—The International Society for Optical Engineering. San Francisco (2013)
3. Jesacher, A., Maurer, C., Schwaighofer, A., Bernet, S., Ritsch-Marte, M.: Near-perfect hologram reconstruction with a spatial light modulator. *Opt. Express* **16**(4), 2597–2603 (2008)
4. Qi, Y.J., Chang, C.L., Xia, J.: Speckleless holographic display by complex modulation based on double-phase method. *Opt. Express* **24**(26), 30368–30378 (2016)
5. Carbonell-Leal, M., Mendoza-Yero, O.: Shaping the amplitude and phase of laser beams by using a phase-only spatial light modulator. *J. Vis. Exp.* **143**, e59158 (2019)
6. Mendoza-Yero, O., Minguez-Vega, G., Lancis, J.: Encoding complex fields by using a phase-only optical element. *Opt. Lett.* **39**(7), 1740–1743 (2014)
7. Arrizon, V.: Complex modulation with a twisted-nematic liquid-crystal spatial light modulator: double-pixel approach. *Opt. Lett.* **28**(15), 1359–1361 (2003)
8. Chang, K.M., Chen, C., Wang, J., Wang, Q.H.: Improved single-random-phase holographic encryption using double-phase method. *Opt. Commun.* **443**, 19–25 (2019)
9. Cai, J.J., Shen, X.S., Fan, C., Kong, D.Z., Huang, F.Y.: High resolution real-time complex modulation using single spatial light modulator. *Laser Phys. Lett.* **16**(6), 066201 (2019)
10. Liu, J.P., Hsieh, W.Y., Poon, T.C., Tsang, P.: Complex Fresnel hologram display using a single SLM. *Appl. Opt.* **50**(34), 128–135 (2011)
11. Zhang, Z., Liu, J., Gao, Q.K., Duan, X.H., Shi, X.L.: A full-color compact SD see-through near-eye display system based on complex amplitude modulation. *Opt. Express* **27**(5), 7023–7035 (2019)
12. Verrier, N., Atlan, M.: Off-axis digital hologram reconstruction: some practical considerations. *Appl. Opt.* **50**(34), H136–H146 (2011)

13. Martinez, J.L., Martinez-Garcia, A., Moreno, I.: Wavelength-compensated color Fourier diffractive optical elements using a ferroelectric liquid crystal on silicon display and a color filter wheel. *Appl. Opt.* **48**(5), 911–918 (2009)
14. Gerchberg, R.W.: A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* **35**, 237–246 (1972)
15. Chen, H.M.P., Yang, J.P., Yen, H.T., Hsu, Z.N., Huang, Y., Wu, S.T.: Pursuing high quality phase-only liquid crystal on silicon (LCOS) devices. *Appl. Sci.* **8**, 2323 (2018)

Chapter 13

An Examination of Obstacle Avoidance by Sound for Visually Impaired Children



Yukiko Matsushima, Teruo Kimura, Feifei Cho, and Noboru Yabuki

Abstract In this paper, our research group examines obstacle avoidance by sound to support visually impaired children, because it is estimated that there are 1.4 million visually impaired children in the world. For that reason, we have developed a walking training device for these children. This paper demonstrates obstacle avoidance by sound as an elemental technology of the device. An obstacle transmission element outputs different sounds depending on the position of the obstacle. In particular, the element calculates the distance from the child to the obstacle by values from another elemental technology. After that, the element selects a sound using the distance and obstacle coordinates, and outputs the sound. For an evaluation of the elemental technology, we conducted an evaluation experiment to confirm that the sound correctly corresponds with the distance to the obstacle. As a result, we confirmed that the device could calculate distance and output different sounds correctly.

13.1 Introduction

It is estimated that there are 1.4 million visually impaired children in the world. Moreover, the number of the children is forecasted to increase by five hundred thousand new patients each year, of which 50% of the children die within 1–2 years [1]. They are slower in motor development such as grabbing, walking and running than healthy children. Their walking start time is about two years old on average. When babies can walk alone, their range of activities expands, but for visually impaired children, they may receive support from guardians or walking trainers because it is feared that they do not know the situation around them [2].

In contrast, according to a survey by Ministry of Health, Labour and Welfare in 2017 [3], there are approximately fourteen thousand certified orthoptists engaged in rehabilitation, about five hundred walking trainers working in schools and facilities for the blind. In addition to this, the number of guide dogs that are partners for

Y. Matsushima (✉) · T. Kimura · F. Cho · N. Yabuki
National Institute of Technology, Tsuyama College, 624-1 Numa,
Tsuyama, Okayama 708–8509, Japan
e-mail: matusima@tsuyama-ct.ac.jp

visually impaired people is also lacking in Japan. The number of guide dogs per one million population is 79.3 in New Zealand, 79.2 in the UK, 49.7 in Nederland, 36.5 in the USA and 28.8 in Ireland, compared to only 7.5 in Japan [4]. This is a very small number, because there are few guide dog breeders and guide dog training facilities.

As mentioned above, the supply of support services has not caught up with the need for the visually impaired people in Japan, so in this paper, our research group examines obstacle avoidance by sound to support visually impaired children. We have developed a walking training device for the children, and this paper demonstrates obstacle avoidance by sound, as an elemental technology of the device. The goal of the device is to reduce fear and resistance to walking, and to support safe walking. Furthermore, when the visually impaired children perform walking training in houses or parks using this device, the burden on guardians and walking trainers is reduced.

The device consists of three elemental technologies. In the first element, an ultrasonic sensor, an infrared sensor and a temperature sensor obtain values indicating an obstacle. In the second element, a 3D distance image sensor obtains the coordinates of the obstacle. The third element, the obstacle transmission element, outputs different sounds depending on the position of the obstacle. In particular, the obstacle transmission element calculates the distance from the child to the obstacle by the values from the first element. After that, the element selects a sound using the distance and the obstacle coordinates from the second element, and outputs the sound.

For an evaluation of the elemental technology, we conducted an evaluation experiment to confirm that the sound correctly corresponds with the distance to the obstacle. As a result, we confirmed that the device could calculate distances and output different sounds correctly every 30 (cm). In addition to this, we examined which frequencies are easy to hear for a child in Okayama School for the blind, and found that frequencies around 300 Hz and around 700 Hz were easy to hear. Using audible sounds is important, because infants with inherent visual impairment have slower sense and motor development than healthy children.

As a related research, Takuno et al. proposed a small device that guides visually impaired people to destinations [5]. They examined the use of this device for walking training. The purpose of this walking training is to help visually impaired people who can walk alone to learn new walking routes, and this device can partly act for walking trainers. They use a pedometer to calculate walking distance and a geomagnetic sensor to detect a direction of walking in order to realize a small, lightweight and inexpensive system.

The rest of this paper is organized as follows: Section 13.2 presents the walking training device. Section 13.3 describes the obstacle transmission unit as an elemental technology of the device. Section 13.4 shows evaluation results. Section 13.5 concludes this paper with a future study.

13.2 Walking Training Device

To use the walking training device, a visually impaired child pushes a handcart to perform walking training, as shown in Fig. 13.1. The child is attached to three types of sensors on the chest, and there are a 3D distance sensor and a laptop on the handcart. The sensors on the chest and the handcart detect obstacles in front of the child, and the speaker of the laptop plays sounds indicating the positions of the obstacles.

In order to ensure effective walking support, our research group set the goals of this device as follows.

- Distance to obstacles from the child: detects obstacles up to 3 (m) in front of the child.
- Obstacle height: detects obstacles up to the height of the child.
- Ground depression: detects grooves, stairs, etc.

Our research group has developed this device into three major elemental technologies. Firstly, we call the first element *the ultrasonic sensor unit*. In this ultrasonic sensor unit, an ultrasonic sensor, an infrared sensor and a temperature sensor obtain values indicating an obstacle to a visually impaired child. Figure 13.2 shows the appearance of the first unit. The first unit mainly detects obstacles in front of the child. The unit uses the ultrasonic sensor as the main sensor and uses the infrared sensor for soft obstacles that are hard to detect by the ultrasonic sensor. Moreover, the temperature sensor measures temperature because the ultrasonic sensor changes its outputs depending on the temperature. By this mechanism linking the sensors, the walking training device can measure the values to an obstacle accurately.

Secondly, we call the second element *the 3D distance image sensor unit*. This 3D distance sensor unit detects some obstacles and acquires their shapes. This unit outputs the shapes of the obstacles in coordinates. Figure 13.3 shows the appearance of the second unit. The walking training device recognizes an obstacle in two steps.

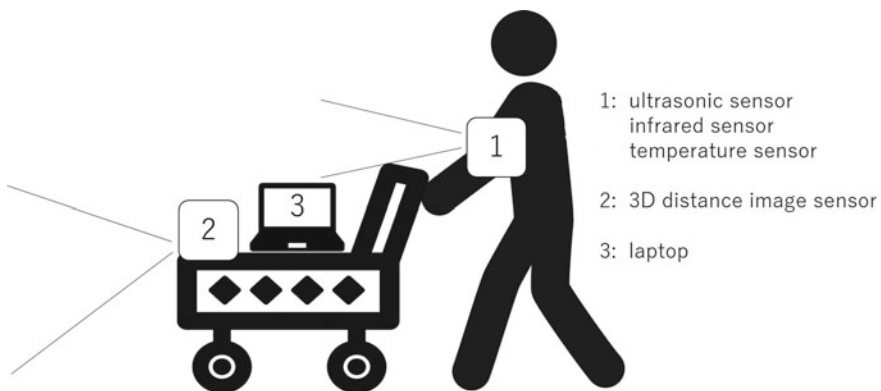


Fig. 13.1 Walking training device

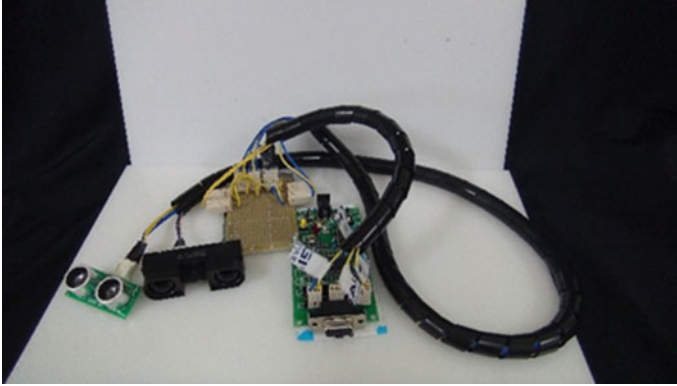


Fig. 13.2 Ultrasonic sensor unit



Fig. 13.3 3D distance image sensor unit

First, the ultrasonic sensor unit detects the obstacle, and after that the 3D distance image sensor unit detects the shape of the obstacle.

Lastly, the third element outputs different sounds depending on the position of the obstacle using inputs from these sensors. We call this element *the obstacle transmission unit*, and this paper demonstrates the third unit as an elemental technology of the device.

13.3 The Obstacle Transmission Unit

13.3.1 Sound Selection Process

The obstacle transmission unit calculates the distance from a visually impaired child to an obstacle by the outputs of the ultrasonic sensor unit. After that, this obstacle transmission unit selects a sound using the distance and the obstacle coordinates from the 3D distance image sensor unit, and outputs the sound.

In addition, as a method of transmitting the position of obstacles to visually impaired children, we examined a sound method by a speaker, a vibration method by a vibrator and a light emission method by LED. As a result, we decided to use the sound method in this research. This is because the sound method does not rely on vision, can easily be extended to output in various forms of information on positions and distances of obstacles, and can be miniaturized because it does not require additional IC chips.

The walking training device uses a microcomputer as a processing device for each sensor of the ultrasonic sensor unit. The obstacle transmission unit performs distance calculation on this microcomputer. This microcomputer transmits a calculation result using a serial communication, and the laptop on the handcart receives it. At the same time, the laptop receives outputs from the 3D distance image sensor unit.

The output of the 3D distance image sensor unit is coordinates that indicate positions and sizes of obstacles seen from the perspective of the visually impaired child. The program on the laptop uses the coordinates to determine the positions of the obstacles seen from the perspective of the child.

At the end of this process, the obstacle transmission unit determines the position of the obstacle in three dimensions according to the position and the distance of the obstacle seen from the child, and selects an appropriate sound for the obstacle avoidance to the child. Figure 13.4 shows these procedures.

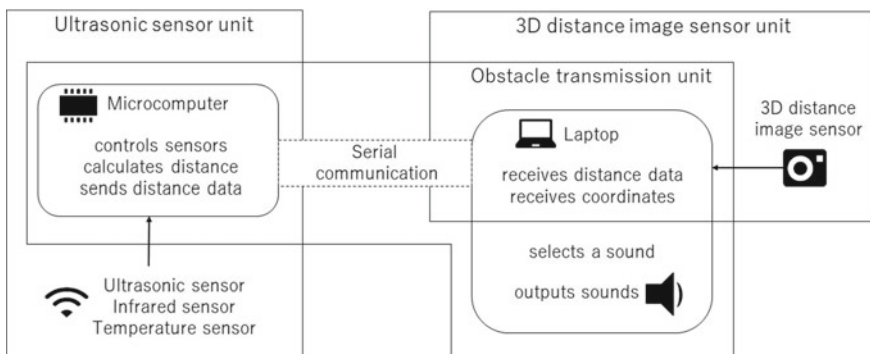
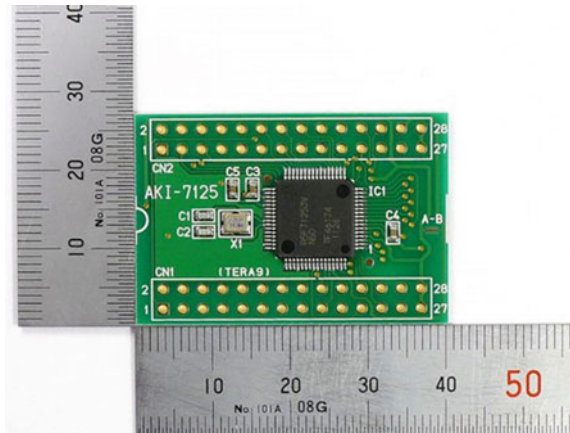


Fig. 13.4 Sound selection process

Table 13.1 Outline of SH7125F

Item	Performance
Power supply voltage	DC 4.0–5.5 (V)
Internal flash memory	128 (KB)
RAM	8 (KB)
Clock rate	50 (MHz)

Fig. 13.5 SH2 Tiny microcomputer SH7125F

The reason why the ultrasonic sensor unit uses the microcomputer is that the ultrasonic sensor unit must be lightweight and compact so that the visually impaired child can wear it on the chest. In this study, our research group uses the SH2 Tiny microcomputer SH7125F. The reason for this choice is that processing is fast and extensibility is high. Additionally, the SH7125F has a power supply voltage of 5 (V), which is common to other components. Table 13.1 shows the outline of SH7125F, and Fig. 13.5 shows the appearance of it.

13.3.2 Positioning of Obstacle

13.3.2.1 Distance Calculation to Obstacle

This section shows how to calculate the distance between a visually impaired child and an obstacle using the measurement data of each sensor from the ultrasonic sensor unit.

Our research group assumes that the distance between the visually impaired child and the obstacle is L (m), the speed of sound is c (m/s), and the time until an ultrasonic sensor emits ultrasonic waves and they bounce back is t (s). Equation (13.1) calculates L .

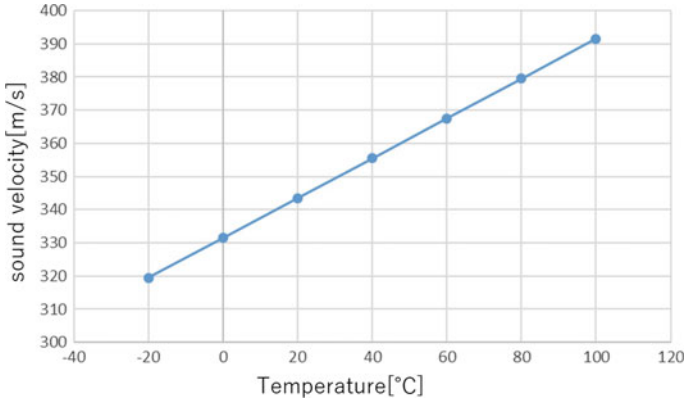


Fig. 13.6 Change of sound velocity by temperature

$$L = c \times t \times \frac{1}{2} (\text{m}) \quad (13.1)$$

Here, as shown in Fig. 13.6, the speed of sound changes depending on the air temperature, so the ultrasonic sensor unit measures the air temperature using the temperature sensor and calculates the speed of sound to obtain a more accurate distance. We assume that the temperature is T (°C), and Eq. (13.2) calculates c .

$$c = 331.5 + 0.6T (\text{m/s}) \quad (13.2)$$

13.3.2.2 Position Classification of Obstacle

This section shows the method to classify the position of an obstacle using the coordinates data from the 3D distance image sensor unit.

The obstacle transmission unit obtains the planar position of the obstacle seen from the visually impaired child as coordinates from the 3D distance image sensor unit. The obstacle transmission unit calculates the width and height of the obstacle from these coordinates. If the 3D distance image sensor unit detects two or more obstacles, the 3D distance image sensor unit can detect the closest one.

The obstacle transmission unit uses these data to classify the position of the obstacle in nine directions seen from the perspective of the child, which are upper left, upper, upper right, left, front, right, lower left, lower and lower right, as shown in Fig. 13.7.

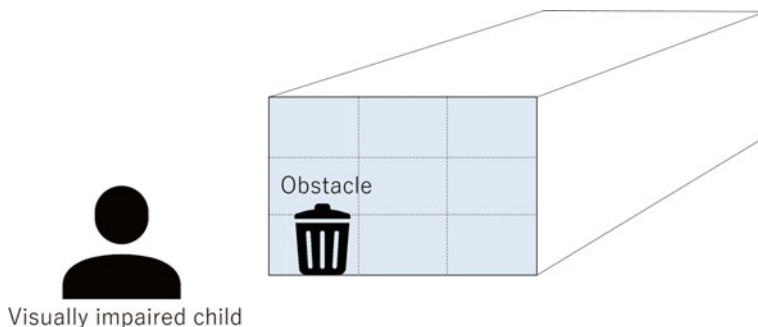


Fig. 13.7 Position classification of an obstacle

13.4 Evaluation Experiment

For an evaluation of the obstacle transmission unit, our research group measured whether a sound correctly corresponds with a distance to an obstacle. In addition to this, we examined which frequencies are easy to hear for a child in a school for the blind.

13.4.1 *Measurement Experiment of Sound Selection Accuracy*

In order to confirm whether the obstacle transmission unit can accurately calculate a distance to an obstacle and select a sound corresponding to the distance by the outputs of the ultrasonic sensor unit, we performed an experiment in which a subject is attached to the ultrasonic sensor unit on the chest and walked toward the obstacle. The temperature of the experiment day was 15 (°C).

The procedure of the experiment was as follows.

1. A subject stands 3 (m) in front of an obstacle.
2. The obstacle transmission unit calculates the distance to the obstacle and outputs the sound corresponding to the distance from the speaker.
3. The subject advances 30 (cm).
4. Repeat step 2 and step 3 until the subject touches the obstacle.

Table 13.2 shows the distance to the obstacle and the corresponding output sound used in this measurement experiment. In the experiment, we set the obstacle transmission unit so that the sound goes up the scale sequentially from C to A every time the subject approaches the obstacle by 50 (cm).

As a result of the measurement experiment, we confirmed that this unit can calculate distances and output different sounds correctly every 30 (cm) as shown in Table 13.3.

Table 13.2 Output sound by distance

Distance to obstacle (cm)	Sound
–75	A
75–125	G
125–175	F
175–225	E
225–275	D
175–	C

Table 13.3 Result of measurement experiment of sound selection accuracy

Distance to obstacle (cm)	Measured distance (cm)
30	30
60	60
90	90
120	121
150	151
180	183
210	212
240	244
270	276
300	305

13.4.2 *Hearing Experiment by Children in School for Blind*

We investigated which frequency band sounds are easy to hear for one child at Okayama Prefectural Okayama School for the Blind. The reason is that infants with inherent visual impairment are slower in development of other senses and motor development than healthy children, so using audible sounds is important.

Table 13.4 shows the information of the child who cooperated in the experiment.

Table 13.4 Information of the child who cooperated in the experiment

Item	Data
School year	Sixth grade in elementary school
Height	123 (cm)
Vision	Right eye: total blindness Left eye: severe blindness

As an experiment method, the speaker of this unit outputted sounds of 300–700 Hz one by one, and we asked the child whether they could hear the difference between these sounds. The results are as follows.

- He could easily hear sounds around 300 and 700 Hz.
- He did not understand differences of the sound around 500 Hz.
- He could find differences by 10 Hz for sounds around 300 Hz.

The child who participated in the experiment was not impaired except for the visual sense and was able to carry out the sound discrimination as instructed by our research group. On the other hand, there are a number of children born with complex disorders affecting the brain, such as visual impairment and physical development delay, so we must broaden the range of subjects and find sounds that are easy for most people to hear.

13.5 Conclusion

Our research group developed a walking training device to reduce visually impaired children's fear and resistance to walking, and to support safe walking. In this paper, we implemented an obstacle transmission unit as an elemental technology and conducted two evaluation experiments. As a result of the measurement experiment, we confirmed that the unit accurately outputted sounds according to distances to the obstacles. As a future development, we are considering the introduction of stereophonic sound so that visually impaired children can intuitively recognize the position of an obstacle.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number JP17872748.

References

1. SightFirst Long Range Planning Working Group: Childhood Blindness Position Paper. Lions Clubs International (2008)
2. Nakamura, M., Oshiro, E.: The Current Situation and Issues for Early Nurturing Support of Young Visually Impaired Children (in Japanese). The bulletin of Akita University (2014)
3. Ministry of Health, Labour and Welfare: Annual Report on Government Measures for Persons with Disabilities (2018)
4. Fukui, R.: Japanese guide dog training industries from the point of view from the world. *Nippon Hojyoken Kagaku Kenkyu* 2(1), 22–25 (2008). (in Japanese)
5. Takuno, S., Shinoda, Y., Tadokoro, Y.: Traveling training using traveling support system for the visually impaired. *IEICE Trans. Inf. Syst.* **J83-D1**(2), 293–302 (2000). (Japanese Edition)

Chapter 14

Efficient Technique of Impulsive Noise Detection and Replacement in Color Digital Images



Bogdan Smolka

Abstract In this paper, a very efficient method of impulsive noise suppression in color digital images is presented. The new filtering design exploits the concept of digital paths connecting the boundary of the filtering window with its center pixel. If the minimum connection cost of a path linking the window center with its border exceeds a predefined threshold, then the pixel is considered as distorted and filtered, otherwise, it is preserved. For finding the path with minimum cost, the Dijkstra algorithm is applied. The noisy pixels are restored using a new approach, which is based on the overdetermined system of equations used in the Laplace inpainting technique. The experimental results show that the application of the impulse detection relying on the connection cost, combined with a robust method of pixel replacement, significantly excels over existing techniques designed for impulsive noise reduction in digital color images.

14.1 Introduction

The reduction of noise in digital images is still one of the most researched subjects, as the correction of image distortions is required in various practical imaging tasks. Color images are frequently degraded by different kinds of impulsive noise, which randomly changes one or all color pixels channels. This kind of noise can be introduced by faulty pixels in the camera sensors, aging of the storage material, flawed memory locations and can be also caused by analog to digital converter errors during the acquisition process and electromagnetic disturbances [1, 2]. The enhancement of color images affected by impulsive noise is the subject of this contribution.

The techniques of noise suppression in color images operate on individual image channels or they treat color pixels as vectors. In the channelwise denoising, the color image components are processed using methods developed for the grayscale pictures and are then combined to obtain the filtered output. Such an approach is

B. Smolka (✉)
Institute of Automatic Control, Silesian University of Technology,
Akademicka 16, 44–100 Gliwice, Poland
e-mail: Bogdan.Smolka@polsl.pl

straightforward and often applied, however, the omission of correlation between image components often creates strong color artifacts, especially noticeable at edges and for this reason vectorial methods are preferred.

Very often, filters used to remove impulses in color images are utilizing the concept of order statistics [3]. The goal is to order the color pixels, treated as vectors, belonging to a processing window, which will be denoted as W . For each element of W , the sum of distances to other pixels from the filtering window is calculated and the aggregated distances serve as a measure of pixel similarity. A small value of the cumulated distances indicates that there are many similar pixels in the processing window, whereas a large value suggests that a pixel is most probably an outlier, which significantly differs from other pixels occupying the window and has to be corrected. The pixel minimizing the cumulated distances to other members of W constitutes the output of the widely used vector median filter (VMF) [4].

The drawback of VMF is that the filter uniformly processes every image pixel and the unaffected pixels are treated in the same way as outliers, which causes unnecessary replacements of undistorted pixels resulting in the loss of image details. The VMF is also not able to suppress the Gaussian noise, as its output is always a pixel from the noisy image. To alleviate this effect, trimming techniques were introduced, which determine the average of the first ordered pixels to enable the smoothing of Gaussian noise contaminating the color image. Additionally, various kinds of color spaces and dissimilarity measures between pixels are being applied [2, 5].

If the image is affected by impulsive noise only, various switching filters, which tend to filter out only the corrupted pixels are used. First, the noisy pixels are detected and in the second step they are recovered by a suitable robust filter, which operates on the uncorrupted pixels from the neighborhood. Generally, the switching filters, that restore exclusively the samples detected as distorted, offer good denoising efficiency, as the clean pixels are left unchanged [5–7].

In [8], a concept of a peer group of pixels belonging to the processing window was proposed. If a pixel is not disturbed, then it should show high similarity to other pixels, otherwise, it is an outlier injected by the impulsive noise. For the detection of noisy pixels, their distances to the neighboring pixels are determined and then sorted. If the difference between successive distance values exceeds a predetermined threshold, the processed pixel is changed by the VMF output, otherwise, it is retained.

The number of pixels which exhibit strong similarity to the center of W is also a reliable measure of pixel corruption. In the method introduced in [6, 9], the number of pixels close to the center of W is counted. If the middle pixel of isolated and in its neighborhood a predefined number of close neighbors cannot be found, then it is replaced by the VMF or by the average of neighboring pixels which were detected as not corrupted.

Another technique of impulsive pixels detection [10] is based on the analysis of the dissimilarity between the central element of W and the average of the pixels with smallest ranks in a sorted sequence obtained using the reduced ordering concept. The cumulative sum of pixel distances to the VMF output can also be used as an approximation of the pixel variance and on this basis the outlying pixels can be

detected [11]. This approach has been modified in [12], where weights which decrease with the pixel rank in an ordered sequence of pixels were introduced.

The signal-dependent rank order mean (SD-ROM) also calculates the distances between the pixels from W , finds the lowest-ranked pixels and compares them with the processed pixel using predefined thresholds. If the distances do not exceed the threshold values, a pixel is recognized as undistorted, otherwise, it is replaced by the VMF output.

The detection of corrupted pixels can be also performed with the use of fuzzy set techniques. These denoising filters are based primarily on fuzzy similarity measures [13, 14], which are utilized to determine the degree of membership of a filtered pixel to its neighborhood.

Recently, algorithms utilizing the cost of digital paths exploring the neighborhood of a pixel were proposed. In [15], the cost of a connection between filtering window border and its central pixel was applied as a measure of pixel distortion. The connection cost was estimated utilizing a two-pass algorithm and used for the calculation of distance transform in binary images [16]. A simplified version of this algorithm, which allowed only the shortest paths to link the central pixel with the window boundary was also proposed [17].

In this paper, a method which is using the Dijkstra algorithm to find the minimum connection cost and a very promising technique of detected noisy pixels replacement is described. The novel filtering design significantly excels over existing methods of impulsive noise removal and due to its low computational burden can be used in various imaging applications.

The rest of paper is organized as follows. Next section presents the construction of the noisy pixel detection and replacement method. Then, the experimental results are described and compared with the competitive denoising techniques. At the end, some conclusions are drawn and future work is briefly outlined.

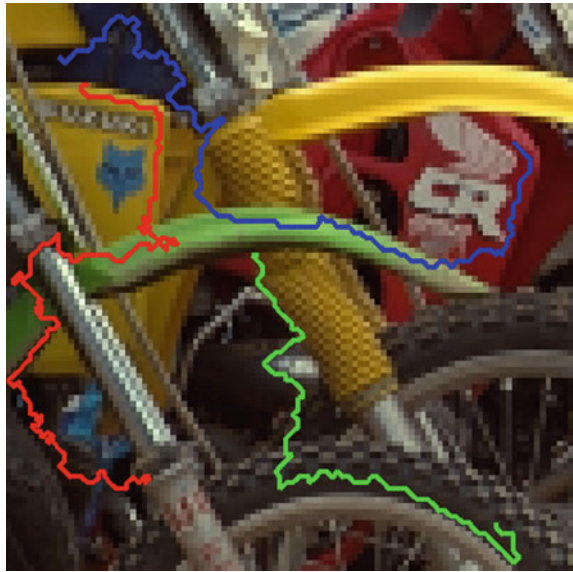
14.2 Noisy Pixel Detection and Replacement

14.2.1 Impulsive Noise Detection

Let us assume that the image is processed using a square window W_i , consisting of $n = (2r + 1)^2$ pixels, which is placed at pixel f_i , where i stands for the pixel location on the image domain. Then, let us analyze the connection costs of digital paths joining the window boundary with its central pixel. The cost \mathcal{C} of a path consisting of pixels f_0, f_1, \dots, f_η will be defined as

$$\mathcal{C}\{f_0, f_\eta\} = \sum_{q=0}^{\eta} D(f_{q-1}, f_q),$$

Fig. 14.1 Exemplary paths of minimum connection costs. Starting and ending points are marked with white circles



where D is a dissimilarity measure between neighboring pixels. Performed experiments show that choosing $D(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|^2$ where $\|\cdot\|$ denotes the Euclidean distance calculated in RGB color space provides satisfying results [15, 18, 19]. The cost of a connection of the central pixel f_i of W_i to its boundary will be denoted as $C(f_i)$. Figure 14.1 shows paths of minimum connection costs joining three pairs of test image pixels. As can be observed the paths tend to propagate along smooth areas, which enables to minimize the overall connection cost.

Figure 14.2 depicts a part of a grayscale image (a) and its pixels intensities (b). For every pixel, the connection cost with the window's border is calculated with the Dijkstra algorithm. In this example, the absolute value of the difference between the intensities of neighboring pixels is taken as the cost of a transition. If the central pixel is an outlier, there is no low value connection with the window border. Even if the noisy central pixel is surrounded by similarly corrupted pixels, the minimum cost is still high, as the cluster of pixels form an "island" which is not connected with the boundary of the filtering window ("mainland"). In this way, we can calculate the minimum cost of all paths linking the center pixel with the boundary of the processing window and we can treat this value as a measure of pixel impulsiveness. The map of minimum connection costs is shown in (c) and its visualization is depicted in (d). Applying a suitable threshold value, the pixel in the center of the processing window, which is corrupted by impulsive noise and needs replacement can be detected.

In this way, the pixel f_i is detected as an outlier if the cost $C(f_i)$ of a connection to the border of the window W_i exceeds a predefined threshold value denoted as h . The output y_i of the outlier detection is formulated as

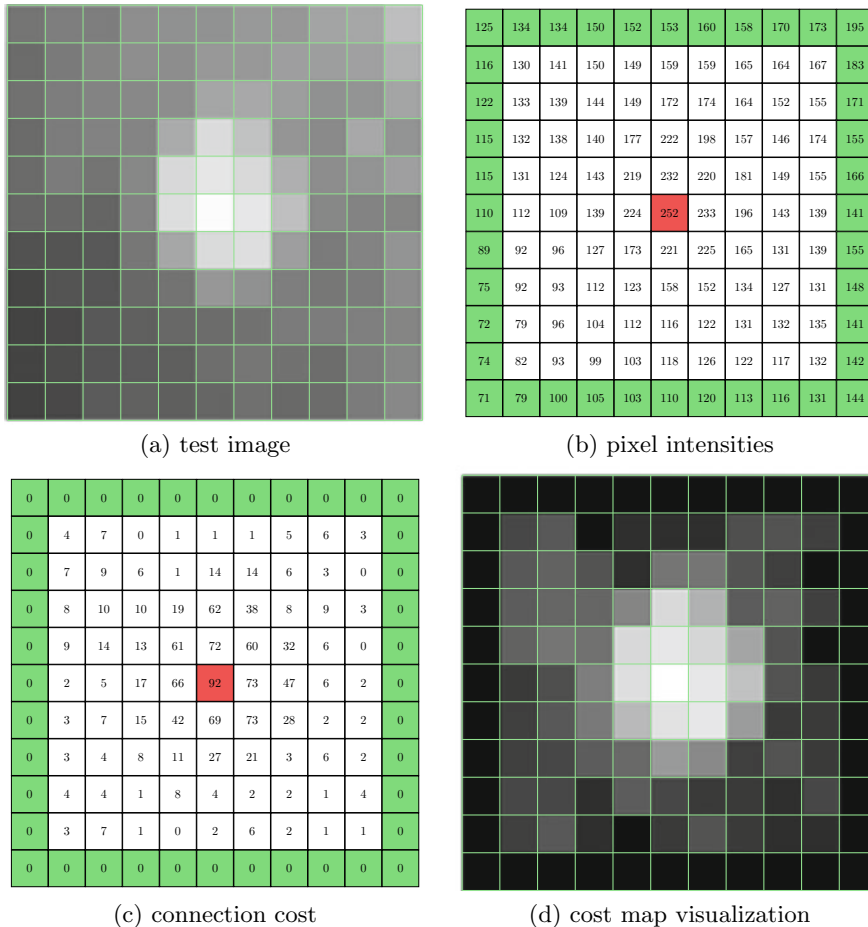


Fig. 14.2 Concept of the minimum connection cost C of the middle pixel of the filtering window (marked red) of size $(2r + 1) \times (2r + 1)$, with its boundary (marked green). In this example, $r = 5$

$$y_i = \begin{cases} f_i & \text{if } [C(f_i)]^{\frac{1}{2}} \leq h, \\ \text{NaN}, & \text{otherwise,} \end{cases}$$

where NaN stands for *Not a Number* data type. The pixels which were assigned the NaN label will be processed in the subsequent step and replaced with a Laplacian-based inpainting technique.

Figure 14.3a presents a part of test image IMG 3 chosen from a set of pictures shown in Fig. 14.6, which was corrupted with 20% random valued impulse noise (b). The random impulsive noise modifies a fraction of image elements (p), in such a way that RGB components are replaced by random values from a uniform distribution in

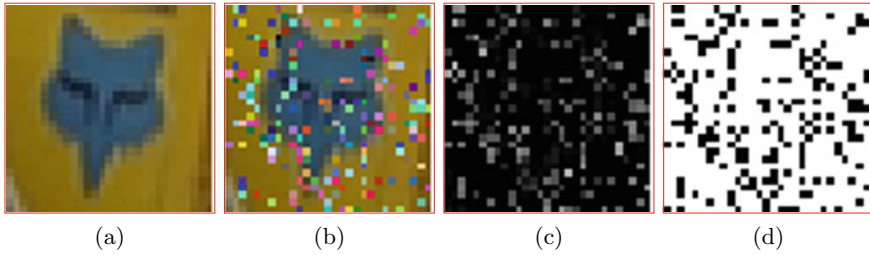


Fig. 14.3 Noisy pixel detection scheme using the minimum connection cost approach: **a** part of test image, **b** degraded image, $p = 0.2$, **c** map of noise intensity and **d** detected impulses which are assigned NaN label

the range $[0, 255]$, assuming a 24 bit color image. The connection costs are depicted in (c) and the pixels whose connection cost exceed a certain threshold are marked in black (d).

14.2.2 Noisy Pixel Replacement

Many inpainting methods were proposed in the rich literature, however, the techniques based on the Laplacian interpolation are popular, as they are simple, fast and deliver good quality results. This kind of interpolation assumes that the missing data can be interpolated using a harmonic function $u(x, y)$ which satisfies $\Delta u = 0$. Harmonic functions have the mean value property, which can be used for determining the missing data points. In case of images, every missing pixel intensity is an average of its neighbors.

The inpainting method based on the Laplacian partial differential equation can be expressed as

$$\begin{cases} \Delta u = 0 & \text{in } \Omega \setminus \Omega_R, \\ u = f & \text{in } \partial \Omega_R, \\ \partial_n u = 0 & \text{in } \partial \Omega \setminus \Omega_R, \end{cases}$$

where f is the *known* and *reliable* image data in the region denoted as Ω_R consisting of pixels whose neighborhood does not contain missing values (NaNs), with boundary $\partial \Omega_R$ on the image domain Ω and ∂_n denotes the derivative in normal direction to the image border [20] (see Fig. 14.4a).

Figure 14.4b shows a part of a grayscale image with two detected impulses marked in black (NaNs), with their direct neighbors marked in yellow. The reliable pixels are marked with green color. To simplify the procedure, the four-neighborhood system is applied. We assume that every pixel detected as an impulse (u_1 and u_2 in the example) and the adjacent pixels (yellow) are equal to the average of the local neighborhood. This assumption leads to a system of equations

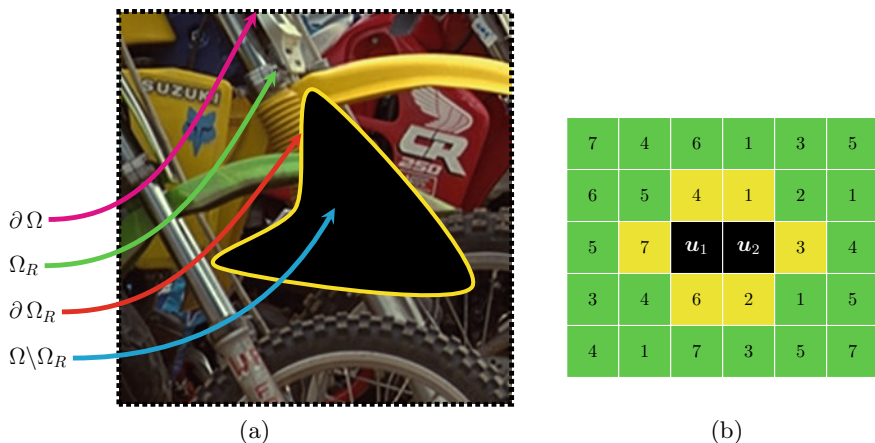


Fig. 14.4 Illustration of the Laplace interpolation method (a) and a part of an exemplary image with two impulses marked in black and their neighbors marked in yellow (b). The remaining *reliable* pixels are colored with green

$$\begin{aligned}
 u_1 &= (7 + 4 + u_2 + 6)/4, & u_2 &= (u_1 + 1 + 3 + 2)/4, \\
 7 &= (5 + 5 + u_1 + 4)/4, & 4 &= (5 + 6 + 1 + u_1)/4, \\
 1 &= (4 + 1 + 2 + u_2)/4, & 3 &= (u_2 + 2 + 4 + 1)/4, \\
 2 &= (6 + u_2 + 1 + 3)/4, & 6 &= (4 + u_1 + 2 + 7)/4.
 \end{aligned}$$

After simplifying, we obtain an overdetermined system of eight equations with only two unknowns, which can be solved by the standard least-squared method

$$\begin{aligned}
 4u_1 - u_2 &= 17, & -u_1 + 4u_2 &= 6, & u_1 &= 14, & u_1 &= 4, \\
 u_2 &= -3, & u_2 &= 5, \\
 u_2 &= -2, & u_1 &= 11.
 \end{aligned}$$

$$\begin{bmatrix} 4 & -1 \\ -1 & 4 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 17 \\ 6 \\ 14 \\ 4 \\ -3 \\ 5 \\ -2 \\ 11 \end{bmatrix} \rightarrow u_1 = 5.58, \quad u_2 = 2.58.$$

The application of the overdetermined system of equations is beneficial, as the final output is more robust to outlying pixel values not properly identified in the

noisy pixel detection step. Additionally, the interpolation of corrupted pixels takes into account the neighboring missing pixels, whose interpolated values are found solving the overdetermined equation system. Such an approach yields much better results than interpolating using only the known values in the local neighborhood, which is a standard procedure applied in the popular filtering designs. For high noise intensity, it may happen that a corrupted (missing) pixel is surrounded by outliers, which in the standard filters lead to the application of a larger window or other technique which allows to replace the pixels recognized as injected by the noise process. Using the proposed technique, such a problem does not exist, as even large parts of the image can be interpolated using the proposed inpainting technique. In this way, the proposed method is capable of first detecting and then interpolating small clusters of impulses which are treated by the conventional methods as tiny image details and are retained.

Figure 14.5 shows the result of the Laplace interpolation method using a part of a test image. The regions which have to be filled are marked in black. As can be observed, the missing pixels are smoothly filled with a harmonic function. The texture details are lost; however, in case of impulsive noise, where the corrupted pixels form only small clusters of pixels, this effect is not noticeable.

14.3 Experiments

In this work, we evaluated the proposed image filtering approach on test images polluted by so-called *random valued* impulsive noise. This noise model presumes that the channels of the corrupted pixels are changed by a random variable drawn from a uniform distribution from the available intensity range, in our case $[0, 255]$ [6, 21], and the relative amount of corrupted pixels is denoted by p .

The efficiency of the proposed denoising method was evaluated utilizing the PSNR and MAE image quality measures defined as

$$\text{MSE} = \frac{1}{3N} \sum_{i=1}^N \sum_{c=1}^3 (f_{i,c} - z_{i,c})^2, \quad \text{PSNR} = 10 \log \left(\frac{255^2}{\text{MSE}} \right), \quad (14.1)$$

$$\text{MAE} = \frac{1}{3N} \sum_{i=1}^N \sum_{c=1}^3 |f_{i,c} - z_{i,c}|, \quad (14.2)$$

where $f_{i,c}, c \in \{1, 2, 3\}$ are color components of the original, (clean) pixels with index i , N denotes the total amount of pixels and $z_{i,c}$ are output pixels. The experiments were performed on test images from the database used in [6] and filtering results are reported using test images as depicted in Fig. 14.6.

Figure 14.7 depicts the dependence of the PSNR measure on the filtering window size r for the test images 1, 2, 5 and 6 (see Fig. 14.6). As can be observed, a filtering window of the size 5×5 ($r = 2$) enables to obtain the best denoising results for low

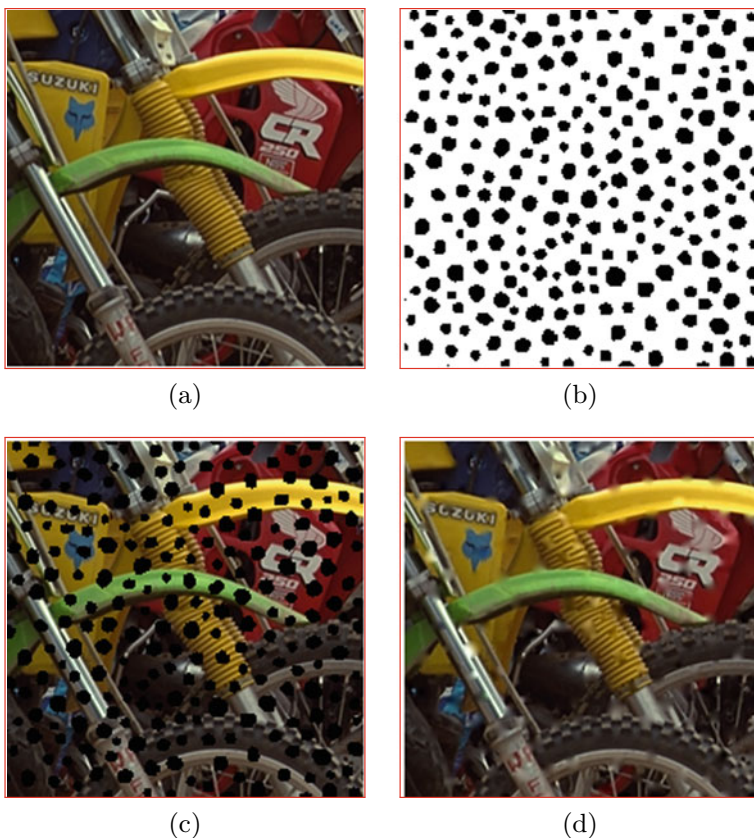


Fig. 14.5 Illustration of the Laplace inpainting: **a** test image, **b** mask of positions (black) for which the RGB values are unknown and have to be interpolated, **c** the test images with superimposed mask, **d** result of Laplace interpolation

and medium contamination level ($p \leq 0.3$). For higher image degradations, a larger window yields slightly better results; however, the gain is not worth the increased computational burden.

In this way, the only parameter of the novel algorithm is the threshold h , which is used to detect the noisy pixels. As can be observed in Fig. 14.8, the dependence of the PSNR on the thresholding parameter h is similar for all test images and does not change significantly with the noise intensity.

The maximum of PSNR is attained for the optimal value of the threshold h . If the threshold is too small, too many pixels are recognized as corrupted and are modified. Otherwise, some impulses are treated as uncorrupted pixels and are retained. Generally, the PSNR measure is varying slowly with the threshold value which is beneficial as a deviation of h from the optimal setting causes only a small decrease in the PSNR performance. Therefore, the setting $h = 20$, for which the PSNR attains



Fig. 14.6 Test images from the publicly available database [6].

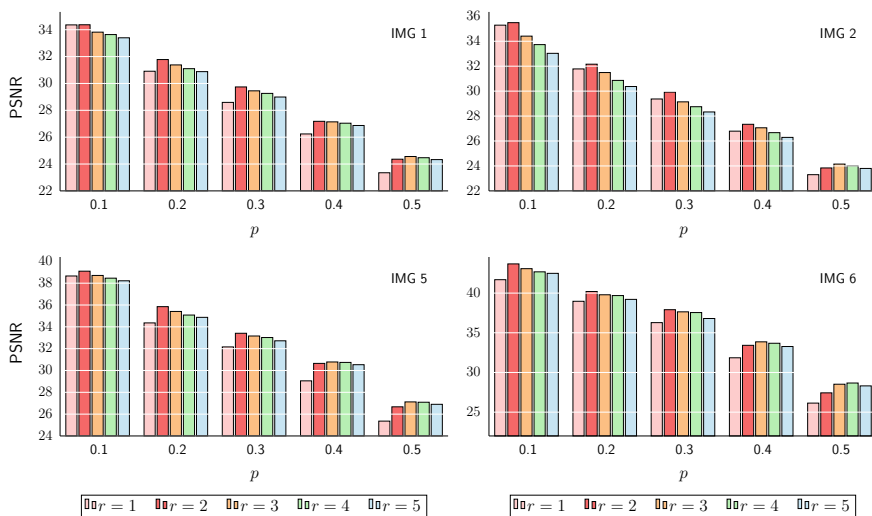


Fig. 14.7 Dependence of PSNR on the window size r for increasing fraction of corrupted pixels $p \in \{0.1, \dots, 0.5\}$ using test images 1, 2, 5 and 6

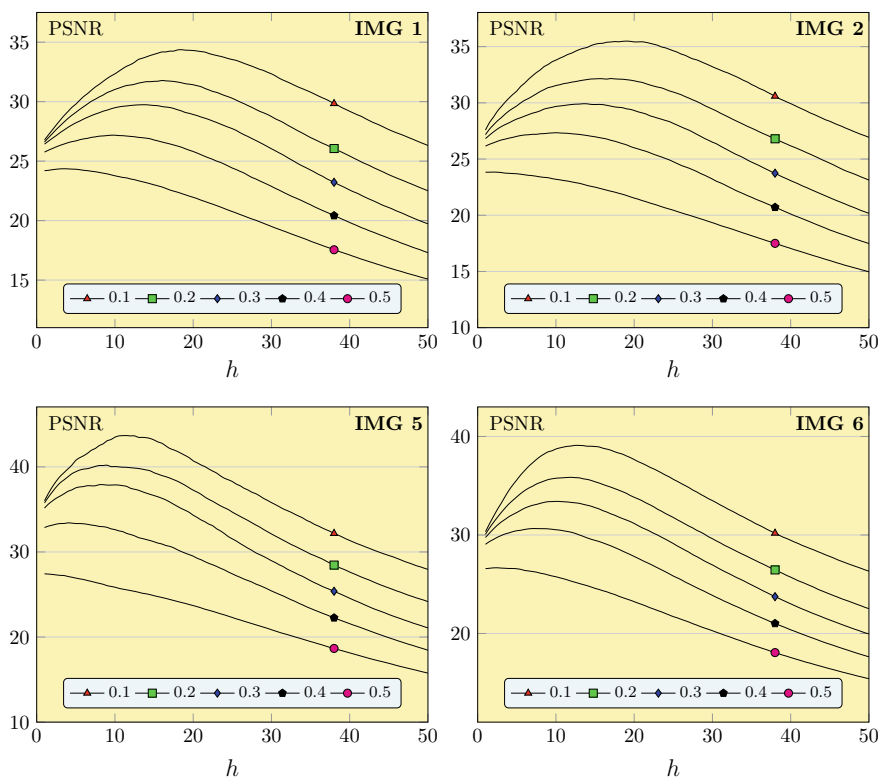


Fig. 14.8 Dependence of PSNR measure on threshold h for various fractions of corrupted pixels: $p \in \{0.1, \dots, 0.5\}$ using test images 1, 2, 5 and 6

mostly a value close to the maximum (see Fig. 14.8), provides satisfactory denoising efficiency and this default value was used for the comparison with the competitive denoising methods.

The following filters were taken for comparisons:

- Soft switching digital path filter (SSDPF) [15],
- Shortest path filter (SPF) [17],
- Adaptive center-weighted vector median filter (ACWVMF) [5],
- Adaptive vector median filter (AVMF) [5, 10],
- Fast averaging peer group filter (FAPGF) [7],
- Fast adaptive switching trimmed arithmetic mean filter (FASTAMF) [6],
- Fast fuzzy noise reduction filter (FFNRF) [5, 14],
- Fast modified vector median filter (FMVMF) [5, 22],
- Fast peer group filter (FPGF) [5, 23],
- Peer group filter (PGF) [8] and
- Sigma directional distance filter (SDDF) [5, 11].

Table 14.1 Comparison of the filtering efficiency of the proposed method with the state-of-the-art denoising techniques

IMAGE	p	NEW	SDDPF	SPF	ACWVMF	AVMF	FAPGF	FASTAMF	FFNRF	FMVMF	FFGF	PGF	SDDF
<i>PSNR [dB]</i>													
1	0.1	34.37	32.95	33.17	32.06	29.92	31.83	33.16	31.46	30.49	30.88	33.01	30.09
	0.2	31.78	30.63	30.86	28.23	26.79	29.89	30.83	29.12	28.49	28.49	29.55	25.31
	0.3	29.74	28.89	29.13	24.33	24.06	28.15	28.79	26.06	25.76	25.73	26.29	21.32
2	0.1	35.49	33.70	33.90	32.06	29.91	32.07	33.50	31.10	30.35	30.65	33.04	29.27
	0.2	32.15	30.93	31.13	27.78	26.90	29.96	30.90	28.83	28.28	28.19	29.27	23.98
	0.3	29.92	29.02	29.18	23.38	23.81	28.17	28.79	25.71	25.00	24.93	25.51	20.14
3	0.1	31.89	30.63	30.82	29.22	28.46	28.89	30.43	28.61	27.32	27.94	30.45	28.09
	0.2	29.14	28.24	28.41	26.06	25.47	26.97	28.05	26.51	25.58	25.76	27.16	23.67
	0.3	26.75	26.36	26.54	22.63	22.79	25.40	26.17	23.92	23.28	23.40	24.23	19.91
4	0.1	40.22	38.27	38.48	36.76	30.84	37.66	38.78	37.04	36.18	35.96	37.26	33.49
	0.2	37.45	35.77	35.99	31.82	27.61	34.96	35.93	33.14	33.19	32.70	32.89	27.65
	0.3	34.72	33.59	33.80	27.15	25.01	32.52	33.22	29.16	29.47	29.24	29.20	23.19
<i>MAE</i>													
1	0.1	0.79	1.53	1.40	1.18	1.50	1.25	1.03	1.24	1.54	1.57	1.03	2.00
	0.2	1.54	2.53	2.32	2.41	3.07	2.19	1.89	2.17	2.52	2.76	2.16	3.57
	0.3	2.45	3.57	3.30	4.50	5.18	3.28	2.87	3.62	4.06	4.47	3.76	6.61
2	0.1	0.62	1.25	1.14	0.98	1.36	1.10	0.91	1.06	1.32	1.43	0.89	1.73
	0.2	1.27	2.19	2.01	2.21	2.75	1.94	1.67	1.88	2.20	2.54	1.95	3.62
	0.3	2.08	3.13	2.89	4.56	4.86	2.95	2.55	3.27	3.89	4.40	3.66	7.08
3	0.1	1.13	2.18	1.96	1.83	1.91	2.01	1.57	1.90	2.65	2.64	1.51	2.54
	0.2	2.24	3.60	3.30	3.47	3.79	3.35	2.78	3.15	4.03	4.36	3.09	4.63
	0.3	3.59	5.08	4.72	6.13	6.33	4.85	4.12	4.99	6.13	6.69	5.22	8.40
4	0.1	0.34	0.70	0.64	0.54	1.21	0.52	0.46	0.50	0.58	0.67	0.52	1.03
	0.2	0.68	1.20	1.12	1.30	2.53	1.03	0.91	1.09	1.14	1.37	1.22	2.16
	0.3	1.11	1.79	1.65	2.68	4.26	1.71	1.47	2.05	2.10	2.44	2.24	4.46

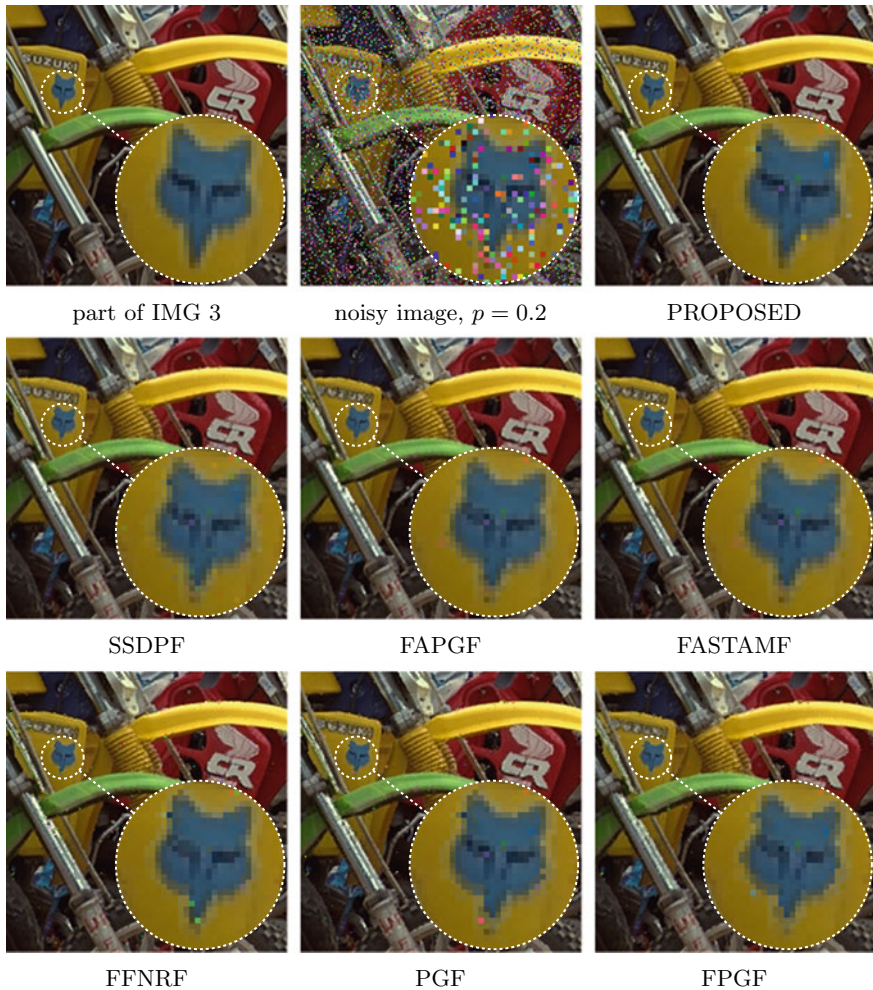


Fig. 14.9 Comparison of the filtering efficiency using IMG 3, ($p = 0.2$)

The denoising results in terms of PSNR and MAE quality measures are presented in Table 14.1. The proposed filter significantly outperforms the techniques based on digital paths (SSDPF, SPF) [15, 17] and other state-of-the-art filtering methods. The increased efficiency is due to the more accurate determination of the connection cost and much better efficiency of the applied inpainting method used for the noisy pixel replacement. The satisfactory denoising capability the proposed filter can be also assessed visually in Fig. 14.9.

14.4 Conclusion

In the paper, a novel filtering design suitable for the removal of impulsive noise in color digital images has been described. The proposed filtering structure is based on the minimal connection cost of digital paths linking the border of the filtering window with its center. This cost determined with the Dijkstra algorithm was used as a measure of pixel impulsivity and applied for the detection of samples contaminated by impulsive noise. This procedure proved to be very efficient and able to discriminate clusters of corrupted samples.

The second step of the proposed algorithm is the noisy pixels replacement. Instead of substituting the noisy pixels with local estimates based on the uncorrupted samples, the Laplacian inpainting scheme was used. Filling the gaps in the noisy images was performed solving an overdetermined system of linear equations. The approach utilizing the inpainting mechanism yields excellent results, both visually and objectively in terms of image restoration quality measures. The comparison with the existing filters show that the described approach offers significantly better suppression efficiency and detail preservation capability. As the proposed filter is fast and only one tuning parameter is needed, it can be applied in various image processing tasks which work with noisy color images. Future work will concentrate on the elaboration of more efficient transition costs working preferably in perceptually uniform color spaces.

Acknowledgements This work was supported by the Silesian University of Technology, Poland, (grant BK 2019/T3) and was also funded by a research grant 2017/25/B/ST6/02219 from the National Science Center, Poland.

References

1. Boncelet, C.: Image noise models. In: Bovik, A. (ed.) *Handbook of Image and Video Processing, Communications, Networking and Multimedia*, pp. 397–410. Academic Press, London (2005)
2. Plataniotis, K., Venetsanopoulos, A.: *Color Image Processing and Applications*. Springer, New York (2000)
3. Lukac, R., Smolka, B., Martin, K., Plataniotis, K., Venetsanopoulos, A.: Vector filtering for color imaging. *IEEE Signal Process. Mag.* **22**(1), 74–86 (2005)
4. Astola, J., Haavisto, P., Neuvo, Y.: Vector median filters. *Proc. IEEE* **78**(4), 678–89 (1990)
5. Celebi, M., Kingravi, H., Aslandogan, Y.: Nonlinear vector filtering for impulsive noise removal from color images. *J. Electron. Imaging* **16**(3), 033008 (2007)
6. Malinski, L., Smolka, B.: Fast adaptive switching technique of impulsive noise removal in color images. *J. Real-Time Image Proces.*, 1–22 (2016)
7. Malinski, L., Smolka, B.: Fast averaging peer group filter for the impulsive noise removal in color images. *J. Real-Time Image Proces.* **11**(3), 427–444 (2016)
8. Kenney, C., Deng, Y., Manjunath, B.S.: Peer group image enhancement. *IEEE Trans. Image Proces.* **10**(2), 326–334 (2001)
9. Smolka, B., Plataniotis, K.N., Chydzinski, A., Szczepanski, M., Venetsanopoulos, A.N., Wojciechowski, K.: Self-adaptive algorithm of impulsive noise reduction in color images. *Pattern Recogn.* **35**(8), 1771–1784 (2002)

10. Lukac, R.: Adaptive vector median filtering. *Pattern Recogn. Lett.* **24**(12) (2003)
11. Lukac, R., Smolka, B., Plataniotis, K., Venetsanopoulos, A.: Vector sigma filters for noise detection and removal in color images. *J. Vis. Commun. Image Representation* **17**(1), 1–26 (2006)
12. Smolka, B., Malik, K., Malik, D.: Adaptive rank weighted switching filter for impulsive noise removal in color images. *J. Real-Time Image Proces.* 1–23 (2012)
13. Hussain, A., Masood Bhatti, S., Jaffar, M.A.: Fuzzy based impulse noise reduction method. *Multimedia Tools Appl.* **60**(3), 551–571 (2012)
14. Morillas, S., Gregori, V., Peris-Fajarnes, G., Latorre, P.: A fast impulsive noise color image filter using fuzzy metrics. *Real-Time Imaging* **11**(5–6), (2005)
15. Smolka, B., Cyganek, B.: Impulsive noise suppression in color images based on the geodesic digital paths. *Proc. SPIE* **9400**, 9400–12 (2015)
16. Rosenfeld, A., Pfaltz, J.: Distance functions on digital pictures. *Pattern Recogn.* **1**(1), 33–61 (1968)
17. Smolka, B., Malinski, L.: Impulsive noise removal in color digital images based on the concept of digital paths. In: *International Conference on Computer Science and Education (ICCSE)*, pp. 1–6 (2018)
18. Smolka, B.: Impulsive noise removal in color images based on the local neighborhood exploration by geodesic digital paths. *Proc. Int. Multi. Sci. GeoConf.* **17**, 359–366 (2017)
19. Smolka, B.: Fast technique of impulsive noise suppression in color images. In: *Proceedings of the 4th IIAE International Conference on Intelligent Systems and Image Processing*, pp. 527–532 (2016)
20. Hoeltgen, L., Kleefeld, A., Harris, I., Breuss, M.: Theoretical foundation of the weighted Laplace inpainting problem. *Appl. Math.* **64**(3), 281–300 (2019)
21. Phu, M., Tischer, P., Wu, H.: Statistical analysis of impulse noise model for color image restoration. In: *6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)*, pp. 425–431 (2007)
22. Smolka, B., Szczepanski, M., Plataniotis, K., Venetsanopoulos, A.: Fast modified vector median filter. In: *Proceedings of the 9th International Conference on Computer Analysis of Images and Patterns (CAIP)*, *Lecture Notes in Computer Science*, vol. 2124 (2001)
23. Smolka, B., Chydzinski, A.: Fast detection and impulsive noise removal in color images. *Real-Time Imaging* **11**(5), 389–402 (2005)

Chapter 15

Development of Walking Support System for Visually Impaired People



Feifei Cho, Tatsuya Ohta, Yukiko Matsushima, Teruo Kimura,
and Noboru Yabuki

Abstract Currently, shortage of trainers for visually impaired people with professional knowledge is a big problem in Japan (Japan Ministry of Health in Labour and Welfare: Annual Report on Government Measures for Persons with Disabilities, 2014; Fukui in *Nippon Hojyoken Kagaku Kenkyu* 2(1), 22–25, 2008) [1, 2]. In addition, children with congenital blindness are so difficult to learn to walk by themselves, because they cannot get the visual image. In this research, the authors aim to develop a new support device for walking training that can rise for visual impairment and visually impaired children's spatial perception. The method involves three stages of training, hand cart type stage, wearable type stage, and final stage. The authors focus on developing a safe, lightweight walking support system that can be worn in children's waist belt at the second and the third stage. In this paper, the authors proposed the wearable type walking support system that consists of infrared sensor and ultrasonic wave sensor. In addition, the equipment is used at each stage so that it can finally walk independently.

15.1 Introduction

Currently, there are concerns about an increase in visually impaired children for three reasons: ① Increase in children with disabilities due to population growth, ② popularization of medical infrastructure in the world, and ③ increase in premature infants. There is also a concern that the shortage of qualified trainers for the visually impaired with expert knowledge due to the increase in children with visual impairment will become more serious than ever. In addition, children with congenital blindness have no irritation on the vision, and independent walking is extremely difficult. In this research, the authors aim to develop a new support device for walking training for mobility and spatial capacity development of visually handicapped and visually impaired children.

F. Cho (✉) · T. Ohta · Y. Matsushima · T. Kimura · N. Yabuki
National Institute of Technology, Tsuyama College, 624-1 Numa, Okayama, Tsuyama 708-8509,
Japan
e-mail: cho@tsuyama-ct.ac.jp

In order to become a support device that can be used at normal family, the cost must be reduced. So, the authors use ultrasonic distance measuring sensor and infrared distance measuring sensor, which is inexpensive and easy to construct system. The authors also use SH7125F microcomputer with high processing speed and low price for control. Moreover, this device will be worn for a long time, and therefore, the authors designed and proposed walking training system with emphasis on weight reduction.

The authors assume three sensor module attachment devices, a backpack type, a stick type, and a push car type. Among three types, the push-pull type has a large payload, so it is supposed to install additional equipment.

In this paper, the authors investigated the present condition of visually impaired people and developed the walking training system that supports vision as a means of solving each problem and its mounting device. The authors devised three sensor module mounting devices and designed the sensor module and developed the system.

15.1.1 Background

Recently, visually impaired people are increasing steadily. As one of the indicators, the research team at Anglia-Ruskin University in England predicts that visually impaired people will increase from current 36 million to 150 million by 2050 unless improving treatment by budget enlargement [3]. The cause is aging, rising age of childbirth, improvement of medical technology. Figure 15.1 [3] shows the visually impaired people with severe and moderate vision disorders among the increasing visual impairments. The vertical axis shows the number of people, and the horizontal

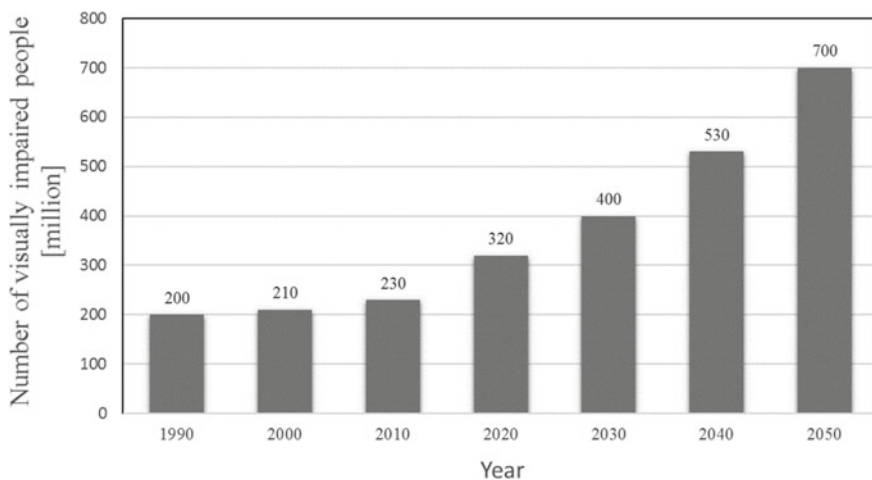


Fig. 15.1 Increase in the number of moderately and severely blind people

Table 15.1 Number of qualified care workers of various types

Profession	Number of people
Physical therapist	120,072
Occupational trainer	70,672
Vision trainer	12,085
An artificial prosthetist	4447
Language hearing expert	23,750

axis shows the year. In their paper, the authors know that Southeast Asia, China, South Africa are the most frequent increased [3, 4].

If the number of people with visual impairment that cannot be self-sustained increases, the burden on social welfare such as child rearing support expenses for visually impaired children and living assistance expenses for visually impaired people becomes immeasurable [5].

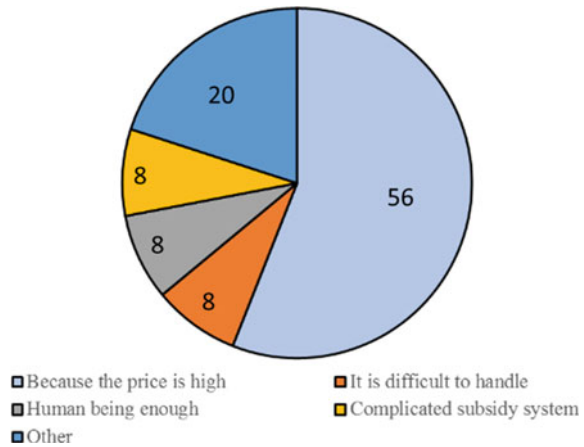
The other problem is the shortage of qualified trainers. As given in Table 15.1, there are 300,000 people with visual impairment in Japan. On the other hand, there are 12,000 qualified trainers [6]. The ratio is 25:1, and the shortage of this qualified trainer is said to become even more severe in the future. Figure 15.2 [6] shows the breakdown of caregivers.

Moreover then, it is extremely difficult to walking training people with visual impairment, it places a heavy burden on trainers and guardians during training such as colliding with obstacles in front of the face, falling into grooves/holes missed by white sticks. Therefore, the ratio of this qualified trainer to the visually impaired is a big problem to be solved [7].



Fig. 15.2 Emphasis on selecting nursing robots

Fig. 15.3 Reason about not to introduce care robot



15.1.2 Opinions from the Field

Next, the opinions of nursing care sites are shown at Fig. 15.2 [8, 9]. This data is on public opinion survey conducted in Japan. The graph is needed for the support device/support robot. The vertical axis represents the item, and the horizontal axis shows the percentage that answered “Yes.” Items are numerous sales results, good reputation, national recommendation, compact, easy to maintain, accident insurance, long-term care insurance, safety, low price, and ease of operation. The authors can see that low cost which is required from more than 60% of nursing side.

Next, Fig. 15.3 [10, 11] shows the reason why the caregiver does not introduce support devices/support robots. Items are in clockwise order, high prices, so handling is difficult, adequate human, complex subsidy system, others, etc. The graph shows that high price is a bottleneck in the introduction of support equipment/support robot.

15.1.3 Purpose

Therefore, the purpose of this study is to solve spatial problems in the training of visually impaired people and accompany various findings accompanying autonomous behavior, making it low cost and easy to handle.

The goal is to develop a new support device for walking training for visually impaired people who can realize the above. In addition, eventually, it will strengthen the connection between the movement ability and the environment of the spatial ability and create an opportunity to learn the perception, cognition, language, society, etc., of the visually impaired and visually impaired children, the social reintegration of visually impaired person. It aims to be useful for independence of children.

15.2 Walking Training System

In this research, the authors use one three-dimensional range image camera and one sensor to recognize surrounding environment. First, confirm the operation and the performance of the 3D range image system. Next, the distance image is acquired, and the object having the same distance from the distance image is extracted as one object. For each extracted object, discrimination between the moving object and the stationary object is performed. After that, the stationary object records the position of the living beings and recognizes it as an obstacle and aims at tracking the movement for the moving object. Furthermore, by detecting the object by distance measurement using a sensor, this is assisted to improve safety.

First, appropriate obstacles are corresponded to obstacles analyzed by image processing. At this time, infants with visual impairment are delayed in the concept of “there are obstacles in walking in the space,” so the authors repeated the experiments at school to determine the timing of the best utterance and the type of sound. Next, the authors will support training of direction sense by changing the sound to be output depending on who is walking toward. As for this, the authors adopt a method which tests both the absolute direction based on the east, west, north, and south and the direction relative to the entrance of the room and has a large blind angle effect.

Table 15.2 below shows the responsible range of the department of collaborative research and the contents of each subject of this research.

15.2.1 Sensor

The device detects approach by two distance measuring sensors. Figure 15.4 [12] shows the infrared distance measuring sensor used, and Fig. 15.5 [13] shows the ultrasonic distance measuring sensor. The infrared distance measuring sensor has high accuracy with respect to price and has few errors. However, there is a weak point that permeate cannot be detected. Therefore, by compensating for that part by the ultrasonic distance measuring sensor, it is possible to measure with high precision

Table 15.2 Division of research with collaborative research members

Affiliation/duties	Sharing
National Institute of Technology, Tsuyama College. Mechanical System Course	Understand the progress of the whole plan, construction of a fall prevention mechanism, introduction of sensors
National Institute of Technology, Tsuyama College. Communication and Information System Course	Image processing
National Institute of Technology, Tsuyama College. Communication and Information System Course	Audio output system

Fig. 15.4 Infrared distance measuring sensor

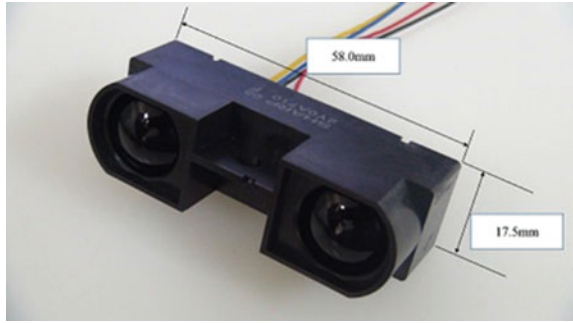
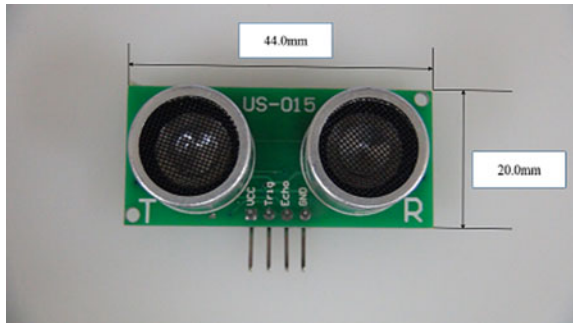


Fig. 15.5 Ultrasonic distance measuring sensor



regardless of the type of the object to be measured. Table 15.3 [12] and Table 15.4 [13] show the performance tables.

The infrared distance measuring sensor uses “sharp distance measuring module GP2Y0A710K.” The power supply voltage V_{cc} is -0.3 to $+7$ V, and the output terminal voltage is -0.3 to $V_{cc} + 0.3$ V. The operating temperature is -10 to $+60$ °C. The infrared distance measuring sensor uses “ultrasonic distance sensor module US-015.” The measurable distance is 0.02–4.0 m, the operating current is 2.2 mA, the power supply voltage V_{cc} is 5 V, and the operating temperature is 0 to $+70$ °C.

Table 15.3 Infrared distance measuring sensor data sheet

Parameter	Rating
Operating voltage	DC 5 V
Working current	2.2 mA
Operating temperature	0 ~ $+70^{\circ}$
Output mode	GPIO
Sensing angle is less	15°
Detection distance	2–400 cm
Detection accuracy	0.1 cm + 1%
Resolution higher	1 mm

Table 15.4 Ultrasonic distance measuring sensor data sheet

Parameter	Rating
Supply voltage	$V_{cc} - 0.3$ to $+7$ V
Output terminal	$V_o - 0.3$ to $V_{cc} + 0.3$ V
Operating temperature	-10 to $+60$ °C
Storage temperature	-10 to $+70$ °C

Therefore, it can be used under normal temperature, has a measurement distance enough to avoid obstacles, and can be fed from the same power supply.

Next, the authors show the distance detection method of each sensor. The infrared distance measuring sensor outputs a voltage corresponding to the distance. This is converted to a digital value by AD conversion. Calculate this value to find the distance. Examples are shown below.

(Ex) When the 10-bit AD conversion value is 400

$$\frac{5}{2^{10}} \times AD(= 400) = 1.953 \dots [V]$$

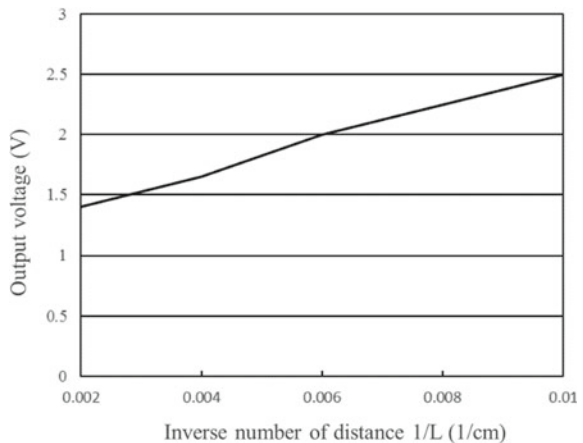
From the graph of Fig. 15.6 [12], the distance at this example is

$$0.006[1/cm] = 166.666 \dots [cm]$$

The distance can be measured from the above.

Next, the ultrasonic distance sensor outputs a pulse of a time corresponding to the distance. Therefore, calculate the distance by calculating the time of the pulse.

Fig. 15.6 Output characteristic of infrared distance sensor



15.2.2 Equipment Device

The mounting apparatus will be described. Figure 15.7 ~9 shows the outline of the device. The authors assume three types of mounting devices: “shoulder type,” “white cane type,” and “hand cart type.” Among them, the hand cart type has a large payload, so it is supposed to install 3D range image camera equipment. The attachment device’s main body is in green, the sensor module is in blue, and the 3D range image camera is in orange to indicate it.

Explain why multiple types of production are assumed? The visual impairment is roughly divided into three types, blindness, lack of vision, and amblyopia. In addition, the each level of symptoms is further divided, additionally visually impaired is often complicated with other disorders, and it is difficult to completely manualize the walking training. Therefore, it is to cope with this by selectively using multiple types or joint use.

Next, each mounting apparatus will be described. The hand cart type shown in Fig. 15.7 shall be used at the initial stage of walking training. This hand cart type guarantees safety, reduces fear of walking, and helps smooth transition to subsequent training. Shoulder type shown in Fig. 15.8 shall be used during walking training. This shoulder type is assumed to be used in combination with other mounting devices. When walking training in hand cart type, with white cane, the authors support walking training by installing it additionally. The white wand type shown in Fig. 15.9 assumes danger avoidance at the final stage of training and assistance during autonomous walking. Also, by combining these as described above, it is possible to support optimal walking training and autonomous walking according to individual levels.

Finally, prototype sensor module is shown in Fig. 15.10 and explained. In the prototype, the authors implemented the sensor most easily. Create a prototype sensor module as a book. The authors are also planning experiments on prototype operation and measurement system.

Fig. 15.7 Hand cart type

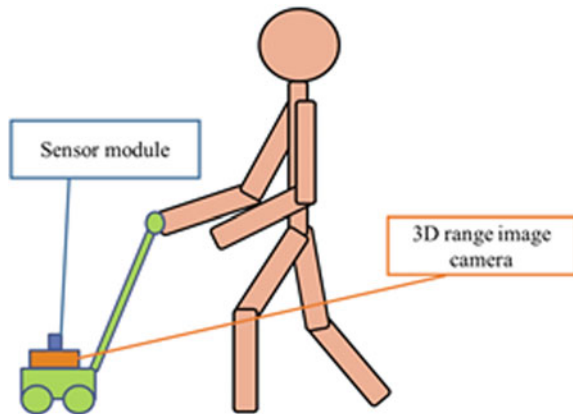


Fig. 15.8 Shoulder type

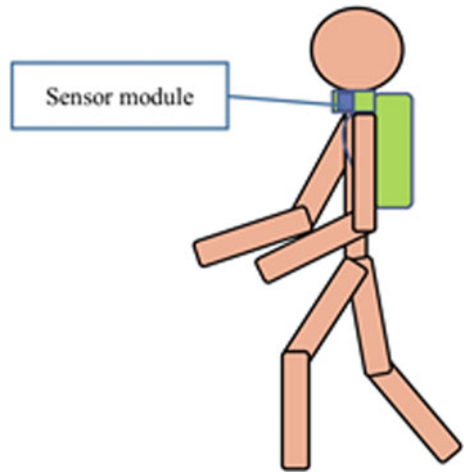


Fig. 15.9 White cane type

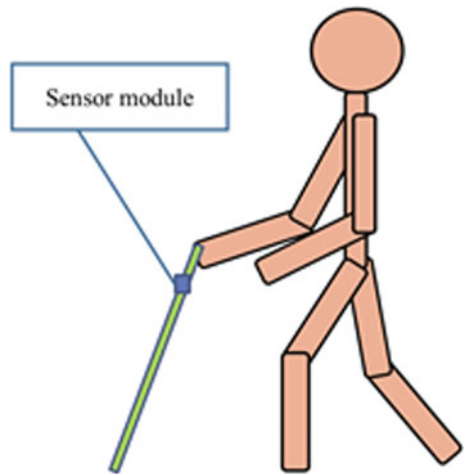
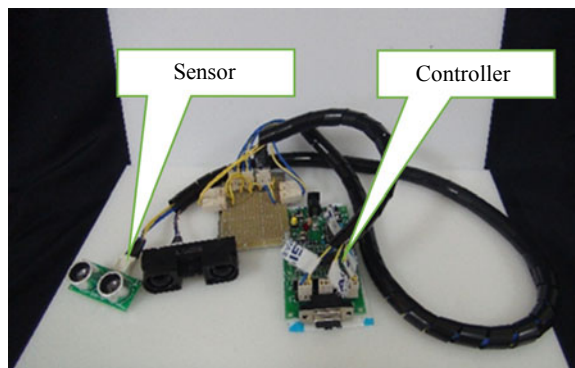


Fig. 15.10 Prototype of support system [14]



15.2.3 Equipment Production

In this research, the design concept of the part in charge of the laboratory is as follows.

1. Make the device wearable.
2. It is a device used for training of single walk.
3. Even children with visual impairment can safely use it.
4. The sensor detects an obstacle in front of the wearer and sounds it to the wearer.
5. There is little burden on the body when worn.

The authors made a device to realize the above concept.

Figure 15.11 shows the 3D CAD of the mounting device. Also, Fig. 15.12 shows the manufactured device, and Fig. 15.13 shows a photograph of it when worn.

Fig. 15.11 Mounting device image

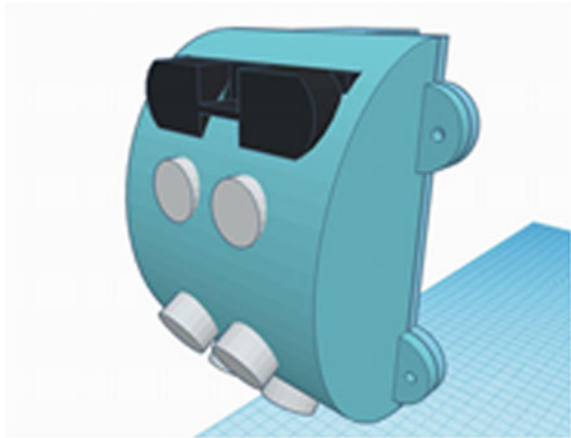


Fig. 15.12 Production equipment outline

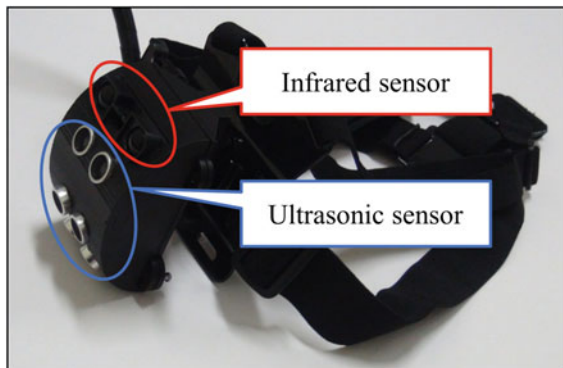




Fig. 15.13 Production equipment mounting photograph

15.3 Experiment

15.3.1 Experimental Method

An experiment was conducted on the performance of the fabricated sensor module and the correction program for the temperature sensor. Perform experiments using a mobile battery as a power source. Use a measure to measure distance. Three types of obstacles are used: concrete wall, acrylic board, and cloth. The reason for adopting an acrylic board is to check whether a transparent object can be detected in each sensor, and the reason for adopting a cloth is to check if a soft object can be detected in each sensor. Figure 15.14 shows an outline of the experimental setup. The detection distance of the measurement object is 1 m, and the detection range of ultrasonic waves is conical with a tip angle of 15° as shown in Fig. 15.14.

Fig. 15.14 Outline of experiment equipment

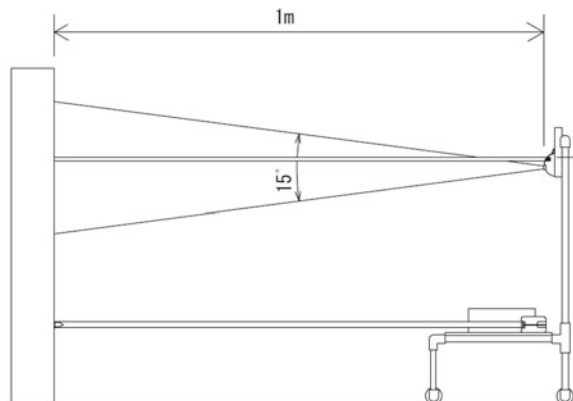


Table 15.5 Results of measurement by each sensor

	Concrete	Acrylic board	Cloth
Ultrasonic sensor average	100.8 cm	98.3 cm	
Same sensor standard deviation	0.71	0.63	
Infrared sensor average	96.7 cm		98.1 cm
Same sensor standard deviation	0.71		0.49

Table 15.6 Comparison of standard deviation of ad value measured by temperature sensor

	AD without correction	AD with correction
Average	377.87	360.70
Standard deviation	66.10	6.19

15.3.2 Results and Discussion

Table 15.5 shows the measurement results of the ultrasonic sensor and the infrared sensor. When the cloth was measured by the ultrasonic sensor, the sensor reacted, but the measurement distance was large, and the value was not stable. The reason for this is that the ultrasonic sensor measures the distance by the echo of the sound, so it is considered that the measured value becomes unstable with the cloth that the sound hardly echoes. In addition, when the acrylic plate was measured by the infrared sensor, the sensor did not react at all even if the object entered the set distance range. The reason for this is that the infrared sensor measures the distance by reflection of infrared light, so it is considered that the infrared light does not reflect, and it cannot detect the passing acrylic plate. From this, it was found that it is necessary to use an ultrasonic sensor and an infrared sensor in combination to detect all obstacles.

Next, Table 15.6 shows the results of experiments on the temperature sensor correction program. The set temperature is 76 °F. The ad value of the temperature sensor is read by a microcomputer and displayed on seven segments. The standard deviation of the ad value of the temperature sensor could be reduced to the original 9.36% after correction. It is considered that this makes it possible to greatly improve the correction of other sensors by the temperature sensor and make the accuracy in the distance measurement of the device more accurate.

15.4 Improvement

Figure 15.15 shows the detection range of the existing sensor module. As shown in the figure, in the existing sensor module, a gap exists between the 15° detection

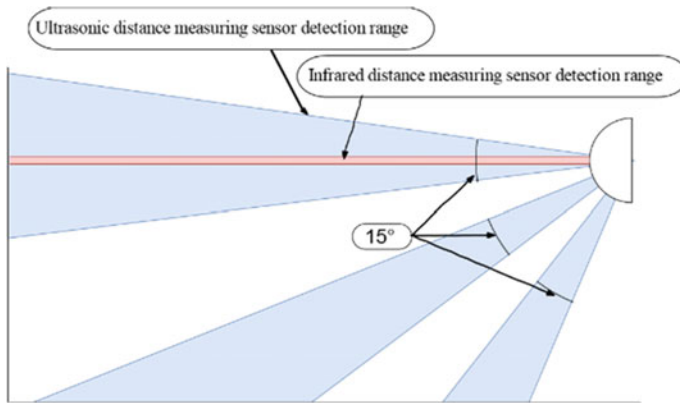
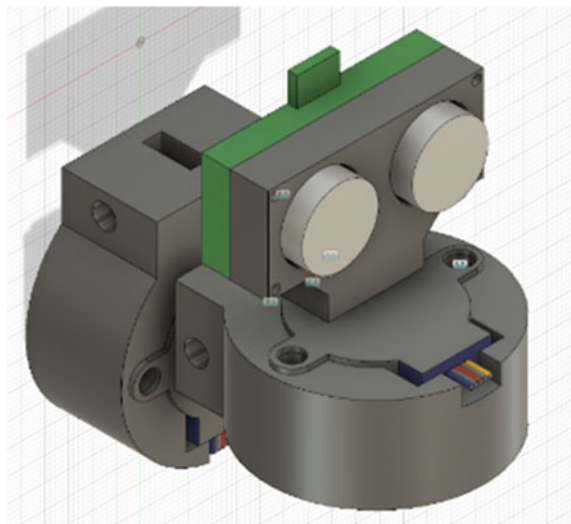


Fig. 15.15 Detection range of prototype

range of each sensor. Therefore, the authors will improve the detection part of the device. The image of the improved part is shown in Fig. 15.16.

In the improvement, the detection range is expanded by reciprocating the sensor in the x and y axial directions using two motors. Figure 15.17 shows the detection range of the improved device. The detection range is 60° vertically and 90° horizontally. This improvement has eliminated gaps in the detection range. In addition, more accurate detection of obstacles is possible. By conducting measurement that concentrated around the detected obstacle by the operation of the motor, it is possible to determine the shape and angle of the obstacle, and it becomes possible for the visually impaired to transmit accurate information of the obstacle.

Fig. 15.16 Image of improved equipment



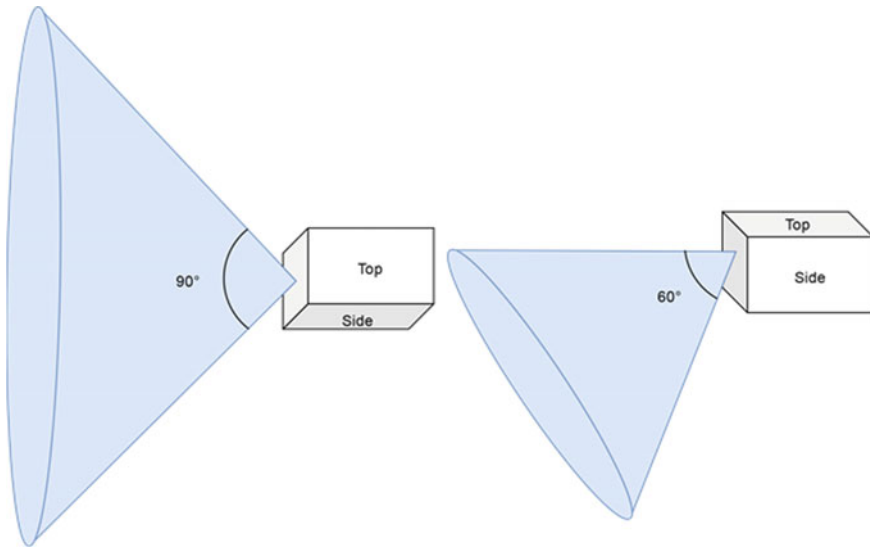


Fig. 15.17 Detection range of improved equipment

15.5 Conclusion

In this study, the authors investigated the present condition of the visually impaired and analyzed the spatial problems in the training of visually handicapped people (including congenital blind people and children with blindness) as a means of solving problems such as increasing tendencies of visually impaired people and lack of trainers. The authors are developing a novel support device for walking training for the visually impaired who can realize the above thing more than making it possible to experience the solution of the autonomous behaviors and various findings accompanying the self-sustaining behavior, making it cheap and easy to handle. The authors devised three sensor module mounting devices, designed a sensor module, and developed a highly safe system using two sensors. In the future, the authors will combine the whole research and production work done in each laboratory, produce prototypes of three sensor module mounting devices, and conduct evaluation tests at the blind school.

References

1. Japan Ministry of Health: Labour and welfare: annual report on government measures for persons with disabilities (2014)
2. Fukui, R.: Japanese guide dog training industries from the point of view from the world. *Nippon Hojyoken Kagaku Kenkyu* **2**(1), 22–25 (2008)

3. Bourne, R.R., Flaxman, S.R., Braithwaite, T., Cicinelli, M.V.: Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *LANCH Global Health* **5**(9), 888–897 (September, 2017)
4. Lions Clubs International: Sight first long range planning working group: childhood blindness position paper (2008)
5. Nakamura, M., Oshiro, E.: The current situation and issues for early nurturing support of young visually impaired children. *The bulletin of Akita University* (2014)
6. Japan Cabinet Office: Japan Cabinet Office Heisei 26th edition white paper of disabled people, Chapter 6, Section 1–8 (2014)
7. Tanai, M., Okura, M.: The present condition and problems of visually impaired person support technology: about single walking. *Measur contr* **34**(2), 140–146 (1995)
8. Japan Ministry of Health, Labor and Welfare: Guide to developing welfare equipment and nursing robots” Chapter 2, Section 3 (2014)
9. Japan Ministry of Health, Labor and Welfare: Welfare equipment and nursing robot practical application support project report Chapter 2 (2018)
10. Nursing Robot Online: What is the penetration rate of nursing care robots? Three factors that inhibit popularization (2019)
11. National Institute of Advanced Industrial Science and Technology: Overall outline of robot nursing care device development/introduction promotion project (2013)
12. SHARP: SHARP product lineup GP2Y0A710K0F data sheet
13. Sain Smart: Ultrasonic module distance measurement transducer sensor DC 5 V, US-015 data sheet
14. Renesas Electronics: Renesas SuperH RISC engine Family SH7124, SH 7125 Data Sheet (2009)

Chapter 16

Research on Video Mosaic Technology Based on Fully Mechanized Face of Coal Mine



Weihu Zhang, Zhihui Tao, and Xu Li

Abstract In the process of coal mine safety production, video surveillance is one of the important technical means to ensure production safety. However, due to the complicated environment of the coal mine itself, video surveillance has some defects, and the means are single, which cannot effectively complete the effective monitoring of the production environment. Therefore, a video monitoring method that meets the actual needs and is suitable for the complex environment of coal mines is proposed. In this paper, in view of the fact that the coal mine fully mechanized mining face needs the whole roadway visual field to monitor the safety of the whole roadway, the whole roadway monitoring can be completed by studying the panoramic video splicing technology, which can meet the production conditions and facilitate the safe mining activities of coal mines. The experimental results show that the method used in this paper has good robustness and timeliness and can be used for panoramic video stitching in coal mines.

16.1 Introduction

At present, video stitching algorithm has some development but it is not applicable to the actual needs of fully mechanized coal face [1]. The video splicing technology based on ORB feature proposed by Huang [2] effectively satisfies the video splicing requirements and reduces the number of feature extractions by dividing the ROI area of interest to improve the splicing rate. However, the coal mine underground environment is complex, and the dust is large which requires more real-time splicing effect. In 2017, Cheng et al. [3] proposed to stitch images through the SURF algorithm and the improved sampling consensus algorithm to solve the problem of ghosting and stitching. In 2017, Feng and Peng [4] and others proposed to improve the SURF algorithm descriptor to optimize the panoramic video splicing technology. The video

W. Zhang · Z. Tao (✉)
Xi'an University of Science and Technology, Xi'an, China
e-mail: txtao123@163.com

X. Li
Xi'an Hezhiyu Information Technology Corporation, Xi'an, China

© Springer Nature Singapore Pte Ltd. 2021
S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation,
Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_16

splicing effect is better but it is not applicable in coal mines. In 2018, Guan et al. [5] proposed improving the SURF algorithm for dimensionality reduction and dividing the region of interest by ROI to improve the efficiency of downhole video stitching. In 2018, Chen and Zhang [6] proposed the importance of panoramic video in coal mining and proposed a feasible method which provided a very constructive opinion for the application of video stitching in coal mines. In view of the current theoretical stage for coal mine video splicing technology in practical applications, there is a problem that the delay is large and more video frames need to be discarded. A method of quickly splicing video is proposed which can meet the actual needs and save more video. Frame information to avoid the loss of real video information in coal mine roadways [7]. And in the real environment of coal mine, the dust is large, the lighting conditions are poor, etc., the corresponding preprocessing technology is adopted, the video frame is denoised and preprocessed, and the video frame information is optimized to provide the basis for subsequent video stitching.

16.2 Introduction to Video Stitching Principle

16.2.1 Technical Route and Block Diagram

In this paper, the video splicing technology collects the video through the coal mine roadway camera, performs video preprocessing, and denoises the video information to provide more feature points extraction for the next video splicing. Then, the Moravec operator was used to extract feature points from different video frames, and FLANN method was used for coarse matching. The RANSAC algorithm performs the pure matching and the model parameter estimation on the extracted feature points, performs affine transformation [8], and then performs video frame fusion and output image. If a scene changes or timeout occurs during video stitching, the matching point is recalculated. If there is no change, the stitching is performed according to the previous affine transform coefficient. This method effectively improves the program running efficiency. The results show that the method has good real-time performance and can meet the actual production needs. The schematic is shown in Fig. 16.1.

16.2.2 Video Stitching Algorithm Flow Introduction

Moravec Operator

This paper detects feature points of video frames by Moravec [9] operator. This operator defines feature points as points with low “autocorrelation”. The algorithm detects each pixel of the image, uses a neighborhood around the pixel as a patch, and detects the correlation of the patch with other patches around it. This correlation is

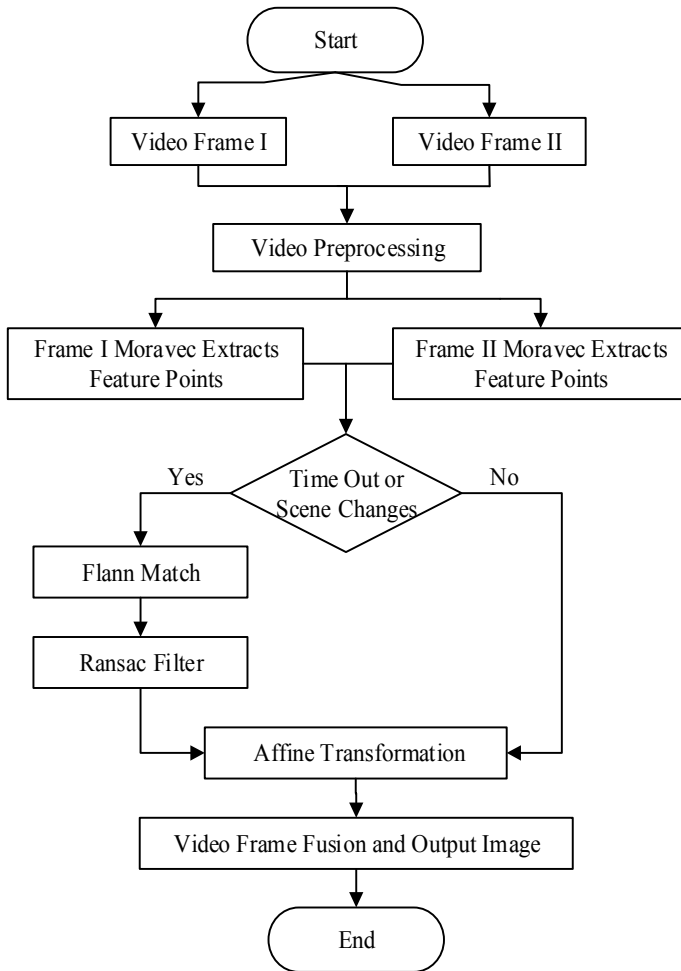


Fig. 16.1 Video mosaic flowchart

measured by the sum of squared differences (SSD) between the two patches, and the smaller the SSD value, the higher the similarity is.

When the Moravec operator calculates feature points, the following occurs: Pixels are in the smooth image area, the surrounding patches are very similar, and the SSD will be small; if pixels are on the edges, the surrounding patches will be in the direction orthogonal to the edges. There are big differences, which are similar in the direction parallel to the edges. SSD is small in one direction, and there are great differences in other directions. Pixels are feature points that change in all directions, and all patches around them are not very similar. In all directions, the SSD will be large. The algorithm steps are as follows.

In the $n * n$ window centered on the image pixel (x, y) , the grayscale variance V in the four directions (horizontal, vertical, diagonal and anti-diagonal) of the image, the average grayscale variation, is calculated using the following formula [9]:

$$\left\{ \begin{array}{l} V_h = \sum_{i=-k}^{k-1} (I_{x+i,y} - I_{x+i+1,y})^2 \\ V_v = \sum_{i=-k}^{k-1} (I_{x,y+i} - I_{x,y+i+1})^2 \\ V_d = \sum_{i=-k}^{k-1} (I_{x+i,y+i} - I_{x+i+1,y+i+1})^2 \\ V_o = \sum_{i=-k}^{k-1} (I_{x+i,y-i} - I_{x+i+1,y-i-1})^2 \end{array} \right. \quad (16.1)$$

where i represents the value of the change and accumulates k , $k = [n/2]$ represents the number of pixels to be calculated in the window, $[n/2]$ represents rounding, $I(x + i, y)$ represents the pixel value of the corresponding position of the image, and the SSD in each direction is calculated by such a method, and the minimum value is selected as the feature point response value.

According to the set threshold of the actual image, let the window traverse the image and use the feature point response value larger than the threshold as the candidate corner point.

Local non-maximum value suppression selects feature points. In a window of a certain size, the candidate feature points selected in the second part are removed from the point where the response value is not the maximum value, and the reserved feature points are the feature points of the region.

FLANN Match

The result of the feature points matching will result in a correspondence list of the two feature sets. The first set of features is called the training set, and the second set is called the query set. FLANN [10] trains a matcher to improve the speed of the match before calling the match function. The training phase is to optimize execution performance. The training set class will build an index tree of feature sets. Match each feature point of the query set with the training set matcher to find the best match; that is, match the trainer one by one from the features of the query set, that is, each feature set of the query set will have one the best match.

RANSAC Extracts Feature Points

The random sampling consensus algorithm uses the idea of iteration to estimate the correct mathematical model parameters from a set of normal data and anomalous data.

RANSAC [11] achieves its goal by iteratively selecting a random set of data in the data. The selected subset is assumed to be intra-office point and verifies with the following method:

1. Randomly select several data contained in the data and set it as an intra-office point;
2. Calculate the model suitable for the intra-site point;
3. Put the data that is not selected in the first step into the model calculated in the second step to determine whether it is an intra-office point;
4. Record the number of points within the office;
5. Loop iteration, repeat the above steps until you find the model with the most in-office points, so that the coarse matching feature points are filtered.

Parameter determination is as follows [11].

In the sampling consensus algorithm, it is necessary to determine the parameter t (which is used to determine whether the data is adapted to the threshold of the model), d (determine whether the model is suitable for the number of data in the data set), and k (the number of iterations of the algorithm). Where t and d are determined by specific experimental conditions, and k can be deducted by theory.

Assuming that the probability that each point is an intra-point is a , then a is equal to the number of intra-points divided by the total amount of data, but it is not actually known what the w is. The n th power of a indicates that the selected point of each of n times is the probability of the intra-point, and $1 - a^n$ indicates the probability that at least one of the n points is not the intra-point, indicating that a bad one is the calculated model. We use p to indicate a model with a good selection and evaluation.

$$1 - p = (1 - a^n)^k \quad (16.2)$$

Number of iterations:

$$k = \frac{\log(1 - p)}{\log(1 - a^n)} \quad (16.3)$$

Image Fusion

In this paper, the video frame is fused and outputted by using the fading-integrated fusion method. The method is to linearly assign weights to the pixels of the two images in the overlapping region [12].

$$\text{frame} = \text{frame1} * d + \text{frame2} * (1 - d) \quad (16.4)$$

The frame represents the pixel point of the fused video frame, and frame1 and frame2 are the pixel points of video frame to be splicing. d is the distance from a pixel point in the overlapping region to the boundary. In this method, the pixel points of the corresponding position video frame were fused into one pixel point, and different video frames were merged into one video frame with a larger observation field.

16.3 Result Analysis

To verify whether video splicing can be performed on scenes with poor lighting in real time, we deal with Figs. 16.2, 16.3, and 16.4 through the Moravec corner points and Figs. 16.5, 16.6, and 16.7 through the SIFT algorithm.

Through program simulation verification, we can get the difference between the number of extracted feature points and the extraction speed of Moravec corner detection operator and SIFT extraction algorithm (Tables 16.1 and 16.2).

According to the results obtained, we can see that the SIFT algorithm is better than Moravec in extracting effect, but the time required is longer than the corner detection time. Considering that in practical applications, we need more in the case of meeting the demand. Focus on real time, so choose the corner detection algorithm.

Fig. 16.2 Moravec test

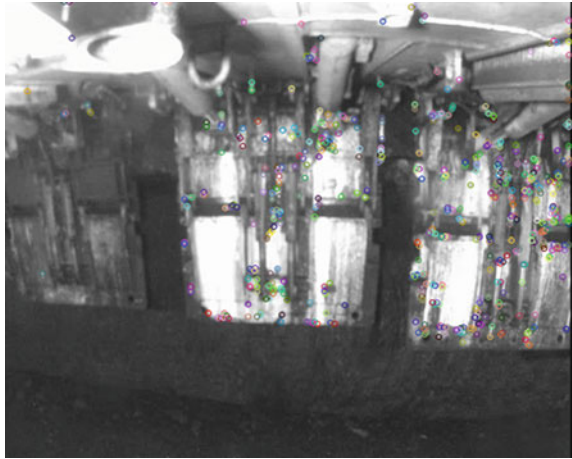


Fig. 16.3 Moravec test

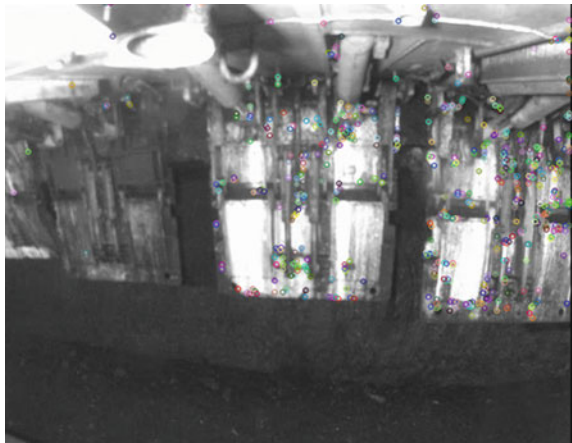
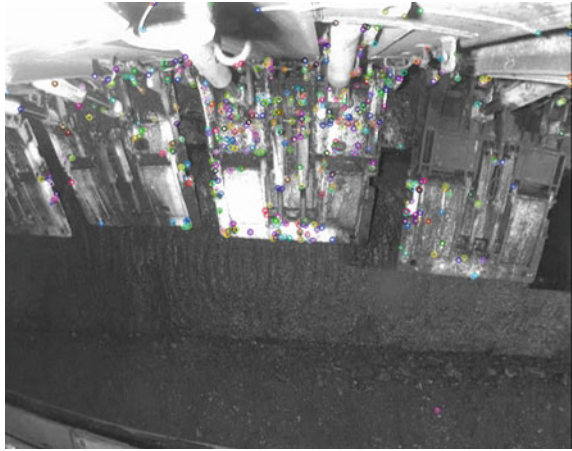


Fig. 16.4 Moravec test**Fig. 16.5** SIFT test

After matching by the FLANN algorithm, the feature matching result filtered by the RANSAC random sampling consensus algorithm. Select two images for splicing as an example, as Fig. 16.8 shows.

Since the video cameras of the intelligent working surface may have angular differences, the viewing angle correction is performed during the preprocessing process of the entire algorithm, thereby ensuring that the video images are in the same plane and ensuring the stitching effect [13]. Figures 16.9 and 16.10 indicate the stitching effect.

It can be seen from Figs. 16.9 to 16.10 that the algorithm effectively meets the actual requirements. The algorithm extracts feature points from the same part of different video frames and then stitches different video frames together by matching

Fig. 16.6 SIFT test

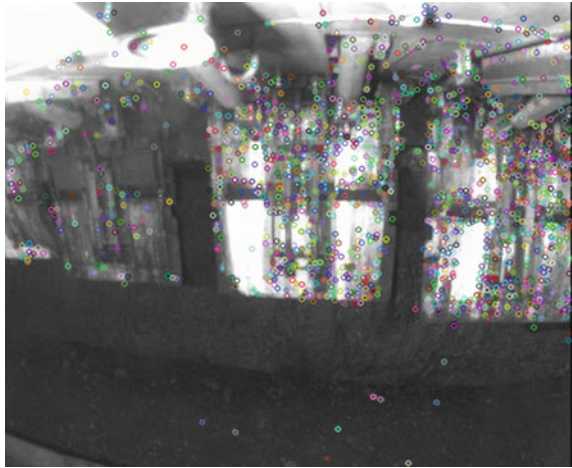


Fig. 16.7 SIFT test

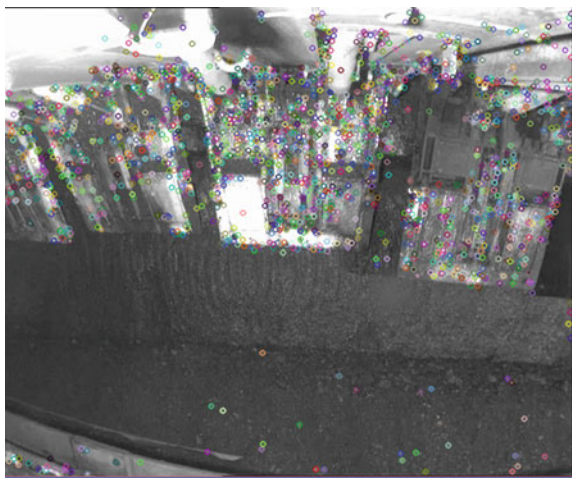


Table 16.1 Moravec test results

	Figure 16.2	Figure 16.3	Figure 16.4
Number of feature points	500	512	618
Speed (ms)	338.021	318.394	319.074

Table 16.2 SIFT test results

	Figure 16.5	Figure 16.6	Figure 16.7
Number of feature points	800	1772	2000
Speed (ms)	753.978	927.605	1052.39

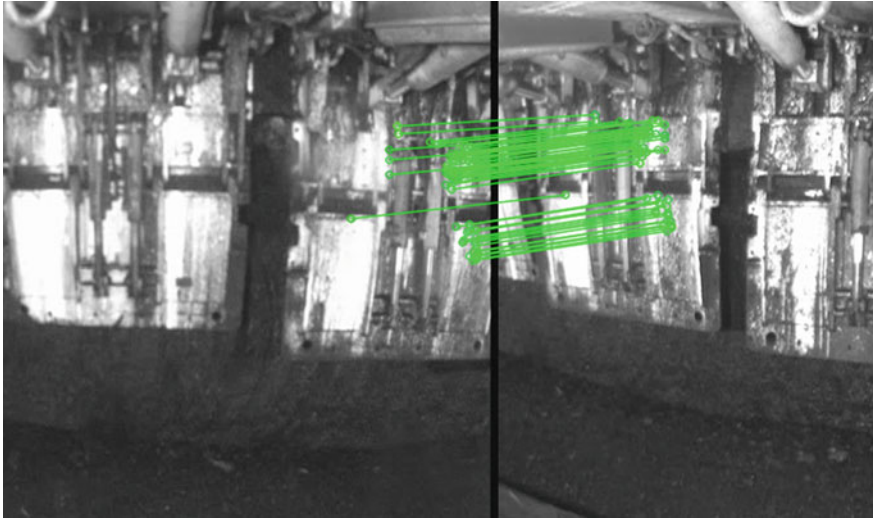


Fig. 16.8 Feature point extraction matching



Fig. 16.9 Stitching result 1

feature points. The results show that the splicing work can be performed on the complex coal mining face, so as to save more video frames and avoid the loss of real video information of coal mine roadway. Moreover, by preprocessing the image, the poor illumination condition can be solved and the roadway in the coal mine can be monitored better.

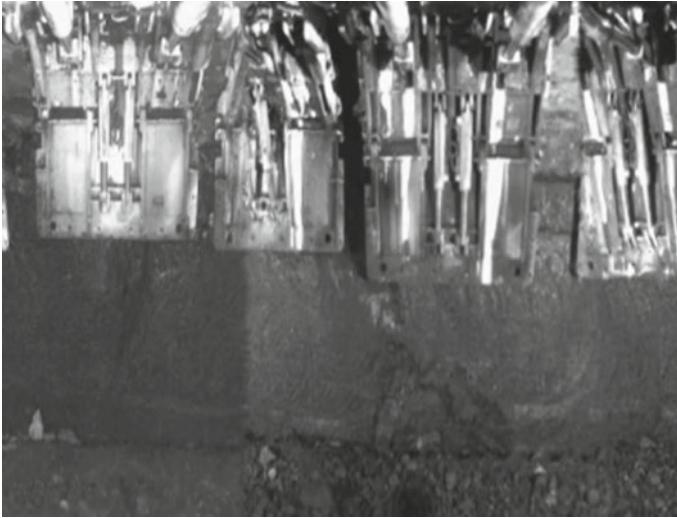


Fig. 16.10 Stitching result 2

16.4 Conclusion

Under the current experimental conditions, compared with the SIFT algorithm, the real-time performance is worse than the Moravec operator, the algorithm proposed in this paper can effectively realize the stitching effect, complete the design and meet the real-time performance, and whether to use the previously retained transform matrix coefficients to perform video in real time by judging whether the scene changes or recalculating the affine transformation matrix within a certain time. But we can also see that the stitching effect is not very good, so next we need to optimize the algorithm under the existing conditions, so that the stitching effect is more impressive.

References

1. Yuan, C.: Development trend of application of coal mine monitoring and control system. *Electr. Technol. Softw. Eng.* **8**, 100–101 (2019)
2. Huang, J.: Improved video real-time stitching technology based on ORB feature. *Electr. Autom.* **46**(3), 212–215 (2017)
3. Cheng, D., Li, H., Huang, X.: Video Mosaic algorithm based on improved random sampling consensus algorithm. *Work. Condition Autom.* **43**(8), 50–55 (2017)
4. Feng, Z., Peng, M.: Research on real-time panoramic video stitching technology. *Softw. Guide* **16**(8), 193–195 (2017)
5. Guan, Z., Gu, J., Zhao, G.: The downhole video mosaic algorithm based on improved accelerated robust feature. *Autom. Work. Autom. Conditions* **44**(11), 69–74 (2018)
6. Chen, L., Zhang, D.: Application of panoramic video splicing technology in coal mining. *Sci. Technol. Econ. Inf.* **26**(11), 10–10 (2018)

7. Zhao, X.: Analysis on the problems and counter measures in the application of coal mine safety monitoring system. *Eng. Technol. Res.* **4**(5), 28–29 (2019)
8. Yu, S.: Classification and application of planar affine transformation. *J. Ningbo University (Sci. Technol.)* **4**(1), 69–72 (2010)
9. Yu, F., Zhu, D., Hu, L., et al.: Research on feature extraction algorithm of aerial photography image. *Softw. Guide* **18**(8), 66–70 (2019)
10. Wang, J., Zhou, Z.: Research on feature extraction based on SIFT image and FLANN matching algorithm. *Comput. Measur. Control* **26**(2), 175–178 (2018)
11. Li, J., Zhang, F., Cui, H.: A homography matrix estimation method based on improved Ransac algorithm. *Softw. Guide* (2019)
12. Shi, M.H.: Summary of the development of image fusion technology. *Comput. Er* **9**, 27–29 (2019)
13. Xiang, H., Bai, X.: Photo image correction technology scheme. *Technol. Econ. Guide* **7**, 33–34 (2016)

Part III
Data Processing and Security

Chapter 17

Optimization Scheme for Traceability of Distributed Denial of Service Attacks Based on Dynamic Probability Packet Marking



Li Chen and Jun Yao

Abstract Aiming at how to improve the efficiency of path reconstruction and reduce the false positive rate, this paper proposes a new optimization scheme based on dynamic probabilistic packet marking (DPPM). First, the IP address information of the router is marked in the form of dynamic probability to some available fields of the packet header. Secondly, in order to reduce the false positive rate during path reconstruction, the IP address of the router and the corresponding hash value are divided into two pieces. Finally, the experimental results show that the improved probability packet optimization algorithm requires 30% of the original convergence packet compared with the dynamic probability packet labeling algorithm; when the attack path length. When it is greater than 12 hops, the false positive rate of the optimization scheme of this paper is 40% of the dynamic probability packet marking scheme. Through experimental comparison, the scheme has good a performance based on DDOS attack traceability.

17.1 Introduction

Distributed denial of service (DDOS) attacks have become a major threat to Internet users due to their ease of implementation, low attack costs, and high risk [1]. DDOS attacks exploit the shortcomings of the TCP/IP protocol during the transmission of data packets. The attacker usually changes the packet header information to use a fake IP address, thereby hiding the true identity of the attacker and posing a huge challenge to the network defense [2]. In order to effectively defend against DDOS attacks, you need to trace the source of the attack. However, most DDOS attacks currently use fake IP addresses to attack, hiding the true identity of the attacker. So IP traceability technology based on DDOS attacks has become a hot issue in network security research.

L. Chen (✉) · J. Yao
College of Communication and Information Engineering, Xi'an University of Science and Technology, Xi'an 710000, China
e-mail: chenli_sun@163.com

In paper [3], a dynamic probability packet labeling scheme for fragmentation correlation of DDOS attack source is proposed, but this scheme is too complicated in the path reconstruction process, and the number of data packets required for path reconstruction is too large. In paper [4], the information of the router is tagged to the data domain of ICMP to achieve traceability. This method needs to generate a large number of ICMP message packets to ensure the success rate of reconstruction. In this paper, the data packet passing through the router node is marked in the form of a dynamic probability mark. The 32-bit IP address information of the router node is divided into two pieces, which are high 16 bits and low 16 bits, respectively. Then convert the 16-bit address information into an 8-bit hash value through the hash function. The hash function is used to convert the 8-bit hash value to ensure the validity of the IP address during the reconstruction of the attack path. Finally, the experimental results show that the scheme can effectively reduce the number of data packets required for path reconstruction, reduce the false positive rate of path reconstruction, and solve the uncertainty of attack source.

17.2 Dynamic Probability Packet Marking Algorithm

17.2.1 Dynamic Probabilistic Packet Marking (DPPM)

In the DPPM scheme, the router in the attack path has a label probability $p = 1/i$ (i representing the number of hops between the current router and the attacker) to mark the data packet w . At this time, the probability that the data packet is finally marked by any router is $1/d$ [5]. Then determine the probability p of the tag based on the value of the TTL field. The initial value of TTL is in the set $\{32, 64, 128, 255\}$ [6]. The method for determining the probability of a router in an attack path is as follows (w_{ttl} represents the value of the *ttl* field in packet w):

1. When $w_{ttl} < 32$, the probability of labeling $p = 1/(32 - w_{ttl})$ at this time;
2. When $32 < w_{ttl} < 64$, the probability of labeling $p = 1/(64 - w_{ttl})$;
3. When $64 < w_{ttl} < 128$, the probability of labeling $p = 1/(128 - w_{ttl})$;
4. When the value of w_{ttl} in the data packet is the flag probability, $p = 1/(255 - w_{ttl})$;

The algorithm interleaves 32-bit router address information and 32-bit hash value in bits and combines them into 64-bit addresses. The odd-numbered bits represent the address information of the router, and the even-numbered bits represent the hash check code. The address information is divided into eight slices on average, each of which is 8-bits, and the corresponding offset is 0–7.

17.2.2 The Problem of Dynamic Probability Packet Marking

Problem 1: The router's IP and its corresponding hash value are divided into eight pieces. In the DPPM scheme, each packet carries only 1/8 of the total information, so at least eight packets need to be collected to reassemble the IP address information of one router node in the attack path. In the process of attack path reconstruction, a large number of data packets are needed and it is very time consuming.

Problem 2: The selection of the marker probability in the DPPM scheme depends entirely on the TTL value in the IP packet, but the attacker can tamper with the TTL value to destroy the attack source. For example, if the attacker changes the value of the TTL field of the IP header to 140, the router r_i closest to the attacker will consider that the packet has passed 115 routers, that is 115 hops ($255 - 140 = 115$), so the labeling probability of the node is for 1/115, the next router will mark the packet with a probability of 1/127, 1/128, ..., but the packet actually passes through only 1 hop. Tampering of TTL values can affect the accuracy of attack path reconstruction.

17.3 Optimization Scheme of Dynamic Probability Packet Marking

17.3.1 Tag Domain Settings

In paper [7], it can be concluded that the probability that IP packets are fragmented will not be greater than 0.25%, and fragmentation can be avoided by automatic MTU discovery [8]. Therefore, in the optimization scheme of this paper, the identification, flags, and offset fields in the packet header are used to store the tag information, which does not affect the forwarding and routing of the packet by the router. This paper selects the 16-bit identification field, the upper 1 of the flags field, and the 13-bit offset field. The tag space is shown in Fig. 17.1.

In this paper, the 30-bit mark space is divided into four fields, corresponding to different tag information. The explanations of the four fields are as follows:

1. IP_frag (16bits): used to store the IP fragment information of the router.
2. hash_frag (8 bits): stores the hash value corresponding to the IP fragment of the router and plays a check role when the path is reconstructed;
3. offset (1 bit): indicates the offset of the fragment of the router's IP address. The two values of 0, 1 indicate the corresponding fragment offset;
4. distance (5 bits): indicates the number of hops between the router and the victim from the attack path; the literature [9] shows that the data packets transmitted through the network will not exceed 32 hops, so distance is enough for the domain to use 5-bit ($2^5 = 32$).

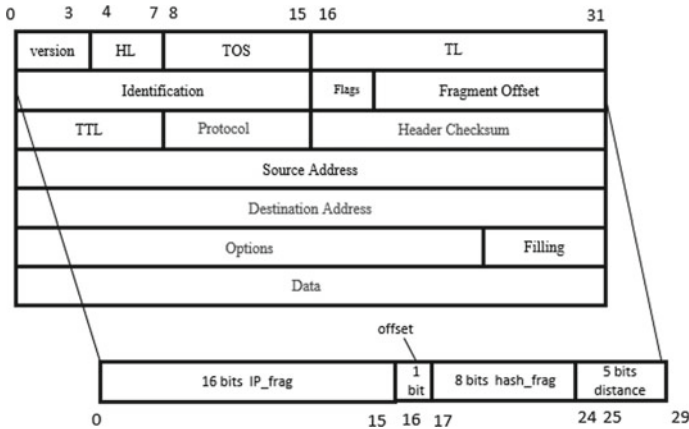


Fig. 17.1 Selection of marker space in the optimization scheme

17.3.2 Generation of Tag Information

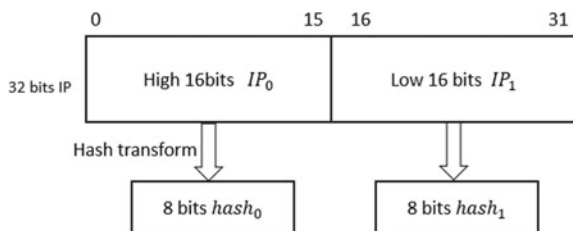
This scheme divides the tag information into two blocks. The processing method of the router IP address is shown in Fig. 17.2.

First: Divide the IP address of the 32-bit router into two blocks, which are the IP_0 of the high 16 bits and the IP_1 of the low 16 bits. When the router decides to mark the data packet, it will correspond according to the value of the offset field. The fragmentation information value is stored in the IP_frag domain.

Second: Two hashes are used to calculate two hash fragments by using the 8-bit hash function, which are $hash_0$ and $hash_1$. $hash_frag_i$ stores $hash_{i+1}$ value (i indicates the corresponding offset of the shard), $hash_frag_i$ and $hash_i$ correspond to the following relationship: $hash_frag_0 = hash_1$; $hash_frag_1 = hash_0$.

Third: When the data packet is marked, the marking device randomly generates a number i between the set $S = \{0, 1\}$, and the i group data is marked according to the marking algorithm.

Fig. 17.2 Processing method of tag information



17.3.3 Marking Algorithm

In order to protect against the impact of path reconstruction on packets that have been tampered with TTL values, this paper sets a uniform initial value for the TTL field, which is represented by T_{init} . When the router node detects that the value of the TTL field in the data packet is greater than T_{init} , the value of $T_{init} - 1$ is written into the TTL field at this time, and the distance between the router and the attacker is considered to be only 1 hop, because the data packet will not exceed 32 hops during transmission. The value of T_{init} is preferably 64 or 32.

When $w_{ttl} < 32$ in the data packet, the probability of labeling is $p = 1/(32 - w_{ttl})$;

When $32 < w_{ttl} < 64$ in the data packet, the probability of the flag is $p = 1/(64 - w_{ttl})$;

The specific description of the marking algorithm is as follows:

1. Obtain the IP address information of the current device, and then divide the obtained IP address information into two pieces;
2. Check the TTL value of the received IP packet header. If $T_{init} < w_{ttl}$ ($T_{init} = 64$, write $T_{init} - 1$ to the TTL field, $w_{ttl} < T_{init}$ goes directly to the next step;
3. Determine the mark probability p . The router obtains the value of the TTL field in the data packet arriving at the node w_{ttl} . When $w_{ttl} < 32$, then $p = 1/(32 - w_{ttl})$; when $32 < w_{ttl} < 64$, $p = 1/(64 - w_{ttl})$;
4. The router node randomly generates a number x between $[0, 1]$ and an integer i of 0 or 1, and compares x with the calculated probability p . If $x < p$, the router needs to mark the data packet and perform step 5; if not, perform step 6;
5. Process the IP address information of the router, and obtain IP_0, IP_1 , $hash_0$, and $hash_1$. According to the random integer i , add ip_i to the IP_addr domain, and add $hash_{i+1}$ to the hash_frag domain, and set the distance field to 0, the fragment offset t is set to i , and the mark is completed;
6. First, check the value of the w.distance field in the data packet. If w.distance = 0, it means that the data packet has been marked by the previous node. In this case, the value of the w.offset field i needs to be read from the data packet; XOR ip_i of this node information with the value of the IP_addr field in the packet, and the result of XOR is rewritten into the IP_addr field in the packet; then the node is $hash_{i+1}$. The value is XORed with the value of the hash_frag field in the packet, the result of the XOR is written to the hash_frag field of the packet, and finally the value of the distance field in the packet is incremented by 1. The IP_addr field and the hash_addr field record the side information between adjacent nodes; if w.distance $\neq 0$, the value of the distance field is directly added to 1 to complete the marking algorithm.

17.3.4 Path Reconstruction Algorithm Running on the Victim Host

1. On the victim host, first, divide the data packet into different sets according to the value of the distance field in the attack data packet (the distance domain is the same as a set); secondly, in the set with the same distance domain value, and then according to the offset the value of the domain i ($i = 0, 1$) is divided into two small sets;
2. Starting from the distance domain $w.distance = 0$, reorganize the IP fragment address. For the packet w , if its $w.offset = 0$, convert the value of its IP_addr field to the same hash function as when it was marked, and comparing it with the hash_frag field value in the $w.offset = 1$ set. If they are the same, they may be from the same node of the router. Secondly, the value of the IP_addr field in the $w.offset = 1$ set is compared with the value of the IP_addr field in the $w.offset = 0$ set by the hash function conversion. After double comparison, if they are the same, they are the same router node information, thus obtaining the complete IP information of a router node on an attack path.
3. Starting from $w.distance = 1$, the values of the IP_addr field and the hash_frag field in the data packet are XORed with the values of the corresponding fields in the data packet of $w.distance = 0$. The $w.distance = 0$ packet stores the last hop router IP address information, and $w.distance \neq 0$ stores the XOR edge information of the neighboring router node, according to the nature of the exclusive XOR operation $a \oplus b \oplus a = b$, in turn, the fragment address information of the previous hop router can be restored, and then the fragment address is reassembled to obtain the complete router node address.
4. Start from $w.distance = 0$ until the largest $w.distance$ node information is restored, thus completing the reconstruction of the entire attack path.

17.4 Experimental Simulation Results

To verify the performance of this article's optimization scheme, this paper uses network simulation software NS2 to analyze and compare the performance of the optimization scheme. Three experiments of convergence, false positive rate, and uncertainty of attack source are selected for simulation experiments.

17.4.1 Convergence Performance Comparison

The convergence in the traceability scheme is the number of packets required to reconstruct the attack path at the attacked host when the length of the attack path is different. In this paper, the router node address information is divided into two pieces,

and one node information can be combined by two data packets when reconstructing the path. In the DPPM scheme, the router IP address information is divided into eight pieces, and the data packet required to reconstruct the attack path is expected to be $E0(N)$ [10]:

$$E0(N) = \frac{K * \ln(kd)}{1/d} = kd * \ln(kd) \tag{17.1}$$

In the formula (17.1), k is the number of fragments of the IP address information, and d is the number of router nodes on the attack path.

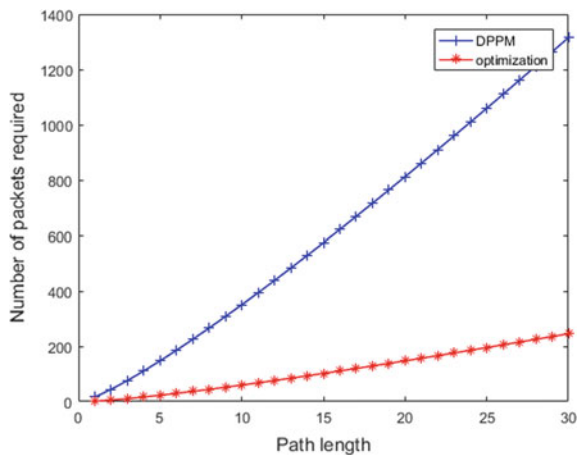
In the optimization scheme of this paper, the node IP address information is only divided into two pieces, and the data packet required to reconstruct the attack path is expected to be $EI(N)$:

$$EI(N) = 2 * d * \ln(2d) \tag{17.2}$$

In order to test the convergence performance of this optimization scheme, the attack path lengths $d = 1, 2, 3, \dots, 0.32$ were selected and simulated, and compared with the DPPM scheme, the attack path length and the number of data packets required to reconstruct the attack path were obtained. Simulation diagram between. As shown in Fig. 17.3.

It can be concluded from Fig. 17.3 that compared with the DPPM scheme, the number of packets required for the optimization scheme of this paper is 30% of the number of packets required by the DPPM. This means that the number of packets required during the reconstruction of the attack path is small and the reconstruction efficiency is high.

Fig. 17.3 Relationship between path length and number of packets required to reconstruct the attack path



17.4.2 False Positive Rate

The false positive rate is an indicator used to measure the robustness of the IP traceability scheme. The false positive is that the attack path reconstructed at the attacked host is not the actual attack path. The literature [11] shows that in a hash function of length h , the probability of receiving a candidate edge of any construct is $\frac{1}{2^h}$. When there are y attackers, at any d , the worst will probably be y different routers, so the false positive rate of arbitrary reconstruction at distance d is P_e [12] as follows:

$$P_e = 1 - \left(1 - \frac{1}{2^h}\right)^{y^k} \tag{17.3}$$

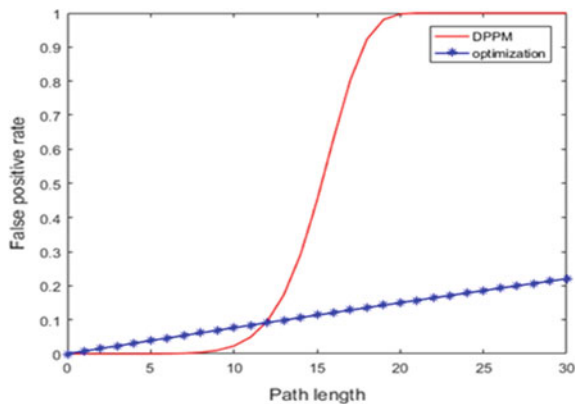
The y^k in the formula (17.3) represents the number of times the combination is required in the worst case. It can be obtained by formula (17.3), the false positive rate of DPPM $P_e(\text{DPPM}) = 1 - \left(1 - \frac{1}{2^{32}}\right)^{m^k}$, where m represents the number of attack paths.

In the optimization scheme of this paper, the node IP address is divided into two pieces, which requires two hash operations. That is to say, when there are m attack paths on the distance i , the false positive rate of finding adjacent fragments is $\frac{m}{2^8}$. A shard can only get two complete comparisons to get the complete address information at most, so the false positive rate of an IP address is obtained during the reconstruction process $P_e = 1 - \left(1 - \frac{m}{2^8}\right)^2$.

In order to test the false alarm rate performance of the traceability of the scheme, the length of the attack path was simulated from 1 to 32, and compared with the DPPM scheme, the relationship between the false positive rate and the attack path was obtained, as shown in Fig. 17.4.

It can be concluded from Fig. 17.4 that when the attack path is less than 12, the false positive rate of the optimization scheme is higher than DPPM, but the false positive rate of the optimized scheme is also kept at a low level; when the attack

Fig. 17.4 Relationship between path length and false positive rate



path is greater than or equal to 12, the false positive rate of the optimization scheme is significantly smaller than DPPM. When the attack path length is close to 20, the false positive rate of the DPPM scheme is close to 1, and the false positive rate of the optimization scheme is only close to 0.15. The false positive rate of the optimization scheme is basically 40% of DPPM. This shows that for the DDOS attack source tracing, the better the false positive rate of the optimization scheme proposed in this paper is when the attacker is farther away from the victim.

17.4.3 Attack Source Uncertainty

In the DPPM scheme, the value of the TTL field in the data packet may have been tampered with by the attacker, affecting the labeling probability of the data packet, which leads to the reconstruction of the attack path in the traceability process, resulting in uncertainty of the attack source. However, in the optimization scheme, the initial value T_{init} of the TTL domain is specified, and when the router node in the attack path detects that the value of the TTL field in the data packet is greater than a predetermined initial value, the value is rewritten into the TTL domain, and the node is considered to have only 1 hop from the attacker. At this point, the uncertainty of the attack source can be well reduced.

17.5 Conclusion

Based on the dynamic probability packet marking DDOS attack source optimization scheme, the optimization scheme reduces the router IP address fragmentation from eight slices to two slices and adopts the strategy of double checksum between two slices by hash code, which greatly reduces the path reconstructing the number of packets required, on the other hand, reduces the complexity and false positive rate of fragmentation reorganization. Finally, through the experimental simulation comparison, the optimization scheme of this paper is much better than the DPPM scheme in terms of convergence and false positive rate.

Acknowledgements CERNET Innovation Project (NGII20160301).

References

1. Foroushani, V.A., Zincir-Heywood, A.N.: TDFA: traceback-based defense against DDOS flooding attacks. In: IEEE International Conference on Advanced Information Networking and Applications (2014)
2. Xiaohu, L., Mingqing, Z., Jun, T., et al.: Design of intrusion tracking model for distributed DDoS attack source. *J. Inf. Eng. Univ.* **15**(2), 242–247 (2014)
3. Jin, H.: Design and implementation of network attack traceability system. Beijing University of Posts and Telecommunications (2017)
4. Xiuzhen, C., Shenghong, L., Yidong, L., et al.: A new method for multi-label IP return tracking for denial of service attacks. *J. Xi'an Jiaotong Univ.* **47**(10), 13–17 (2013)
5. Jin, N.: Research on packet marking optimization scheme under DDOS attack. University of Electronic Science and Technology (2013)
6. Chen, D.: Research on DDOS attack source tracking technology based on grouping marker. Hunan University (2017)
7. Stoica, I., Zhang, H.: Providing guaranteed services without per flow management. In: Conference on Applications. ACM (1999)
8. Jing, Y., Tu, P., Wang, X., et al.: A DDoS attack source tracking model based on reverse confirmation. *Comput. Eng.* **2**, 127–129 + 154 (2007)
9. Theilmann, W., Rothermel, K.: Dynamic distance maps of the Internet. In: Proceedings of the 2000 IEEE INFOCOM Conference (March 2000)
10. Sun, S.: Research on DdoS attack source and detection method. Liaoning University (2016)
11. Savage, S., Wetherall, D., Karlin, A., et al.: Practical network support for IP traceback. In: Proceedings of ACM SIGCOMM, pp. 295–306. Stockholm, Sweden (2000)
12. Yang, Y.: Research and algorithm implementation of packet-based DRDoS attack source tracing. East China Normal University (2016)

Chapter 18

Processing Initial Data for the Agent-Based Model of the Russian Federation Spatial Development



Aleksandra L. Mashkova

Abstract In our research, the preparation of initial data for the simulation model of the spatial development of the Russian Federation is being conducted. Within the model database, information about population, economy, and social institutions from the official open sources: Web sites of the ministries, federal, and regional statistical yearbooks, is integrated. While information about population and education is quite detailed, referring to the production system it is disaggregated and needs processing to be converted into required structure. In this paper, structure of initial data for simulating production in the model and methods of matching it with available open data are presented. Iterative proportional fitting technique is implemented for aggregating information about employment and equipment of organizations in different regions. Within the procedure of generation, the model database is filled with objects created on the basis of initial data.

18.1 Introduction

Our research is aimed at developing a tool for evaluating alternative managerial decisions by simulating different scenarios of the spatial development of the Russian Federation. Creation of such a model requires integration of population, economic structures, and social institutions to reflect a wide range of control actions, including tax, monetary, and investment policies. Significance of practical application of the model directly depends on completeness and detailing of its input data, which ensure that the model environment corresponds to the real world and the forecast estimates obtained on its basis are adequate.

Initial modeling data on the population, production, infrastructure, financial state of organizations, and households can be found in federal and regional statistical

A. L. Mashkova (✉)

Orel State University named after I.S. Turgenyev, Komsomolskaja St. 95, 302026 Orel, Russian Federation

e-mail: aleks.savina@gmail.com

Central Economics and Mathematics Institute, Russian Academy of Sciences, Nakhimovskiy Av. 47, 117418 Moscow, Russian Federation

© Springer Nature Singapore Pte Ltd. 2021

S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation, Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_18

227

collections [1, 2], open data of ministries [3–5], and results of sociological surveys [6, 7]. Since this information was collected by different sources and for different reasons, it needs processing to be converted into required structure of initial modeling data tables.

Of particular difficulty is the search for information in a spatial context. To analyze demographic processes, one can rely on census data [1], which provides a full range of necessary information about population in regional, urban, and rural context. Detailed information on higher and secondary vocational education is presented in the reports of the Ministry of Education and Science [5]. For economic structures, obtaining information in a spatial context is fraught with considerable difficulties. The main source of information is the Federal State Statistics Service Web site [2] and its annual collections, but information provided in them is significantly generalized. More detailed information is available in regional collections, but there are a few difficulties that determine the need for initial data preprocessing:

1. Detailing of sector structure of gross regional product is presented in collections for only 14 regions out of 82.
2. In these 14 collections, some economic activities are not detailed; for example, economic activity ‘personal services’ is detailed only in collections of Karelia and Arkhangelsk.
3. Often, 2–3 related sectors are combined into a complementary set (e.g., food and tobacco industry); moreover, composition of complementary sets differs from one regional collection to another.

The purpose of the presented work is integration of data from different sources, which retains high detail for the regions that presented the most complete information and uses proportional fitting techniques for others. The solution of the indicated problems is only the first step in the study, since in the future it is planned to integrate data on individual organizations and corporations within the model database. The task of statistical data preprocessing, however, is urgent for the development of the demo version of the model and debugging of its algorithms.

18.2 Research Methodology

For modeling spatial development of the Russian Federation agent-based approach was chosen [8], which includes heterogeneity, bounded rationality, and global dynamics as a result of micro-level agents’ interactions. Based on these assumptions, agent-based models become computer laboratories to test the effects of policies on macroeconomic and spatial dynamics [9]. Agent-based models have already been implemented in different policy areas such as fiscal [10], monetary [11], financial [12], and labor market policy [13].

Loading real data into agent-based models remains an urgent task. There are examples of solving this task for different regions, for example, for Leeds, the UK [14], and Saint Petersburg, Russia [15]. Due to the regional scale of these models, only

regional management measures are simulated in these models, while macroeconomic effects are treated as the external environment.

Within creation of the model of spatial development of the Russian Federation sex-age structure of the population, infrastructure, production, administration, and educational institutions in each region are reconstructed. Dynamics of the system is simulated through behavior of agents and decisions of organizations.

In this paper, methods of processing initial modeling data used for reconstructing production system of the Russian Federation in the spatial context are considered. For data processing, iterative proportional fitting (IPF) is used, which is a procedure for assigning values to internal cells based on known marginal totals in a multidimensional matrix [16]. In different branches of academic research, IPF is called bi-proportional fitting or RAS algorithm [17]. IPF is widely used in spatial microsimulation studies for integrating geographically aggregated data sources [18].

18.3 Structure of the Agent-Based Model

In our research, the agent-based model is being developed, which simulates demography, production, employment: financial, education, and budgetary system (see Fig. 18.1). Population and organizations are related to regions of the Russian Federation [19].

In the module, ‘population’ demographic processes are reflected, including birth and death, as well as creation of new households connected with marriages and divorces. Within other modules, agents act as students, employees and employers, taxpayers and consumers.

Organizations in the model are aggregated: one organization in the model responds to a set of organizations of one economic sector in the region [20]. There are commercial, budgetary, and financial organizations in the model. Commercial organizations interact within ‘production and service’ module. Financial institutions in the model

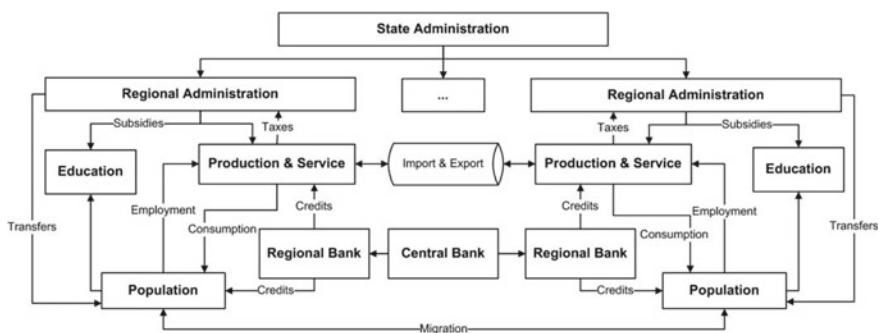


Fig. 18.1 Main processes in the agent-based model of spatial development of the Russian Federation

are the Central Bank (that sets the interest rate and issues bonds) and regional banks, which deal with deposits and credits of households and organizations. The state administration is responsible for tax collection, social transfers, and federal investment programs. The regional administration implements their functions through budgetary organizations in social security, education, and medicine [19].

18.4 Data Processing Algorithms

In the further paragraphs, methods of processing data from the portal of the Federal State Statistics Service [2], federal and regional statistical yearbooks are considered. Information on regional production, employment, property, and equipment is presented in the economic activities scale, which is much less detailed than the sector scale used in the federal input–output table [19]. Organizations in the model, however, represent sector structure of regional economies. Table 18.1 shows that available data requires preprocessing to match initial modeling data structure.

18.4.1 Production and Service

Firstly, production volume of each sector in the complemented set $P_r^{\sum_{i=1}^k s_i}$, presented in the regional statistical yearbook, is calculated.

$$\widehat{d}_r^s = \frac{P^s}{\sum_{i=1}^k P^i} \quad (18.1)$$

\widehat{d}_r^s —share of sector s in the complemented set in region r ; P^s , P^i —gross product of sectors in Russia, k —number of sectors in the complemented set.

$$P_r^s = P_r^{\sum_{i=1}^k s_i} * \widehat{d}_r^s \quad (18.2)$$

P_r^s —product of sector s in region r ; $P_r^{\sum_{i=1}^k s_i}$ —product of the complemented set of sectors.

Results of processing data about production in economic activity ‘transport and communication’ in region 1 (Belgorod) are presented in Table 18.2. Supplementary sets are: ‘land transport service’ and ‘water transport service’; ‘additional transport service’ and ‘communication’.

Table 18.1 Initial modeling data for simulating spatial distribution of production

Initial modeling data tables	Table content	Available open data tables	Table content
Production	P_r^s —product of sector s in region r	Input–output table	P^s —product of sector s , billion RUR
		Aggregated gross regional product structure	P_r^e —product of economic activity e in region r
		Detailed gross regional product structure	P_r^s —product of sector s in region r
Export and import	$e^s; i^s$ —share of export and import in the output of sector s	Input–output table	$E^s; I^s$ —volume of export and import in sector s
Supply	a_{ij} —volume of sector i product used in production of a unit of sector j	Input–output table	x_{ij} —cost of sector i product used in total production of sector j
		Product of sectors in physical terms	P_r^s —product of sector s in physical terms (standard units)
Property, plants and equipment	PE_r^s —volume of property, plant, and equipment in sector s in region r	Property, plant, and equipment in different regions	PE_r —volume of property, plant, and equipment in region r
		Property, plant, and equipment of organizations of different economic activities	PE^e —volume of property, plant, and equipment in economic activity e
Employment	L_r^s —number of employees in sector s in region r	Number of employees in different regions	L_r —number of employees in region r
		Number of employees of organizations of different economic activities	L^e —number of employees in economic activity e

Total product of sector s in regions from 1 to u with detailed statistical data is:

$$V_s^{\text{det}} = \sum_{r=1}^u P_r^s \quad (18.3)$$

V_s^{det} —total product of sector s in regions from 1 to u ; u —number of regions with detailed statistical data about sector s ; P_r^s —product of sector s in region r .

For different economic activities u accepts values from 2 ('personal services') to 14 (manufacturing). In economic activity 'transport and communication' $u = 11$.

Table 18.2 Results of data processing (transport and communication in Belgorod region)

Sectors in economic activity 'transport and communication'	$\sum_{i=1}^2 s_i$, billion RUR	P^s , billion RUR	\hat{d}_r^s	P_1^s , billion RUR
Land transport service	25.6	5596.5	0.97	24.8
Water transport service		191.3	0.03	0.8
Air and space transport service	0	928.9	0	0
Additional transport service and communication	13.8	2669.5	0.58	8.0
Communication		1971.0	0.42	5.8

For each sector its total product in other regions is:

$$V_s^{\text{ost}} = V_s - V_s^{\text{det}} \quad (18.4)$$

V_s^{ost} —total product of sector s in regions from u to 82.

For regions from u to 82, approximate share of sector s in the corresponding economic activity type is calculated:

$$\widehat{d}^{s-e} = V_s^{\text{ost}} / \sum_{r=u+1}^{82} P_r^e \quad (18.5)$$

\widehat{d}^{s-e} —approximate share of sector s belonging to economic activity e ; P_r^e —product of economic activity e in region r .

For further operations, share of sectors in economic activities in different regions is calculated:

$$d_r^{s-e} = \frac{P_r^s}{P_r^e} \quad (18.6)$$

d_r^{s-e} —share of sector s in economic activity e in region r , P_r^s —product of sector s in region r , P_r^e —product of economic activity e in region r .

18.4.2 *Export and Import*

Share of export in the output of sector s is:

$$e^s = \frac{E^s}{P^s} \quad (18.7)$$

e^s —share of export in the output of sector s ; E^s —volume of export in sector s ;
 P^s —product of sector s .

Share of import in consumption of product sector s is:

$$i^s = \frac{I^s}{P^s} \quad (18.8)$$

i^s —share of import in sector s ; I^s —volume of import in sector s ; P^s —total consumption of product of sector s .

Since more detailed information is not available, it is assumed that share of export in output and import in supply of each organization is equal to the average share for the sector, which the organization belongs to.

18.4.3 *Supply*

In the federal input–output tables, deliveries between sectors are presented in monetary terms. In order to preserve quantitative relationships in the real economy in the period of changes in currency exchange rates and inflation, accounting of production in standard units for each sector is used. For this task, the price of a standard unit of each sector is calculated:

$$\text{Price}^s = \frac{P^s}{Pr^s} \quad (18.9)$$

Price^s —price of a standard unit of sector s ; P^s —product of sector s in monetary terms; Pr^s —product of sector s in physical terms (standard units)

Supply matrix in the model should be presented in physical terms as well:

$$a_{ij} = \frac{x_{ij}}{Pr^j * \text{Price}^i} \quad (18.10)$$

a_{ij} —element of supply matrix (volume of sector i product used in production of a unit of sector j); x_{ij} —cost of sector i product used in total production of sector j (presented in the federal input–output table); Pr^j —product of sector j in physical terms; Price^s —price of a standard unit of sector s .

18.4.4 Property, Plants, and Equipment

For aggregating information about property, plant, and equipment of organizations, iterative proportional fitting algorithm (IPF) is used. Initial data includes PE_r —total volume of property, plant, and equipment in region r (table ‘property, plant, and equipment of organizations in different regions’ presented on Federal State Statistics Service Web site) and PE^e —total volume of property, plant, and equipment in economic activity e (table ‘property, plant, and equipment of organizations of different economic activities’ presented on Federal State Statistics Service Web site).

The aim of calculations is defining PE_r^e —total volume of property, plant, and equipment in economic activity e in region r , so that:

$$PE^e = \sum_{r=1}^{82} PE_r^e \quad (18.11)$$

$$PE_r = \sum_{e=1}^{15} PE_r^e \quad (18.12)$$

$$PE = \sum_{r=1}^{82} PE_r = \sum_{e=1}^{15} PE^e \quad (18.13)$$

where $PE = 74,662.4$ billion RUR—total volume of property, plant, and equipment in Russia [7], PE^e and PE_r are presented in appropriate line and column in Table 18.3.

At first, preliminary volume of property, plant, and equipment in economic activities in each region is calculated:

$$PE_r^e(\text{preliminary}) = PE^e * P_r^e \quad (18.14)$$

$PE_r^e(\text{preliminary})$ —preliminary volume of property, plant, and equipment in economic activity e in region r , PE^e —total volume of property, plant, and equipment in economic activity e , P_r^e —share of region r in production volume of economic activity e .

Iteration 1 is calculating variation of preliminary volume of property, plant, and equipment in regional scale:

$$v_r(i1) = \frac{PE_r}{\sum_{e=1}^{15} PE_r^e(\text{preliminary})} \quad (18.15)$$

PE_r —baseline volume of property, plant, and equipment in region r ; $PE_r^e(\text{preliminary})$ —preliminary volume of property, plant, and equipment in economic activity e in region r .

Table 18.3 Initial data and results of IPF algorithm for calculating property, plant, and equipment in different regions and economic activities

Parameter	PE_r , billion RUR	$\sum_{e=1}^{15} PE_r^e(i4)$ billion RUR	$PE_r^e(i4)$, billion RUR					
			Agriculture and forestry	Fishery	Mining	Manufacturing	...	Personal service
PE_1 (Belgorod)	651.4	651.4	93.1	0.0	66.6	88.4	...	11.6
PE_2 (Bryansk)	288.7	288.7	27.0	0.0	0.2	34.2	...	5.0
PE_3 (Vladimir)	376.7	376.7	13.8	0.0	1.2	72.9	...	12.5
PE_4 (Voronezh)	673.9	673.9	57.0	0.0	2.6	57.6	...	10.6
...
PE_{82} (Chukotka)	54.8	54.8	0.6	0.0	20.3	0.1	...	2.0
PE^e	74,662.4	-	2209.9	61.0	6950.9	7199.4	...	1567.6
$\sum_{r=1}^{82} PE_r^e(i4)$	-	74,662.4	2221.0	62.4	7174.1	7173.9	...	1560.0
$v^e(i4)$, %	-	0	0.5	2.2	3.1	0.4	...	0.5

Maximal variation $v_r(i1)$ is 127% in Chechen Republic. Since variation is very high, the volume of property, plant, and equipment in economic activity e in region r using $v_r(i1)$ is calculated and used as correcting coefficient:

$$PE_r^e(i1) = PE_r^e(\text{preliminary}) * v_r(i1) \quad (18.16)$$

$PE_r^e(i1)$ —volume of property, plant, and equipment in economic activity e in region r at the first iteration.

Iteration 2 Calculation of variation of volume of property, plant, and equipment in the scale of economic activities:

$$v^e(i2) = \frac{PE^e}{\sum_{r=1}^{82} PE_r^e(i1)} \quad (18.17)$$

PE^e —baseline volume of property, plant, and equipment in economic activity e ; $PE_r^e(i1)$ —volume of property, plant, and equipment in economic activity e in region r at the first iteration.

Maximal variation $v^e(i2)$ is 19.1% in economic activity ‘fishery,’ so the data needs further processing:

$$PE_r^e(i2) = PE_r^e(i1) * v^e(i2) \quad (18.18)$$

$PE_r^e(i2)$ —volume of property, plant, and equipment in economic activity e in region r at the first iteration.

Repeating iterations, on the fourth the deviation in regional scale goes down to zero and maximum 3.1% in economic activity scale (Table 18.3).

Further information about property, plant, and equipment of organizations is detailed to the sector scale:

$$PE_r^s = PE_r^e(i4) * d_r^{s-e} \quad (18.19)$$

PE_r^s —property, plant, and equipment of an organization, representing sector s in region r ; $PE_r^e(i4)$ —volume of property, plant, and equipment in economic activity e in region r at the fourth iteration.

18.4.5 Employment

IPF is also applied for aggregating information about employment, which is presented separately in regional scale and in scale of economic activities. Starting from maximum variation 232% in regional scale and 22% in economic activity scale, on the fourth iteration a zero and 2.5% variation accordingly is reached.

Number of employees of organizations in different sectors is also calculated:

$$L_r^s = L_r^e(i4) * d_r^{s-e} \quad (18.20)$$

L_r^s —number employees of an organization, representing sector s in region r ;
 $L_r^e(i4)$ —number employees in economic activity e in region r at the fourth iteration.

18.5 Further Data Processing

The aim of the first stage of our research methodology is to prepare and process initial data into detailed and interconnected objects of the database. The population and organizations at the base year of modeling are reconstructed by filling the model database step by step. Structure of the model database is presented in [19].

Algorithm of artificial society reconstruction is considered in [20]; it includes the following steps:

1. Set geographical structure of the Russian Federation (82 regions).
2. Create the original generation of agents in accordance with the sex-age structure of population and structure of households in each region (on the basis of All Russian Population Census data).
3. Generation of aggregated organizations in the regions. After that the property, plant, and equipment, financial state (credits and deposits) and accounting of each organization are initialized, as well as its supplies. Each organization has employees, which are assigned to it through workplaces with different qualification and salaries
4. Generation of educational places and assigning agents of the appropriate age to them. Number of educational places in higher and secondary vocational education is presented in the reports of the Ministry of Education and Science [5].

Procedure of synthetic population, organizations, and institutions generation was programmed on C# in Microsoft Visual Studio 2015. Generated objects are stored in the model database for later access via SQL-queries. Modeling results are presented in [20].

18.6 Results and Discussion

In this paper, methods of data preprocessing for the agent-based model of spatial development of the Russian Federation have been presented. At the first stage of the research, the population, production, and social institutions in the regions of Russia were reconstructed. This stage required collection of large amounts of information from various sources: All Russian Population Census, federal and regional statistical yearbooks, official information on the portals of the ministries and survey results.

While information on demographic situation and educational system in the regional context is quite fully represented in open official sources [1, 5], information on regional production and service requires processing. In official collections, information on employment and equipment is presented separately in the regional scale and in the scale of economic activities, while for simulating spatial development both of these aspects are crucial. The iterative proportional fitting algorithm was used to integrate such data arrays. Another problem was incompleteness of information in regional collections; for its solution, a method of sequential detailing was used, which maintained high accuracy for the regions that provided complete information.

The resulting arrays of initial data were converted using the generation algorithm [20] to the synthetic society that represents population of the Russian Federation in 2014, production, property, and equipment of organizations, employment, and educational system.

The task of loading real data is a crucial part of creating the model of the Russian Federation spatial development, since it gives the opportunity to evaluate impact of federal and regional policy on different categories of population, taking into account age, education, income and status in marriage, as well as their preferences and beliefs. The concept of integration large data arrays can be implemented in different countries which would seriously improve prognostic capabilities of the social simulation models.

Acknowledgements The reported study was funded by the Russian Foundation for Basic Research within the research project № 18-29-03049.

References

1. All-russian population census 2010 official website. <http://www.gks.ru>. Last accessed 22 Mar 2019
2. Russian Federation Federal State Statistics Service. <http://www.gks.ru>. Last accessed 15 Mar 2019
3. The Central Bank of the Russian Federation official site. <http://www.cbr.ru/eng/>. Last accessed 22 Mar 2019
4. Ministry of Economic Development of the Russian Federation official website. <http://economy.gov.ru/minec/main>. Last accessed 22 Mar 2019
5. Ministry of science and higher education Russian Federation official website. <https://minobrnauki.gov.ru/>. Last accessed 21 Feb 2019
6. Internet recruitment portal Head Hunter, <https://hh.ru/article/research>, last accessed 2019/02/06
7. National credit bureau official website. <https://www.nbki.ru/company/news/>. Last accessed 16 Mar 2019
8. Lebaron, B., Tesfatsion, L.: Modeling macroeconomies as open-ended dynamic systems of interacting agents. *Am. Econ. Rev.* **98**, 246–250 (April, 2008). <https://doi.org/10.1257/aer.98.2.246>
9. Fagiolo, G., Roventini, A.: Macroeconomic policy in DSGE and agent-based models redux: New developments and challenges ahead. *J. Artif. Soc. Social Simul.* **20**(1), 1 (2017). <https://doi.org/10.18564/jasss.3280>. <http://jasss.soc.surrey.ac.uk/20/1/1.html>

10. Napoletano, M., Roventini, A., Gaffard, J.: Time-varying fiscal multipliers in an agent-based model with credit rationing. *Economics Discussion Papers* 112, Kiel (2017). <http://hdl.handle.net/10419/172323>
11. Gatti, D., Desiderio, S.J.: Monetary policy experiments in an agent-based model with financial frictions. *J. Econ. Interact. Coord.* **10**(2), 265–286 (2015). <https://doi.org/10.1007/s11403-014-0123-7>
12. Ashraf, Q., Gershman, B., Howitt, P.: Banks, market organization, and macroeconomic performance: an agent-based computational analysis. In: Working Paper 17102, National Bureau of Economic Research (June 2011). <https://doi.org/10.3386/w17102>. <http://www.nber.org/papers/w17102>
13. Napoletano, M., Dosi, G., Fagiolo, G., Roventini, A.: Wage formation, investment behavior and growth regimes: an agent-based analysis. *Revue de l'OFCE* **124**(5), 235–261 (2012). <https://doi.org/10.3917/reof.124.0235>
14. Ballas, D., Kingston, R., Stillwell, J.: Using a spatial microsimulation decision support system for policy scenario analysis. In: *Recent Advances in Design and Decision Support Systems in Architecture and Urban Planning*. Springer, Dordrecht (2004)
15. Sushko, E.: Multi-agent model of the region: concept, design and implementation. Tech. Rep. Preprint WP/2012/292, CEMI RAS, in Russian (2012)
16. Cleave, N., Brown, P.J., Payne, C.D.: Evaluation of methods for ecological inference. *J. R. Stat. Soc. Series A (Stat. Soc.)* **158**(1), 55–72 (1995). <https://doi.org/10.2307/2983403>. <https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2983403>
17. Lahr, M.L., de Mesnard, L.: Biproportional techniques in input-output analysis: table updating and structural analysis. *Econ. Sys. Res.* **16**(2), 115–134 (2004). <https://doi.org/10.1080/0953531042000219259>
18. Lovelace, R., Birkin, M., Ballas, D., van Leeuwen, E.: Evaluating the performance of iterative proportional fitting for spatial microsimulation: New tests for an established technique. *J. Artif. Soc. Social Simul.* **18**(2), 21 (2015). <https://doi.org/10.18564/jasss.2768>. <http://jasss.soc.surrey.ac.uk/18/2/21.html>
19. Mashkova, A.L., Savina, O.A., Banchuk, Y.A., Mashkov, E.A.: Using open data for information support of simulation model of the russian federation spatial development. In: Chugunov, A., Misnikov, Y., Roshchin, E., Trutnev, D. (eds.) *Electronic Governance and Open Society: Challenges in Eurasia*, pp. 401–414. Springer International Publishing, Cham (2019)
20. Mashkova, A.L.: Reconstructing an artificial society on the basis of big open data. In: *Supplementary Proceedings of the 7th International Conference on Analysis of Images, Social Networks and Texts (AIST-SUP 2018)*, vol. 2268, pp. 241–251. Moscow, Russia, 5–7 July 2018

Chapter 19

Design and Implementation of Total Station Wireless Data Transmission System



Xiaohu Yin and Mihuan Wang

Abstract At present, the data transmission of total station is transmitted by wired mode, which requires the laying of transmission line hardware and has a high cost. In this paper, the total station wireless data transmission system is designed by using TMS320F28335 DSP and nRF905 wireless transceiver module. The system consists of wireless module sending data node and receiving node, displaying the data collected by the total station on the upper computer and applying it. The hardware design of the system includes the interface circuit between total station and DSP, the hardware circuit between wireless module and DSP, and the serial communication between DSP and host computer, and the software part includes nRF905 software, DSP control program software, and CSI serial communication software. The experimental results of wireless transmission waveform and PC serial port debugging show that the design can achieve wireless transmission of total station data, and the results are accurate.

19.1 System Design

The total station wireless data transmission system is a wireless data transmission device based on radio frequency technology. It is mainly divided into four parts: total station, DSP controller, nRF905 wireless transceiver module, and host computer. The system designed the connection between the DSP and the total station interface RS232. The data measured by the total station is loaded into the nRF905 wireless transmitting module through the interface under the control of the DSP for GFSK modulation [1], and the data was then sent to the nRF905 wireless receiving module. Next, it is transmitted by the DSP controller of the receiving end to the upper computer through the interface for processing. The receiving end receives the data transmitted by each transmitting module according to the actual needs in an address-sharing way [2]. The diagram of the system's structural frame is shown in Fig. 19.1.

X. Yin · M. Wang (✉)
Xi'an University of Science and Technology, Xi'an, Shanxi, China
e-mail: 375616915@qq.com

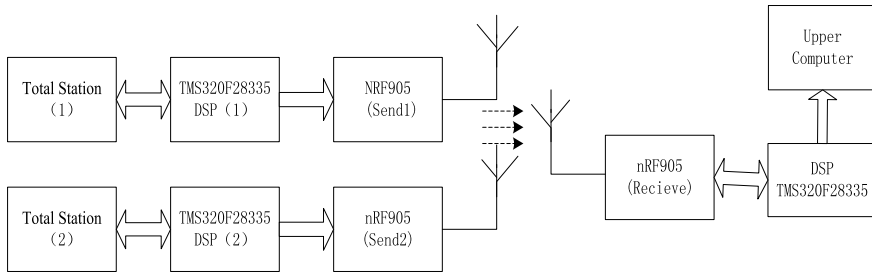


Fig. 19.1 Diagram of the system's structural frame

19.2 Hardware Design of the System

The hardware of the total station wireless data acquisition system mainly includes: each part of the power circuit, the interface connection circuit, and the nRF905 transceiver module circuit. For each part of the power circuit, it can be determined as the total station end and the upper machine end according to different situations. This design uses the Nikon series of total stations. The total station power supply adopts 18650 lithium battery pack, which has the advantages of small size, lightweight, high efficiency and energy-saving, multiple cycles, and many other advantages. The working time can reach 6 h, and the battery voltage range is 6–8.4 after combination with 2 strings and 2 combinations. V. The chip AMS1117-5V is used to convert the 5 V voltage required by the DSP, and the AMS1117-3.3 is used to convert the 3.3 V voltage required for the nRF905 operating mode. The same applies to the power circuit of the upper computer, which is not described here.

19.2.1 Interface Circuit Design

The interface for the Nikon series of total stations is a circular serial physical connector that uses the very broad RS-232C interface standard. RS-232C is a long-established standard (RS stands for recommendation; 232 stands for identifier; C stands for number of modifications), which describes the physical interface and protocol for lower-rate serial data communication between computers and related devices [3]. The names and functions of the pins of the Nikon Total Station RS-232C interface are shown in Table 19.1 [4].

For control data transmission, the selected DSP model is TMS320F28335 [5]. This type of digital signal processor is a TMS320C28X series floating-point DSP controller from TI. Compared with the previous fixed-point DSP, the device has high precision, low cost, low power consumption, high performance, high peripheral integration, large data and program storage, and more accurate and faster A/D

Table 19.1 Name and function of each pin of the RS-232C interface

Pin position	Signal	Input/output	Description
1	Public	NULL	Applied
2	NULL		
3	SD (TXD) data reception	1	Applied
4	RD (RXD) data reception	1	Applied
5	NULL		
6	NULL		

Table 19.2 Name and function of each pin of the RS-232C interface

Pin	Definition	Function
1	DCD	Carrier detection
2	RXD	Receive data
3	TXD	Send data
4	DTR	Ready for data terminal
5	SG	Signal position
6	DSR	Ready for data
7	RTS	Request to send
8	CTS	Clear sending
9	RI	Ringing prompt

conversion. The TMS320C28335 DSP also uses the RS-323C interface. Unlike the total station, it uses a 9-pin specification [6], as shown in Table 19.2.

It can be seen from Tables 19.1 and 19.2 that to achieve the goal of connecting the total station to the DSP, only the data line connection needs to be selected. The RS-232C interface connection between the total station and the DSP is shown in Table 19.3. At the same time, the selection of the data line is about 10 cm. The principle of the connection is that the total station is used as the output terminal and the DSP is used as the input pin connection [7]. It is also an essential line.

Table 19.3 Total station circular interface and DSP 9-pin interface

Serial port	Corresponding pin		
	Total station	1	2
DSP	2	3	5

Table 19.4 Name and function of each pin of the RS-232C interface

PWR_UP	TRX_CE	TX_EN	Working mode	
0	X	X	Power down and SPI programming	Saving mode
1	0	X	Standby and SPI programming	Saving mode
1	1	0	ShockBurst RX	Normal mode
1	1	1	ShockBurst TX	Normal mode

19.2.2 nRF905 Transceiver Module Circuit

The nRF905 used in this system is a single-chip RF transmitter chip from Nordic, Norway [8]. The 433 MHz band is freely used in China. The nRF905 consists of a frequency synthesizer, a receiver demodulator, a power amplifier, a crystal oscillator, and a modulator. The nRF905 can communicate with any MCU using the SPI interface, where the address, output power, and communication channel can be configured through the program, so it can be used for multi-machine communication [9]. The nRF905 incorporates ShockBurst™ technology, which automatically processes the packet headers and has built-in CRC check to ensure reliable data transfer [10]. The nRF905 consumes very low power. When transmitting at -10 dBm, the operating current is only 11 mA. The corresponding receiver's operating current is only 12.5 mA, which is widely used. The working mode of the nRF905 is shown in Table 19.4.

As the core control center, DSP has the control function for the nRF905 transceiver module [11]. Because the data transmission of TMS320F28335 and nRF905 is mainly applied to the MISO and MOSI pins in this circuit connection, the clock signal is output by the DSP, and other parts only need to be high. The low level is used. The power supply required by the nRF905 is 3.3 V, and the power supply of the TMS320F28335 is 5 V. It has been solved in the power supply circuit before and is not described here. The two ground lines can be directly connected, and the specific connection circuit is shown in Fig. 19.2.

19.2.3 Block Diagram of Interface Circuit Between DSP and PC

Nowadays, most PC computers do not have RS232 9 pin serial port, so USB is needed to realize communication. The USB specification describes bus characteristics, protocol definitions, programming interfaces, and other features required for system design and construction. USB is a master-slave bus. When working, the USB host is in the master mode and the device is in the slave mode [12]. The only system resources needed by the USB system are the memory space used by the software of the USB

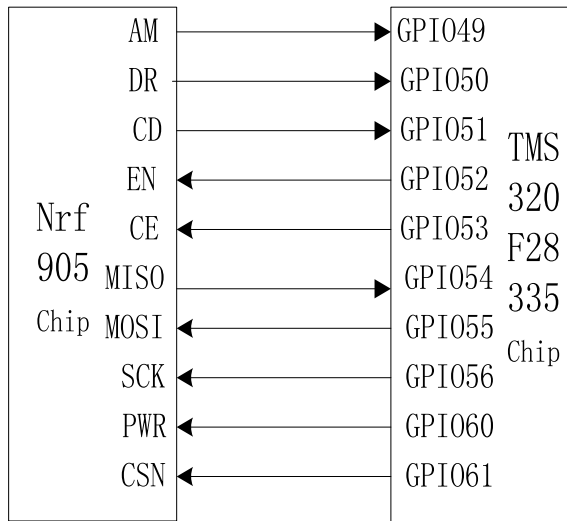


Fig. 19.2 TMS320F28335 DSP and nRF905 circuit connection

system, the memory address space (I/O address space) used by the USB master controller, and the interrupt request (IRQ) line. USB data transmission requires device drivers to identify devices, etc. It also needs a device driver to virtualize the serial port and imitate the real serial port. The driver of this system is USB POS provided by Microsoft, which allows applications to access USB devices. At the same time, the 9-pin serial port of the DSP has been explained above, and it will not be further elaborated. From RS232 to the host computer, USB needs to use chip conversion; the chip model used here is PL2303, produced in Taiwan, with good performance. The specific circuit block diagram is shown in Fig. 19.3.

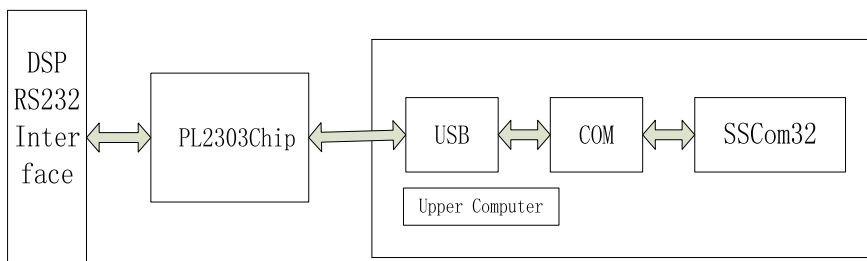


Fig. 19.3 Interface circuit block diagram

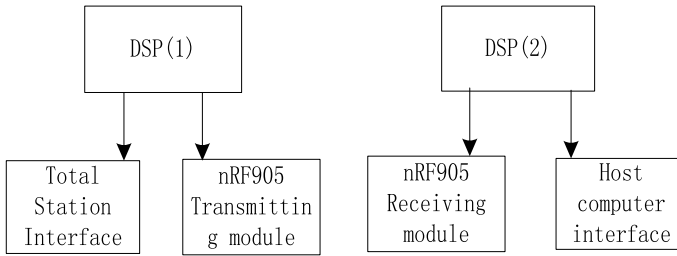


Fig. 19.4 DSP program control block diagram

19.3 Software Design of the System

19.3.1 Software Programming

The communication protocol of TMS320F28335 and nRF905 in this system adopts SPI protocol. Each interface communication in the system requires software implementation to match the hardware. Specifically, the program software of the interface between the controller DSP and the total station, the program software of the nRF905 transceiver module, the program software of the controller DSP and the host computer interface, the above programs need to be completed in the DSP controller, and the specific block diagram is shown in Fig. 19.4.

19.3.2 Software Program Implementation

19.3.2.1 Receiving and Transmitting nRF905

In the process of designing and implementing the system, the data transmission is divided into one master node and several slave nodes. In the actual use of the total station, the total station data of the different slave nodes is transmitted to the master node PC. The data in the master node is received by the nRF905 to the DSP, and the slave node is the same [13]. When the DSP controls the nRF905 wireless transmission program design, the software we use here is CCS3.3. The specific SPI configuration procedure is as follows:

```

void spi_init()
{SpiaRegs.SPICCR.bit.SPISWRESET = 0; // reset SPI
SpiaRegs.SPICCR.bit.CLKPOLARITY = 0;
SpiaRegs.SPICTL.bit.CLK_PHASE = 1; // Input data latched on rising edge
SpiaRegs.SPICCR.bit.SPICHR = 0x07; // One-character move-in is removed
to 8 bits
SpiaRegs.SPICTL.bit.MASTER_SLAVE = 1; // DSP is set to host mode
  
```

```

SpiaRegs.SPICTL.bit.TALK = 1;
SpiaRegs.SPIBRR = 0x0063;
SpiaRegs.SPICCR.bit.SPISWRESET = 1; // SPI is ready to work
SpiaRegs.SPIPRI.bit.FREE = 1; // automatically running }

```

The control of the nRF905 chip by TMS320F28335 mainly includes the initial configuration, working mode, transmitting data, and receiving stored data of nRF905. During initialization, configure the center frequency, operating band, output power, retransmission data enable, RX and TX address width, RX and TX data valid width, output clock frequency, crystal oscillator frequency setting, in the RF configuration register of the nRF905. The CRC check allows a series of configuration parameters such as enable bit and CRC mode. The specific procedures are as follows:

```

Unsigned char Rfconfig[11] = {
0x00, // configuration command
0x4c, // CH-NO, configure the frequency band at 430 MHZ
0x0c, // output power is 10 db, no retransmission, the node is in normal mode
0x44, // accept the address width setting, 4 bytes
0x20, 0x20, // Receive valid data length is 32 bytes
0xCC, 0xCC, 0xCC, 0xCC, // receive address
0x58}; // CRC fill, 8-bit CRC check, external clock signal is not enabled, 16M
crystal

```

In the master node, the reception of data is mainly for the protocol operation of the nRF905 radio. The specific operation is as follows: When the nRF905 enable bit (TRX_CE) is set to high level and the mode select bit (TX_EN) is low level in the CCS3.3 software, the ShockBurst™ receive mode is entered at this time; after 650 μs, the nRF905 detects this time. After the 340 MHz band, the carrier pin is set high; for the matched address, the address match bit AM is set high; after the data packet is received, the header, address, and CRC check bit are automatically removed, and the data is received. Bit DR is high; the enable bit (TRX_CE) is low and enters the idle mode; nRF905 transmits the received data to the DSP through the SPI interface, and the received data pin (DR) bit and address match are transmitted after the transfer is completed. Pin AM is set low. The software part of the program is:

```

Void RxPacket(unsigned char *RXBUF) // receive packet
{
Int i = 0;
TRX_DIS // register read and write mode
NRF905_CSN_L; // chip select, active low
spi_write_BYTE(RRP); // receive command word
For(i = 0; i < 4; i++)
{*RXBUF++ = spi_read_BYTE();}
DELAY_US(2);
NRF905_CSN_H; // chip select pull high
TRX_EN // register read and write mode}

```

In the slave node, mainly the DSP controls the nRF905 to transmit data. The specific operation is as follows: After the DSP controls the total station data transmission, the address and data of the master node receiver are transmitted to the nRF905 through the SPI interface; the DSP set enable bit (TRX_CE) and the mode control bit (TX_EN) bit are high. The nRF905 enters the ShockBurstTM transmission mode; in the transmission mode, the RF register is automatically turned on, the prefix and the CRC check code are automatically added, and then the data packet is transmitted. When the transmission is completed, the data bit DR is set to a high level. AUTO_RETRAN is asserted high, and data is continuously retransmitted. After the enable bit (TRX_CE) is asserted low, the transmit process is complete and the idle mode is entered. Since there are many similarities between the main program and the receiving end on the transmitting end, they are not described here.

19.3.2.2 Data Transmission Between DSP and PC

Serial data communication between DSP and PC is realized by transferring total station data to PC's serial port software through serial port. SSCOM32 serial port software is used in this system. SCI serial communication is used in DSP [14]. In order to transmit data correctly, it is necessary to keep the transmission parameters of DSP and SSCOM32 consistent; that is, the baud rate is 9600, the invalid test bit, the 8-bit data bit and the 1 stop bit. The SCI setup code is as follows:

```
void scic_echoback_init()
{
    // in the InitSysCtrl() function
    ScicRegs.SCICCR.all = 0x0007; // 1 stop bit, No loopback
                                // No parity, 8 char bits,
                                // async mode, idle-line protocol
    ScicRegs.SCICTL1.all = 0x0003; // enable TX, RX, internal SCICLK,
                                // Disable RX ERR, SLEEP, TXWAKE
    ScicRegs.SCICTL2.all = 0x0003;
    ScicRegs.SCICTL2.bit.TXINTENA = 1;
    ScicRegs.SCICTL2.bit.RXBKINTENA = 1;
    #if (CPU_FRQ_150 MHZ)
        ScicRegs.SCIHBAUD = 0x0001; // 9600 baud @LSPCLK = 37.5 MHz.
        ScicRegs.SCILBAUD = 0x00E7;
    #endif
    ScicRegs.SCICTL1.all = 0x0023; // Relinquish SCI from Reset
}
```

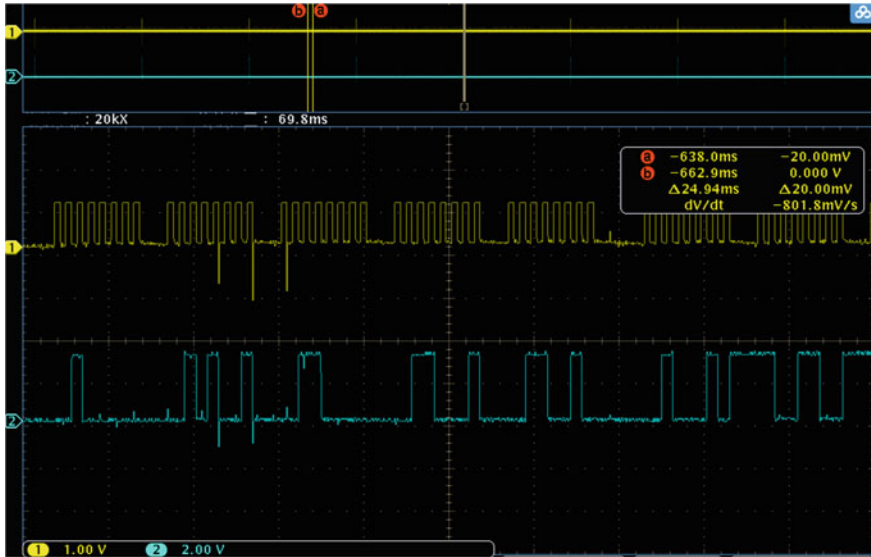


Fig. 19.5 Transmitter data waveform

19.4 System Testing and Results

19.4.1 Wireless Transmission Data Testing

In this system, the data collected by the total station (horizontal angle, horizontal distance, vertical angle, slant range, etc.) is tested during the DSP control nRF905 wireless data transmission process, and the bugs in the program are modified in time according to these test results. Continuous optimization data can be transmitted smoothly. The waveform of the data transmitted at the transmitting end of nRF905 is as shown in Fig. 19.5.

Similarly, the waveform of receiving data at the receiving end is shown in Fig. 19.6.

In Figs. 19.5 and 19.6, the channels 1 and 2 of the four-channel oscilloscope are the clock and transmission data of the transmitting end, respectively, and the channels 3 and 4 are the clock and receiving data of the receiving end, respectively, and the data transmission is consistent by the 2 and 4 channels.

19.4.2 Display Result of PC Serial Port Software

In the actual test, the data of the total station is displayed on the upper computer. In this system, the baud rate, stop bit, check bit, and data bit are set by using ssc32 serial port software. The data obtained is shown in Fig. 19.7.

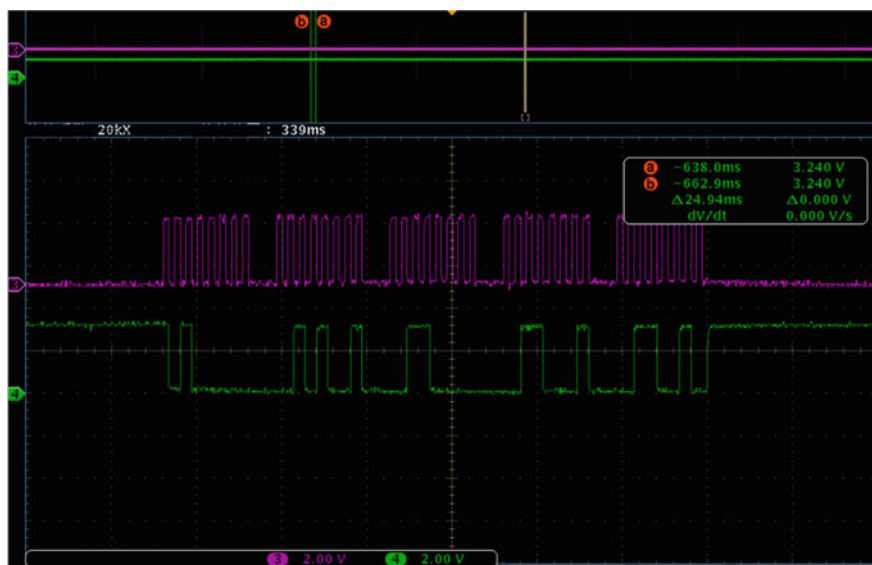


Fig. 19.6 Receiver data waveform

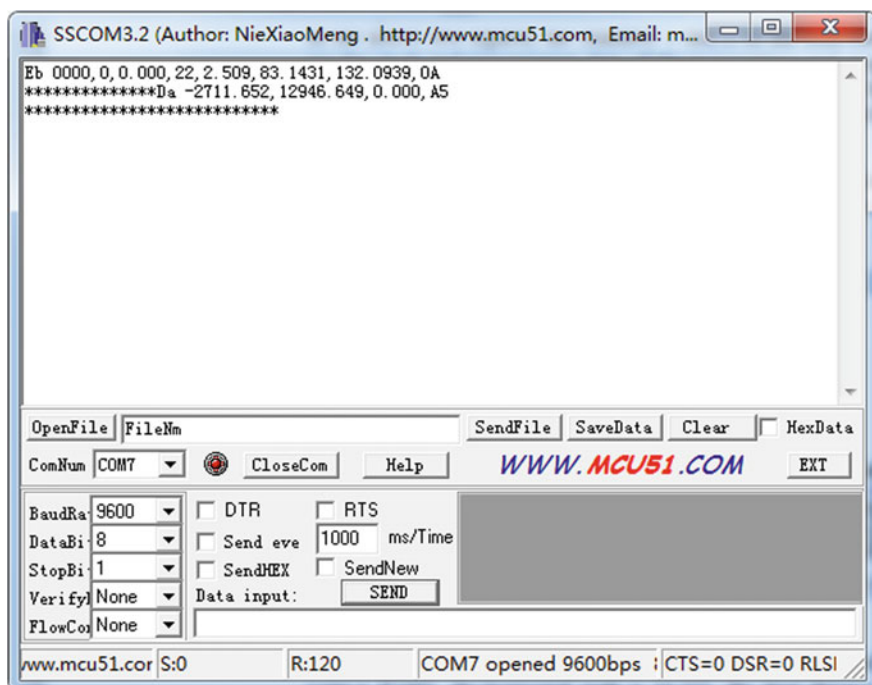


Fig. 19.7 Serial port data

According to the data sheet, it can be seen from Eb that the vertical angle of the target point is 83.14° , the horizontal angle is 132.09° , the slant distance is 2.51 m, and the PPM is 22.2. As Da knows, the N coordinate of the measuring point is -2711.652 , the E coordinate of the measuring point is 19,246.649, and the Z coordinate of the measuring point is -0.0 (the coordinate of the measuring point is set manually). As can be seen from Fig. 19.7, in the system, the command word DaCRLF and EbCRLF are sent to obtain the required data. After viewing the specific meaning and comparing with the actual total station data, it can be concluded that the data transmission is accurate during the data transmission process. The transmission error rate is 0.

19.5 Conclusion

In this paper, the various parts of the total station wireless data transmission system are designed. NRF905 wireless data module and DSP controller are used for circuit design and software design. After debugging, the data to be transmitted is obtained in the serial port software, and explained and utilized, which shows that the wireless information transmission of the total station can be realized, and the bit error rate is 0, which satisfies the design requirements. Because of its small size, the system can be directly applied to most of the total station. In the future construction of field roads, bridges, and houses, the workload of surveying workers is greatly reduced, the speed of the project is increased, and the overall contribution is tremendous.

References

1. Wang, S., Li, X.: Design of multi-channel wireless temperature acquisition system based on nRF905 and DS18B20. *Commun. Power Technol.* **28**(6), 55–57 (2011)
2. Dai, L., Liu, T.: Design of temperature acquisition system based on nRF905 wireless module. *Mod. Electron. Technol.* **38**(3), 21–23 (2015)
3. Liu, W., Xu, B.: Radio communication research between total station and PC-E500 computer. *Transp. Eng.* **6**(3), 106–108 (2012)
4. Hong, C., Zheng, Y.: RS232 serial communication application in PC and MCU communication. *South. Agric. Mach.* **2**(24), 175–176 (2019)
5. Liu, Q., Long, J.: Wireless data transmission and application between PDA and total station based on RF CC1101. *J. Nantong Ship. Vocat. Tech. Coll.* **4**(3), 52–55 (2011)
6. Li, Q.: Teach You to Learn DSP-Based on TMS320F28335. Beihang University Press, Beijing (2018)
7. Yi, Z., Liu, S.: Research on communication between total station, PDA and bluetooth. *Surv. Map. Bull.* **3**(7), 33–35 (2010)
8. Nordic VLSI ASA: Single Chip 433/869/915 MHz Transceiver nRF905 (1) (2017)
9. Zhao, Y.: Multi-point temperature wireless acquisition system based on MCU and nRF905. *Electron. Technol.* **2**(6), 99–101 (2017)
10. Yuan, J., Cao, J., Qiu, Z.: Wireless monitoring system for temperature and humidity of grain depots based on WSN. *Meas. Control Technol.* **31**(4), 77–81 (2012)

11. Wang, S., Li, Y.: Research on wireless WiFi transmission based on DSP image data. *Commun. World* **3**(18), 28–29 (2015)
12. Duan, Z., He, Q.: Design of communication interface between DSP and PC based on USB. *Inform. Comput.* **2**(24), 106–108 (2010)
13. Dong, X., Zhao, C.: Design and implementation of wireless remote control system based on AT89S52 and nRF905. *J. Zhongyuan Univ. Technol.* **2**(4), 27–29 (2016)
14. Xue, Y., Wu, Y.: Research on the development of serial communication interface based on DSP. *Aeronaut. Comput. Technol.* **1**(48), 95–97 (2018)

Chapter 20

The Comprehensive Evaluation of “Five Aspects” Based on Coefficient-of-Variation-Modified G1 Combination Weighting



Wei Ren and Hao Jian

Abstract In order to test the scientificity of evaluation criteria and evaluation indexes, this paper firstly uses principal component analysis method to analyze the evaluation indexes under the evaluation criteria. Then, this paper uses the ratio of the coefficient of variation of each evaluation index to replace the experts' subjective ratio of the importance degree, building a combination weighting method based on coefficient-of-variation-modified G1. Finally, with this method, this paper makes a comprehensive evaluation of development of “Five Aspects” in Guangdong Province.

20.1 Introduction

As is known, the key to scientific and successful comprehensive evaluation is to scientifically and rationally endow different weight for each evaluation index. At present, important methods for determining the weight of evaluation indicators are subjective weighting method, objective weighting method and combined weighting method. The subjective weighting method can better reflect the subjective intentions or experiences of decision makers or experts, but it cannot accurately reflect the objective real data information of the evaluation indicators. The objective weighting method is opposite. People wanted to find out a scientific evaluation method, which could not only take in the advantages of subjective and objective weighting methods in scientific decision-making but could also solve the problem that single method was incapable of reflecting experts' experience or the objective information of the indexes. In this situation, combination weighting method came into being.

The combination weighting used for comprehensive evaluation absorbs both the advantages of the subjective weighting and the objective weighting. Nevertheless, it overcomes both of their shortcomings. However, the combination weighting method is difficult to allocate the combination coefficient scientifically and reasonably. In order to overcome the shortcomings of combination weighting, many scholars have

W. Ren · H. Jian (✉)

School of Business Administration, Guangdong University of Finance and Economics, 510320
Guangzhou, China
e-mail: rajonhao@163.com

taken different approaches and used the objective weighting method to revise the subjective weighting method. Certainly, the effectiveness of such a revised combination weighting method has been proved. For example, Li [1] used the method of entropy-revised G1 combination weighting to evaluate the development of science and technology. In 2012, he put forward another combination weighting method based on the standard variance-revised group G1 [2]. Zhu and his colleagues [3] used the entropy-revised AHP combination weighting method to evaluate the development of “Five Aspects” in Jilin Province. Nevertheless, Zhu and his colleagues also put forward modified-G2 weighting method based on improved CRITIC [4] and coefficient of variation [5]. Besides, Jia and her colleagues bring about an entropy-modified G2 weighting method [6]. As for improved-G1 method, Xing and her cooperators come up with an improved CRITIC-G1 weighting method [7]. Zhu and his partners put forward a modified G1 method based on information gain ratio [8]. These studies have contributed to the development and application of comprehensive evaluation methods.

The *Opinions on Accelerating the Construction of Ecological Civilization* proposes to coordinate the promotion of new industrialization, informatization, urbanization, agricultural modernization and greening. In order to scientifically evaluate the objective status quo of regional “Five Aspects”, based on the comprehensive evaluation index system established by Zhu and his colleague [3], following the principles of scientificity, comparability, operability, comprehensiveness and the availability of data to select the indexes, we firstly construct a comprehensive evaluation index system which can fully reflect the development of the “Five Aspects”. Secondly, a coefficient-of-variation-modified G1 combination weighting method is proposed. Finally, our new combination weighting method is used to empirically analyze the development status of the “Five Aspects” in Guangdong Province.

20.2 Construction of the Index System

20.2.1 *Basis of the Construction of the Index System and Establishment of the Criterion Layer*

The *Opinions on Accelerating the Construction of Ecological Civilization* proposes to coordinate the promotion of new industrialization, informatization, urbanization, agricultural modernization and greening. Therefore, these “Five Aspects” are used as the criterion layer of the index system. Among them, “new industrialization, informatization, agricultural modernization” reflect the principle of innovation-driven, “urbanization” reflects the principle of people-oriented, and “greening” reflects the principle of green and low-carbon.

20.2.2 Construction of the Index System

Drawing on the existing researches, following the principles of scientificity, comparability, operability, comprehensiveness and the availability of data, we establish a comprehensive evaluation index system with five criteria and 25 indexes (Table 20.1). In order to know whether the information of the indexes of each criterion can fully symbolize the criterion, we firstly performed principal component analysis. If all indexes of one criterion could only produce one principal component, it indicated that the principal component could integrate all the information of each index well. In our study, principal component analysis was performed with SPSS 24.0, and selection of the principal component was based on the characteristic root which was more than or equal to 1. All indexes under each criterion layer could only produce one principal component, and each principal component could account for most of the variance variation (close to or beyond 80%), indicating that it is scientific and reasonable to use these indexes to reflect each criterion.

20.3 Coefficient-of-Variation-Modified G1 Combination Weighting Method

20.3.1 Introduction of the Traditional G1 Method

G1 [8–10] is a typical subjective weighting method. The weight of indexes depends on the subjective experience of experts or decision makers. When using the traditional G1 method, the weight of the index layer to the criterion layer is determined first, and then, the weight of the index layer to the goal layer is determined. While using G1 method, the orders of indexes are determined by the experts. Then, the ratio of the importance degree (r_k) of the adjacent index X_{k-1} and X_k is determined according to the determined order, and the weight of each index under the criterion layer is determined according to the value of r_k . The value of r_k is generally referred to Table 20.2.

It can be seen that while using the traditional G1 method, the ratio of the importance degree (r_k) of the adjacent index X_{k-1} and X_k is determined by the experts' experience, and it fails to effectively reflect the information utility contained in the objective data itself. Therefore, the coefficient-of-variation-modified G1 method is used to determine the ratio of the importance degree (r_k) of the adjacent index X_{k-1} and X_k , which can reflect the experts' experience as well as the magnitude of data information through r_k .

Table 20.1 Comprehensive evaluation index system for the development of “Five Aspects”

No.	Criterion layer	Index layer	Direction	Principal component interpretation variance (%)	2013	2014	2015	2016	2017
1	New industrialization X_1	Value-added of industry X_{11}	Pos.	89.51	26,894.54	30,079.24	31,290.75	32,650.89	35,291.83
2		Gross industrial output value X_{12}	Pos.		119,139.72	130,081.02	135,308.14	144,926.09	148,173.99
3		Proportion of the gross industrial output value in GDP X_{13}	Pos.		0.43	0.44	0.42	0.40	0.39
4		Value-added of high-tech manufacturing industry X_{14}	Pos.		6654.38	7083.66	7537.34	8475.25	9507.81
5		Proportion of value-added of high-tech manufacturing industry X_{15}	Pos.		0.25	0.24	0.24	0.26	0.27
6	Urbanization X_2	Urbanization rate X_{21}	Pos.	97.02	67.80	68.00	68.71	69.20	69.90
7		Per capita housing construction area X_{22}	Pos.		30.27	31.88	32.25	32.74	33.09

(continued)

Table 20.1 (continued)

No.	Criterion layer	Index layer	Direction	Principal component interpretation variance (%)	2013	2014	2015	2016	2017
8		Proportion of non-agricultural population X_{23}	Pos.		53.69	54.32	54.96	55.60	56.25
9		Per capita disposable income X_{24}	Pos.		29,537.29	32,148.11	34,757.16	37,684.25	40,975.14
10		Per capita consumption expenditure X_{25}	Pos.		21,621.46	23,611.74	25,673.08	28,613.33	30,197.91
11	Greening X_3	Industrial wastewater X_{31}	Neg.	92.04	17.05	17.76	16.15	13.20	12.38
12		Total volume of industrial waste gas emission X_{32}	Neg.		28,434.00	29,793.00	30,923.00	38,846.00	41,997.50
13		Living wastewater X_{33}	Neg.		69.13	72.68	74.93	80.60	83.75
14		Per capita urban public green area X_{34}	Pos.		15.94	16.28	17.40	17.87	18.24
15		Number of public transportation vehicles X_{35}	Pos.		65,844.00	61,685.00	62,947.00	68,965.00	73,888.00

(continued)

Table 20.1 (continued)

No.	Criterion layer	Index layer	Direction	Principal component interpretation variance (%)	2013	2014	2015	2016	2017
16	Informatization X ₄	Number of mobile telephones subscribers X ₄₁	Pos.	79.95	14,706.06	14,943.37	15,009.75	14,348.96	14,798.85
17		Broadband subscribers of internet X ₄₂	Pos.		2154.28	2243.87	2285.19	2850.60	3288.15
18		Popularization rate of mobile telephones X ₄₃	Pos.		138.16	139.35	138.35	130.46	132.48
19		Pieces of express mail service X ₄₄	Pos.		210,670.00	335,555.90	501,335.00	767,241.56	1013,468.00
20	Agricultural modernization X ₅	Annual average number of newspapers and magazines subscribed per 100 persons X ₄₅	Neg.		10.58	8.94	8.40	6.25	4.99
21		Total agricultural machinery power X ₅₁	Pos.	88.14	2545.30	2611.78	2696.80	2390.50	2410.77

(continued)

Table 20.1 (continued)

No.	Criterion layer	Index layer	Direction	Principal component interpretation variance (%)	2013	2014	2015	2016	2017
22		Total area cultivated using machinery X_{52}	Pos.		3488.03	3612.69	3741.81	3969.17	4014.05
23		Water-saving irrigated area X_{53}	Pos.		268.35	281.77	295.86	301.49	326.19
24		Total area sown using machinery X_{54}	Pos.		204.84	232.88	264.76	315.78	342.21
25		Value-added of agriculture X_{55}	Pos.		3047.51	3118.39	3275.05	3593.64	3712.71

Table 20.2 r_k Value

r_k Value	Notes
1.0	X_{k-1} is as important as X_k
1.2	X_{k-1} is slightly more important than X_k
1.4	X_{k-1} is obviously more important than X_k
1.6	X_{k-1} is especially more important than X_k
1.8	X_{k-1} is extremely more important than X_k

20.3.2 Coefficient-of-Variation-Modified G1 Method

20.3.2.1 Scoring the Evaluation Index

Suppose P_{ij} be the j th index score of the i th evaluation object, V_{ij} the raw data of the j th index of the i th evaluation object, and n means the number of objects to be evaluated. Here comes the scoring formula for positive indexes:

$$P_{ij} = \frac{V_{ij} - \min_{1 \leq i \leq n} \{V_{ij}\}}{\max_{1 \leq i \leq n} \{V_{ij}\} - \min_{1 \leq i \leq n} \{V_{ij}\}}$$

The economic meaning in the formula is the relative distance of the deviation between the j th index value and the minimum value of the i th evaluation object with respect to the maximum–minimum deviation, and higher the score indicates that the index is better. As for the negative indexes, we have another scoring formula:

$$P_{ij} = \frac{\max_{1 \leq i \leq n} \{V_{ij}\} - V_{ij}}{\max_{1 \leq i \leq n} \{V_{ij}\} - \min_{1 \leq i \leq n} \{V_{ij}\}}$$

20.3.2.2 Calculating the Coefficient of Variation

Suppose CV_k be the coefficient of variation of the k th evaluation index, then

$$CV_k = \frac{\delta_k}{\bar{x}_k}$$

δ_k represents the standard deviation of the k th evaluation index, and \bar{x}_k represents the mean value of the k th evaluation index. The larger the value of CV_k , the more information the evaluation index contains, indicating that the index is more important.

20.3.2.3 Combination Weighting

1. According to the value of CV_k , we determine the ratio of the importance degree (r_k) of the adjacent index X_{k-1} and X_k . The formula is

$$r_k = \begin{cases} \frac{CV_{k-1}}{CV_k}, & \text{while } CV_{k-1} \geq CV_k; \\ 1, & \text{while } CV_{k-1} < CV_k. \end{cases}$$

2. According to the value of r_k , we calculate the coefficient-of-variation-modified G1 combination weight s_m of the m th index under the criterion layer. The formula is

$$s_m = \left(1 + \sum_{k=2}^m \prod_{i=k}^m r_i \right)^{-1}$$

3. According to the value of s_m , we can calculate the weight of the other index. The formula is

$$s_{k-1} = r_k \cdot s_k, \quad k = m - 1, m - 2, \dots, 1, 2, 3$$

4. Suppose α_k be the weight of the k th index under the j th criterion layer to the total goal, s_k be the weight of the k th index to the j th criterion layer under the j th criterion layer, $s^{(j)}$ be the weight of the j th criterion layer to the total goal. Then, we have the weight of the index to the total target α_k :

$$\alpha_k = s_k \cdot s^{(j)}$$

20.3.2.4 Calculating the Score of the Evaluation Object

Suppose P_i be the score of the i th evaluation object. According to the weight and the score of the index, we have

$$P_i = \sum_{j=1}^n P_{ij} \cdot a_i$$

20.4 Empirical

20.4.1 Evaluation Objects and Data Sources

We take Guangdong Province as the research object and select the relevant indexes of Guangdong's "Five Aspects" during 2013–2017 as sample. The data needed are from the statistical yearbook of Guangdong Province. Among them, data of industrial wastewater discharge, industrial waste gas discharge and urban domestic sewage discharge are only from 2013 to 2016. According to their development trend, the average growth rate is used to supplement the data of 2017, same as the non-agricultural population. The method predicts the data for three years, from 2015 to 2017. There is data of agricultural machinery, mechanical farming area, water-saving irrigation area and mechanical planting area for 2015–2017, and the average increase is also based on its development trend. The data for 2013 and 2014 will be supplemented (Table 20.1).

20.4.2 Evaluation Index

The data of each index in 2013–2017 are standardized and scored with the evaluation index scoring formula in Sect. 20.3.2.1 (Table 20.3).

20.4.3 Calculating the Coefficient of Variation and the Combination Weight

According to the formula in Sect. 20.3.2.2, we calculate the coefficient of variation of each index. By summing up the coefficient of variation of each index under the same criterion, we calculate the coefficient of variation of each criterion (Table 20.3). According to the coefficient of variation of index and criterion, based on the formula in Sect. 20.3.2.3, we calculate the weight of the index, criterion. Finally, we calculate the comprehensive weight of the index to the total goal (Table 20.3).

20.4.4 Calculating the Score of Evaluation Object

We calculate the scores of each evaluation index according to the formula in Sect. 20.3.2.4. The summation sum is then used to calculate the scores for each evaluation criterion (Table 20.4). For the convenience of analysis, all scores of evaluation indexes time 100.

Table 20.3 Standardized scores and weights of indexes

No.	Index	2013	2014	2015	2016	2017	CV of index	r of index	Weight of index	CV of criterion	r of criterion	Weight of criterion	Comprehensive weight
1	X ₁₁	0.00	0.38	0.52	0.69	1.00	0.10	1.153	0.465	0.43	1.321	0.377	0.175
2	X ₁₂	0.00	0.38	0.56	0.89	1.00	0.09	1.800	0.403				0.152
3	X ₁₃	0.74	1.00	0.60	0.23	0.00	0.05	1.000	0.224				0.085
4	X ₁₄	0.00	0.15	0.31	0.64	1.00	0.15	2.642	0.224				0.085
5	X ₁₅	0.35	0.00	0.16	0.71	1.00	0.06	–	0.085				0.032
6	X ₂₁	0.00	0.10	0.43	0.67	1.00	0.01	1.000	0.245	0.33	1.000	0.285	0.070
7	X ₂₂	0.00	0.57	0.70	0.88	1.00	0.03	1.852	0.245				0.070
8	X ₂₃	0.00	0.25	0.49	0.75	1.00	0.02	1.000	0.132				0.038
9	X ₂₄	0.00	0.23	0.46	0.71	1.00	0.13	1.000	0.132				0.038
10	X ₂₅	0.00	0.23	0.47	0.82	1.00	0.14	–	0.132				0.038
11	X ₃₁	0.13	0.00	0.30	0.85	1.00	0.16	1.000	0.294	0.54	1.000	0.285	0.084
12	X ₃₂	1.00	0.90	0.82	0.23	0.00	0.18	2.283	0.294				0.084
13	X ₃₃	1.00	0.76	0.60	0.22	0.00	0.08	1.334	0.129				0.037
14	X ₃₄	0.00	0.15	0.63	0.84	1.00	0.06	1.000	0.096				0.028
15	X ₃₅	0.34	0.00	0.10	0.60	1.00	0.07	–	0.096				0.028
16	X ₄₁	0.54	0.90	1.00	0.00	0.68	0.02	1.000	0.468	1.10	2.275	0.285	0.134
17	X ₄₂	0.00	0.08	0.12	0.61	1.00	0.19	6.457	0.468				0.134
18	X ₄₃	0.87	1.00	0.89	0.00	0.23	0.03	1.000	0.072				0.021
19	X ₄₄	0.00	0.16	0.36	0.69	1.00	0.58	2.033	0.072				0.021
20	X ₄₅	0.00	0.29	0.39	0.77	1.00	0.28	–	0.036				0.010
21	X ₅₁	0.51	0.72	1.00	0.00	0.07	0.05	1.000	0.375	0.48	–	0.125	0.047

(continued)

Table 20.3 (continued)

No.	Index	2013	2014	2015	2016	2017	CV of index	r of index	Weight of index	CV of criterion	r of criterion	Weight of criterion	Comprehensive weight
22	X ₅₂	0.00	0.24	0.48	0.91	1.00	0.06	1.000	0.375				0.047
23	X ₅₃	0.00	0.23	0.48	0.57	1.00	0.07	1.000	0.375				0.047
24	X ₅₄	0.00	0.20	0.44	0.81	1.00	0.21	2.396	0.375				0.047
25	X ₅₅	0.00	0.11	0.34	0.82	1.00	0.09	–	0.156				0.020

Table 20.4 Evaluation scores of the development status of the “Five Aspects” in Guangdong

Year	New industrialization	Urbanization	Greening	Informatization	Agricultural modernization	Comprehensive evaluation
2013	7.37	0.00	14.10	9.02	2.37	32.86
2014	22.11	7.33	10.73	15.77	6.77	62.70
2015	25.86	13.32	13.60	17.89	11.92	82.59
2016	35.12	19.38	13.80	10.43	12.40	91.12
2017	44.41	25.33	13.88	26.02	16.37	126.01

20.4.5 Analysis of the Status Quo of the Development of “Five Aspects” in Guangdong Province

It can be seen from Table 20.4 that the development status of the “Five Aspects” in Guangdong Province is increasing year by year and maintaining a good momentum. The comprehensive score of 2017 “Five Aspects” is nearly four times that of 2013. In order to better present the development trend of the “Five Aspects” in each year, the year is taken as the abscissa, and the “Five Aspects” score is taken as the ordinate to draw the development trend of the “Five Aspects” in Guangdong Province (Fig. 20.1).

It can be seen from Fig. 20.1 that the comprehensive score of “Five Aspects” in Guangdong Province in the five years from 2013 to 2017 has increased year by year, showing the good momentum of the development of “Five Aspects” in Guangdong Province. In concrete terms, new industrialization, urbanization and agricultural modernization have shown a good momentum of development year by year.

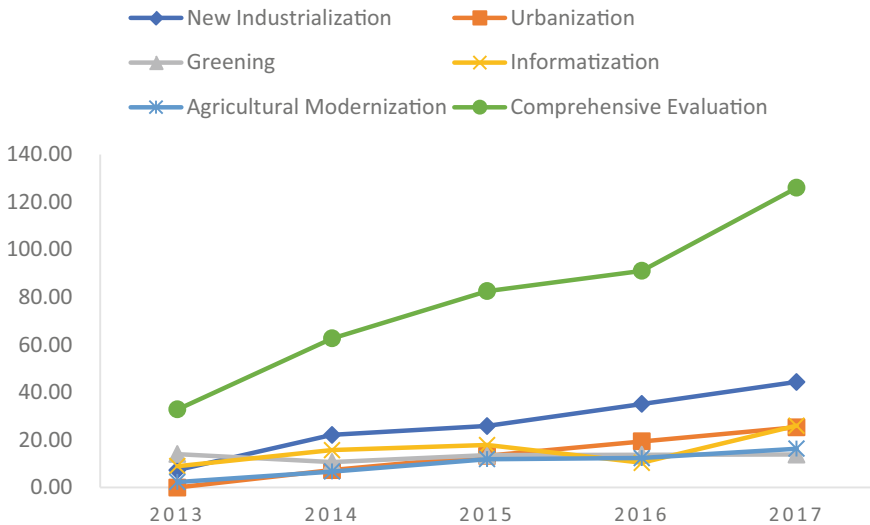


Fig. 20.1 Development trend of “Five Aspects” in Guangdong

The development of new industrialization in 2014–2017 has surpassed other “four aspects”, indicating that new industrialization provides a continuous stream of development for the “Five Aspects”. This is in line with Guangdong’s status of the country’s first major economic province.

The development of informatization has experienced ups and downs, but overall, it still shows growth momentum, which can be said that it develops in twists and turns. It should be noted that although the state of the development of greening during the period of 2014–2017 has improved, it still does not reach the level of 2013. During the period of 2013–2017, although the discharge of industrial wastewater in Guangdong Province has been declining year by year, the discharge of industrial waste gas and urban domestic sewage has increased year by year, resulting in a decline in the development of “greening” in Guangdong. Greening dragged the hind legs of the “Five Aspects”, indicating that the development of new industrialization, urbanization, informatization and agricultural modernization was at cost of the environment.

20.4.6 Comparison

In previous research, we analyzed the development of the “Five Aspects” based on coefficient of variation weighting method. The results show that during 2013–2017, new industrialization, urbanization and agricultural modernization have shown a good momentum of development year by year. Besides, even though the development of greening during the period of 2014–2017 has improved, it still does not reach the level of 2013 (seen from Fig. 20.2). These findings are consistent with what we found

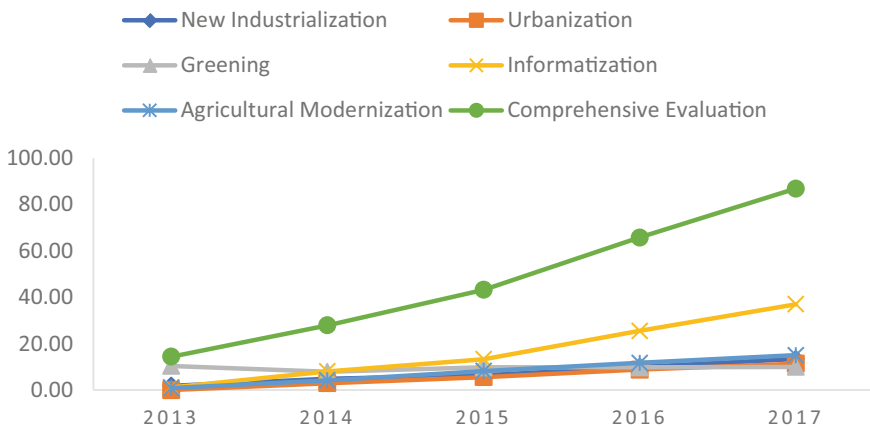


Fig. 20.2 Development trend of “Five Aspects” in Guangdong based on coefficient of variation weighting

based on the coefficient-of-variation-modified G1 weighting method. The difference exists in the informatization.

On the one hand, informatization is the first impetus for the development of the “Five Aspects” based on coefficient of variation weighting, while new industrialization takes the place based on the coefficient-of-variation-modified G1 combination weighting. On the other hand, informatization develops year by year based on the former method, but it develops in twists and turns based on the latter method.

Seen from the historical experience, industrialization promotes the development of urbanization and agricultural modernization, while it influences the development of greening. Nevertheless, new industrialization brings a great need of development of informatization. As a consequence, new industrialization should be the first impetus for the development of the “Five Aspects”. Therefore, experts’ experience should be taken into consideration while evaluating. Comparatively, the results based on the coefficient-of-variation-modified G1 combination weighting method are more scientific and reasonable.

20.5 Conclusion

In this paper, the coefficient-of-variation-modified G1 combination weighting method is constructed by using the coefficient of variation of the evaluation index. The actual importance degree of the index is determined by the coefficient of variation. Coefficient-of-variation-modified G1 combination weighting method can not only reflect the subjective intention of the experts or the decision makers, but it can also reflect the objective information of the index and solve the problem of reasonable distribution of weights. By using this method, this paper uses the development data of “Five Aspects” of Guangdong Province to conduct an empirical analysis.

Based on our analysis, Guangdong should continue to promote the development of new industrialization and maintain a strong growth momentum in the future development process of the “Five Aspects” while strengthening construction of greening and intensifying efforts to rectify industrial waste emissions on the basis of further reduction of industrial wastewater. Besides, try to raise people’s environmental awareness, encourage the recycling of domestic sewage, reduce urban domestic sewage discharge. What is more, increase the intensity of afforestation, promote the use of new energy. In a word, we want gold and silver mountains as well as green water and mountains.

References

1. Li, G.: The science and technology evaluation model and its empirical research based on entropy-revised G1 combination weighting. *Soft Sci.* **24**(05), 31–36 (2010)
2. Li, G.: Research on method of determining combination weights based on the standard deviation revised group-G1. *J. Syst. Eng.* **27**(01), 9–18 (2012)
3. Zhu, Z., Zhang, J., Zhang, G., et al.: Comprehensive evaluation model and its empirical research based on entropy-revised AHP combination weighting. *Stat. Decis.* **34**(13), 47–51 (2018)
4. Zhu, Z., Zhang, G., Zhang, J.: Modified-G2 weighting method based on improved CRITIC and its solid evidence. *Stat. Decis.* **34**(18), 33–38 (2018)
5. Zhu, Z., Zhang, Z.: Modified G2 weighting method and demonstration based on coefficient of variation. *Stat. Decis.* **35**(02), 70–74 (2019)
6. Jia, B., Zhao, T., Zhu, Z.: Comprehensive evaluation method based on entropy value correction G2 weighting and its empirical analysis. *Stat. Decis.* **35**(08), 30–35 (2019)
7. Xing, Y., Wang, J., Ma, W., et al.: China's "Wuhua" coordinated development measure based on improved CRITIC-G1. *Stat. Decis.* **35**(14), 42–46 (2019)
8. Zhu, Z., Zhou, L., Zhang, G.: Weighting method measurement model and demonstration of system coordinated development based on modified G1 method by information gain ratio. *Stat. Decis.* **35**(13), 24–28 (2019)
9. Chi, G., Qi, F., Li, G.: The evaluation model of scientific development concept for Chinese provinces based on combination weighting of improved group-G1 and its application. *Syst. Eng. Theory Pract.* **33**(06), 1448–1457 (2013)
10. Chi, G., Zhu, Z., Zhang, Y.: The science and technology evaluation model based on entropy and G1 and empirical research of China. *Stud. Sci. Sci.* **26**(06), 1210–1220 (2008)

Chapter 21

An Inertia Weight Variable Particle Swarm Optimization Algorithm with Mutation



Mengying Zhao, Yuqi Ni, Tao Chao, and Ke Fang

Abstract To improve the precision, accuracy, and speed of the particle swarm optimization (PSO) algorithm, the influence of the function concavity of the inertia weight reduction strategy on the performance of the algorithm is analyzed by heuristic and contrast experiments. The advantages and disadvantages of different mutation strategies are analyzed, and a PSO algorithm is proposed based on the summary of the experimental results. The algorithm introduces a new type of inertia weight change strategy and mutation mode, so that the particles have different mutation probability in different periods of the algorithm, and the inertia weight is linearly reduced in segments by increasing the absolute value of the slope. The experimental results show that the proposed algorithm has better performance than the original algorithm in most of the selected test functions. In other words, the improvements to the PSO algorithm can improve the precision, speed, and correct rate of PSO algorithm.

21.1 Introduction

Optimization problems are common in our daily life. After proper mathematical modeling, such problems can be described as extreme point finding problems. Generally, the objective function is more complicated in this kind of problems, so it is almost impossible to solve this problem conveniently by just using the analytical method. The optimization problem applies to the following mathematical description (21.1):

$$\begin{cases} \min y = F(x) = (f_1(x), f_2(x), \dots, f_m(x))^T \\ \text{s.t. } g_i(x) \leq 0, \quad i = 1, 2, \dots, q \\ h_j(x) = 0, \quad j = 1, 2, \dots, p \end{cases} \quad (21.1)$$

M. Zhao
Harbin Electrical Machinery Company Limited, Harbin, China

Y. Ni · T. Chao (✉) · K. Fang
Control and Simulation Center, Harbin Institute of Technology, Harbin, China
e-mail: chaotao2000@163.com

Each $f_m(x)$ equation represents an objective function, while $g_i(x) \leq 0$ represents a series of inequality constraints on the optimization space, and $h_j(x) = 0$ represents the equality limit that needs to be taken into account. Single-objective optimization can be considered as a special case when the m value is 1, which will be the topic of this paper. Meanwhile, there are many ways to deal with the constraints, which will not be the topic of this paper.

For the sake of solving these problems, Kennedy and Eberhart [1] proposed the basic particle swarm algorithm in 1995. The algorithm is derived from the observation of the clustering behavior of birds, and abstracts the members of the flock into flying particles without collision volume. The algorithm has the advantages of rapid convergence and simple parameter setting, but it also faces the inherent drawbacks of convergence too early and falling into local extreme points and limited convergence accuracy. Later generations proposed many revisions based on the basic particle swarm algorithm [2–5]. In terms of inertia weight alone, there are many change strategies such as linear decline and exponential decline.

In 1998, Shi and Eberhart introduced the concept of inertia weight for particle swarm optimization in his work [6] and pointed out that when the inertia weight is constant, the value is appropriate in the interval [0.9, 1.2], which is inertia. The iteration effect is best when the weight is 0.9. In the following year, Shi proposed a linear weight change strategy in another paper [7]. It is also pointed out that when inertia weight starting with a initial value close to 1 then linearly decreasing to 0.4, the algorithm will have the best optimization effect. Then adaptive inertia weights and exponential decline change strategies have also come out [8, 9]. These algorithms have different improvements and have different efficiency.

In the aspect of algorithm innovation, it is also proposed to mix particle algorithm and genetic algorithm [10] and introduce mutation algorithm to let the particle have a certain probability to speed or position initialization after state update.

In this paper, a class of inertia weight reduction strategy based on the segmentation line method is proposed, and the start and stop conditions of the mutation algorithm are improved. The following is mainly written in the following order: 1. Introduce several existing inertia weight setting methods and conventional mutation algorithms. 2. Through the method of heuristic and comparison, this paper proposes a folding line inertia weight reduction strategy and improved mutation algorithm with better optimization effect. 3. Through the optimization experiment on the selected test function, verify the performance improvement effect of each improved part on the algorithm. A particle swarm optimization algorithm based on the off-line inertia weight reduction strategy and the improved mutation algorithm is summarized.

21.2 Particle Swarm Optimization Algorithm and Its Improved Strategy

21.2.1 Basic Particle Swarm Optimization

The particle swarm optimization algorithm is a heuristic algorithm that starts from the random initial solution and updates the state through a specific update rule to search for the point with the best function value in the solution space. Denote the speed of the i th particle in the population at the $k + 1$ th iteration as v_{in}^{k+1} , and its position as x_{in}^{k+1} , then the update process of the particle state can be describe by using the following two Formulas (21.2) and (21.3):

$$v_{in}^{k+1} = w \cdot v_{in}^k + c_1 \cdot r_1 \cdot (p_{in}^k - x_{in}^k) + c_2 \cdot r_2 \cdot (p_{gn}^k - x_{in}^k) \quad (21.2)$$

$$x_{in}^{k+1} = x_{in}^k + v_{in}^k \quad (21.3)$$

In the formula, w represents the inertia weight, which determines the degree of influence of the velocity state in the previous iteration on the particle velocity in this iteration, and represents the inheritance of the velocity state. p_{in}^k represents the individual historical optimal position of the i th particle under the current iteration number, and p_{gn}^k represents the overall historical optimal position of the particle population. r_1 and r_2 are two random coefficients whose values are independent of each other between 0 and 1.

It can be seen that the velocity of the particle is mainly composed of three components—the influence from the previous velocity, the spatial distance from the optimal position of the individual history, and the spatial distance from the optimal position of the population history [11].

21.2.2 Variable Inertia Weight Reduction Strategy

The linear decline of inertia weight is a more common adopted change strategy nowadays, and its expression is as shown in Eq. (21.4):

$$w = w_s - (w_s - w_e) * \frac{i}{T} \quad (21.4)$$

where i is the number of the current iteration when the state is updated, and T is the total number of iterations to be made. Since the inertia weight is continuously decreasing during the optimization process, w_s is both the maximum inertia weight and the initial value of the inertia weight. Similarly, w_e means both the minimum inertia weight and the final value of the inertia weight change.

In some adaptive PSO algorithms, an “adaptive inertia weight” strategy is adopted. In this case, the inertia weight is dynamically adjusted in the iterative process as follows

$$\begin{cases} w_e - \frac{(f - f_{\min})}{(f_{\text{avg}} - f_{\min})} * (w_s - w_e), & f < f_{\text{avg}} \\ w_s, & f \geq f_{\text{avg}} \end{cases} \tag{21.5}$$

This formula is suitable for a particle swarm optimization algorithm that seeks the minimum value, where f_{\min} represents the current minimum function value of the particle population, and f_{avg} represents the average function value of the population. The meanings of w_s and w_e are the same as in the previous formula.

In addition, there is a type of exponential descent method, in which c_3 is the adjustment coefficient, and the expression is (21.6)

$$w = w_e * (w_s / w_e)^{1 / (1 + c_3 * i / T)} \tag{21.6}$$

Improved variable inertia weight reduction method Through experimental experience, it is found that if the value of the inertia weight is a continuous function with the period as the independent variable, the concavity and convexity of the function will affect the performance of the algorithm. The exponential decline method is a type of descent method with concave function characteristics. It is conceivable that the algorithm will have better performance when the inertia weight decreases with a certain concavity. In order to study and discuss the most suitable way to reduce the inertia weight, this paper uses the method of “Using straight line to replace curve” to divide the total iteration period of the algorithm into four time periods. The inertia weight decreases linearly according to different slopes in each time period. By adjusting the ratio between the slopes of each stage, it is possible to approximate the curves with different concave and convex features. When the absolute value ratio of the slope has a decreasing trend, the polyline function adopted by the scheme can be approximated as a concave function curve. Conversely, when the absolute value of the slope of the polyline has an increasing trend, the polyline function can be approximated as a convex function curve. The concave function strategy and the convex function strategy are used to refer to the two types of schemes. Both types of schemes will be referred to using a concave function strategy and a convex function strategy. Denote the slope ratio of the four periods as $a : b : c : d$ and let $k = a + b + c + d$, then we have the following inertia weight expressions (21.7):

$$w = \begin{cases} w_s - (w_s - w_e) * \frac{a * i}{T * k / 4}, & i \leq \frac{T}{4} \\ w_s - (w_s - w_e) * \frac{b * (i - T/4) + a}{T * k / 4}, & \frac{T}{4} < i \leq \frac{T}{2} \\ w_s - (w_s - w_e) * \frac{c * (i - 0.5T) + a + b}{T * k / 4}, & \frac{T}{2} < i \leq \frac{3T}{4} \\ w_s - (w_s - w_e) * \frac{d * (i - 0.75T) + a + b + c}{T * k / 4}, & \frac{3T}{4} < i \leq T \end{cases} \tag{21.7}$$

This method is called the piecewise linear descent method.

21.2.3 Introduction and Improvement of Mutation Algorithm

The mutation algorithm is an algorithm widely used in intelligent algorithms. It is attached to the regular status update. Because it is as unpredictable as the mutations produced in the genome, it is also called the mutation algorithm. The mutation algorithm can bring more randomness to the optimization process, so it often makes the population jump out of the local optimum [11–13].

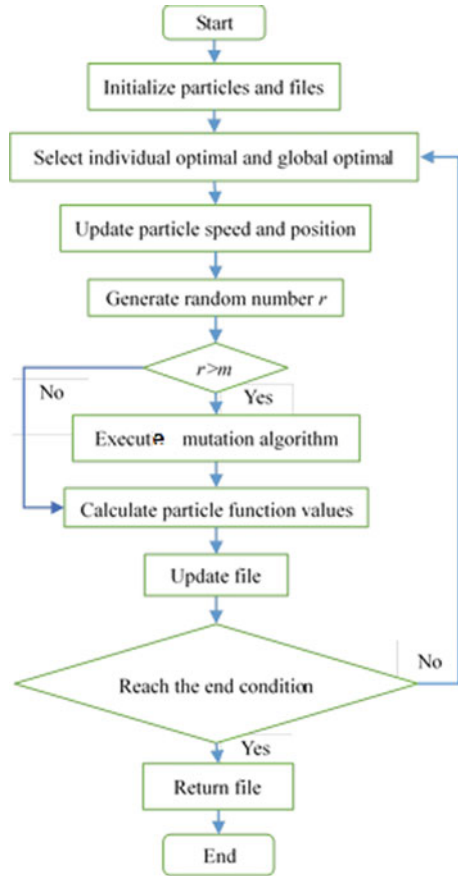
Mutation algorithm in particle swarm optimization In the particle swarm optimization algorithm, the mutation algorithm specifically refers to that when all the particles of the population update the current state of the individual in an updated generation, there is a certain probability to randomly extract one particle in the population, so that a certain dimension or all of its position. The dimension is reinitialized. In theory, it can expand the active area of the particle to a certain extent, reducing the risk that the algorithm falls into local optimum. The variation frequency of the conventional mutation algorithm is fixed. The principle is to preset a constant m between [0, 1]. The algorithm generates a value between 0 and 1 after each state of a particle is updated. The random number r , when $r > m$, randomly extracts a certain position dimension of the particle to execute the mutation statement. Therefore, the lower the m value, the greater the probability that the particle will mutate after the updated state. After adding the mutation algorithm, the flow of the particle swarm algorithm is shown in Fig. 21.1.

Improved mutation algorithm The improved particle swarm algorithm only mutates in the early stages of the algorithm. The working period of the particle swarm algorithm can be roughly decomposed into three parts, a less constrained development period, a stable search period, and a dense collection period. Considering that the particle's range of activity is maximized in the early stage, the population can be prevented from falling into local optimum, and the particle can have a higher frequency of variation at this stage. In the middle of the algorithm, the particle swarm wants to search in a reasonable step size so as not to cross the optimal solution [12]; at this time, if the method of maintaining high-frequency variation has a certain risk, the particle will end the search for part of the region ahead of time and make a mistake. The particles skip the optimal solution [13]. Therefore, the improved mutation algorithm will cause the particles to mutate at high frequency in the early stage. The frequency of the mutation decreases with iteration, and the mutation process gradually stops in the middle. This process is realized by modifying the m value:

$$m = \begin{cases} 5i/T, & i \leq 0.2T \\ 1, & 0.2T < i \leq T \end{cases} \quad (21.8)$$

Equation (21.8) is the m expression of the improved mutation algorithm. In the first quintile of the optimization, the probability of mutation linearly decreases from 100% to 0. When a particle mutates, its position dimension is reinitialized in the solution space.

Fig. 21.1 Flow chart of the mutated PSO algorithm



21.3 Experiments and Discussion

In order to test the performance of the improved algorithm, the test functions [14, 15] such as Schaffer, Rastrigin, Ackley, Rosenbrock, Shubert, and Griewank were selected to test the algorithm before and after the improvement. The experiment was carried out in two steps. First, the improved performance of the piecewise linearity was tested, followed by an improved test of the mutation algorithm. The parameter $c_1 = c_2 = 1.49445$ remains constant. In the exponential descent strategy, the adjustment coefficient c_3 is set to 10. The population size is set to 20. At the maximum speed, an expression of $V_{\max} = 0.5 * |X_{\max} - X_{\min}|$ is used. Each group of experiments was independently repeated 10,000 tests, and the experimental results were taken as arithmetic mean.

21.3.1 Experiment on Inertia Weight Strategy

Schaffer function and related tests The expression of the Schaffer function is shown in (21.9):

$$f_1 = \frac{\left(\sin \sqrt{x_1^2 + x_2^2}\right)^2 - 0.5}{\left(1 + 0.001 \times (x_1^2 + x_2^2)\right)^2} + 0.5, \quad x_i \in [-100, 100] \quad (21.9)$$

When the number of iterations is 30, the experimental data is given in Table 21.1.

For the convenience of timing, Table 21.1 lists the total duration of the algorithm cycle 5000 times.

Griewank function and related tests The expression of the Griewank function is shown in (21.10):

$$f_2 = \frac{1}{4000} \sum_{i=1}^n x_i^2 - \prod_{i=1}^n \cos\left(\frac{x_i}{i^{\frac{1}{2}}}\right) + 1, \quad x_i \in [-600, 600] \quad (21.10)$$

The global minimum value is 0. In the experiment, the optimization dimension is set in 10 dimensions, and the average optimization results obtained are shown in Table 21.2.

Rosenbrock function and related tests The Rosenbrock function expression is shown in (21.11):

$$f_3 = \sum_{i=1}^n \left[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2 \right], \quad |x_i| \in [0, 2.048] \quad (21.11)$$

Table 21.1 Schaffer function optimization results

Decline method	Inertia weight interval	Optimization result	Total time (s)
Self-adaptive	0.95–0.4	−0.9927705072	109.230513
	1.3–0.8	−0.9995883456	110.438924
Exponential	1.3–0.8	−0.9959402668	81.202922
Piecewise linear (0:0:0:1)	0.95–0.4	−0.9975228308	78.750133
	1.3–0.8	−1	80.388402
Piecewise linear (0:0:1:10)	0.95–0.4	−0.9970870357	78.832222
	1.3–0.8	−1	78.437102
Linear	1.3–0.8	−0.9992841241	78.325553
	0.95–0.4	−0.9932311664	78.915517

Table 21.2 10-dimensional Griewank function optimization results

Decline method	Number of iterations: 30		Number of iterations: 500	
	Inertia weight interval: 0.95–0.4	Inertia weight interval: 1.3–0.8	Inertia weight interval: 0.95–0.4	Inertia weight interval: 1.3–0.8
Self-adaptive	3.414528011	3.148078106	0.625433003	0.018070961
Exponential	2.223207988	3.793199799	0.877454287	0.217150569
Piecewise linear 8:4:2:1	2.192629299	3.182405984	0.316146847	0.270820153
Piecewise linear 0:0:0:1	2.75245875	1.588307524	0	0.294277477
Piecewise linear 0:0:1:10	2.794118027	1.400035895	0.036250868	0.33845017
Linear	2.211153999	1.931402254	0.130402658	0.271730083

The global minimum value of the function is 0. In the experiment, the optimization dimension is set in 10 dimensions.

Analysis and conclusion In addition to the Schaffer function, the Griewank function, and the Rosenbrock function, the Ragtrigin, 10-dimensional Ackley, Shubert, and 10-dimensional Sphere functions are also tested using the weighting intervals [0.4, 0.95] and [0.8, 1.3].

Among them, the test group with the ratio of 0:0:0:1 and 0:0:1:10 can reach the accuracy requirement first on the Schaffer test function. The experimental results mainly show two characteristics: First, the algorithm optimization effect of the convex function strategy is better than the algorithm optimization effect using the concave function; secondly, the optimization effect on the inertia interval [0.8, 1.3] is better than that on inertia interval [0.4, 0.95]. A similar conclusion was obtained when testing the 2-dimensional Rastrigin.

In the optimization test of 10D (10-dimensional) Griewank and 10D Ackley and 10D Sphere three kinds of functions, the mean value does not meet the accuracy requirement only when iterating 30 times, but the experimental group with a slope ratio of 0:0:1:10 is more linear than the linear one. The experimental error of the drop was reduced by approximately 27.5%. The optimization effect on the inertia weight [0.8, 1.3] interval is also better than that on the [0.4, 0.95] interval.

In fact, take 100 results from each experimental group of the Griewank function as observation samples, it can be noted that the optimization effect under the convex function strategy is significantly two-level differentiation—either reaching e–10 accuracy or falling into extreme local optimum. The results of those severe failures constrain the accuracy of the average optimization value. However, only consider the correct rate, the effect of optimization under the convex function strategy is much higher than that of the concave function strategy control group.

On the 10-dimensional Rosenbrock and Shubert functions, the test group using the concave function strategy has better optimization precision than the convex function

only when iterating 30 times. Moreover, the optimization results performed under the [0.4, 0.95] interval are generally better than those performed in the interval [0.8, 1.3]. This situation is especially evident in the test group of the convex function.

However, when the number of iterations is appropriately extended, such as 500 generations or even 1000 generations, most of the above conclusions are no longer applicable. The accuracy of the convex function strategy over the [0.4, 0.95] interval exceeds its accuracy over the [0.8, 1.3] interval, while the concave function strategy is more applicable to the remaining functions except the Sphere function over the [0.8, 1.3] interval. On all selected test functions, when the convex function strategy is matched with the [0.4, 0.95] inertia interval, its accuracy has been roughly flat or even exceeds the optimization effect under the concave function strategy.

The contribution of adaptive inertial strategy to accuracy is limited, but it takes the longest time. In Table 21.1, the accuracy of adaptive inertial strategy is less than $1e-10$, but the average time spent is 38.7% more than that of the experimental group with the ratio of 0:0:0:1.

Therefore, when the number of iterations is small, such as only iterating for 30 generations, either using a concave function strategy with a variation interval of [0.4, 0.95] or using a convex function strategy with a variation interval of [0.8, 1.3] has a better optimization result than default linear descent strategy. The specific choice is mainly influenced by the nature of the test function itself. After 500 generations or even 1000 generations of iterations, from the perspective of optimization stability and optimization precision, the convex function optimization strategy matched with an interval of [0.4, 0.95] can obtain the most ideal effect.

21.3.2 Mutation Algorithm-Related Experiments and Analysis

For proving the superiority of the improved mutation algorithm, the experiment uses the inertia weight reduction method with slope ratio of 0:1:2:12 in each stage to test and sets parallel groups using different mutation methods, which are grouped as follows:

Group 1: no mutation is used; group 2: m value is always maintained at 0.8, particle resets all-position dimension when mutation; group 3: m is constant at 0.5, particle resets all-position dimension when variation; group 4: m is based on Formula (21.8) Variation, the particle resets the full position dimension when the mutation, this group is the modified mutation algorithm; the group 5: m changes strictly according to the (21.8), the mutation only resets the single dimension; the group 6: m is mutated according to Formula (21.8) in the first five fifths, and the probability of variation is 0.5 in the last fifth of the period; and group 7: the value of m mutate in the first five fifths according to the (21.8), while the probability of mutation remains at 0.8 in the last quintile.

On the 30-dimensional Rosenbrock function, each experimental group algorithm performs 500 generations, and each group of algorithms is cycled 10,000 times to obtain an average test result as shown in Table 21.4.

It can be seen from the records in Table 21.4 that the use of the mutation algorithm has a certain positive effect on the optimization. However, groups 2 and 3, which have mutated behaviors in each period, are inferior in accuracy to group 4 that only mutated in the first quintile. Comparing the results of group 4 with group 5, it can be seen that the effect of making the full-scale variation of the particles is better than changing the effect of the single dimension. Comparing group 4 with group 6 and group 7, it can be considered that if the variation behavior is continued in the later stage, the optimization precision of the algorithm will be damaged.

Then, the algorithm is tested on Ackley, Rastrigin, Griewank, and Sphere. The minimum function value of these four functions is 0, and the solution space dimension is set at 30 dimensions. When the error between the optimization result and the true optimal function value of the test function is within $1e-7$, the number of iterations that the algorithm has been running at this time is called the average necessary number of iterations. The control group used the linear decline strategy without any mutation algorithm. Both experimental groups A and B used an improved mutation algorithm, group A used an inertial variation strategy with a ratio of 0:1:2:12, and group B used a linear reduction strategy. Here, the maximum number of iterations is 500. Since the final average optimization results of group 1 and group 2 are both 0, they are not listed, and Table 21.5 is obtained.

Table 21.4 Rosenbrock function results

Group	Average optimization result
1	77.23301137
2	0.0000003313468723
3	1.010948118E-12
4	5.01817044E-25
5	45.01632424
6	0.0000001284314917
7	0.00000006138500513

Table 21.5 Optimization results of the four types of function

Test function	Average necessary number of iterations		Average optimization results of the control group
	Group A	Group B	
Ackley	37.59766667	40.0345	9.637387932
Rastrigin	35.96	38.20933333	124.953894
Griewank	37.186	40.72566667	26.69177666
Sphere	37.79066667	40.00366667	7.635740014

It can be seen from Table 21.5 that, when the mutation algorithm is not used for improvement, the average optimization result of the particle swarm algorithm does not reach the desired accuracy in the four types of test functions. The use of the improved particle swarm optimization algorithm enables the particle swarm algorithm to both achieve the accuracy requirements of the algorithm and reduce the number of iterations necessary.

21.4 Conclusion

This paper proposes an improved particle swarm optimization algorithm. The improved algorithm is based on two aspects of inertia weight and increasing mutation algorithm. It is pointed out that the optimization strategy of convex function is more suitable for high-dimensional test function, and the parameter setting strategy of improved mutation algorithm is proposed. By adopting the convex function descent strategy, supplemented by the mutation algorithm only in the initial stage, the improved particle swarm algorithm can complete the accurate optimization of 30-dimensional Ackley, Rastrigin, Griewank, and Sphere in 500 generations. In the cycle of about 37 generations, the optimization result is accurate to $1e-7$.

But in fact, when combining different improvement strategies, there will be a certain coupling effect inside the algorithm. The impact of different improvement measures on the performance of the algorithm is not simply additive, but may be multiplication or even mutual restraint. In the follow-up work, the author will consider more of this impact and conduct more detailed research on existing findings, looking for a set of more general applicability of the proportion of polyline. The future research will mainly lie in replacing the polyline with a smoother descent curve. It is expected to integrate the existing conclusions into the research work of multi-objective optimization in the next stage.

References

1. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proceeding of IEEE International Conference on Neural Networks, pp. 1942–1948. Piscataway (1995)
2. Yang, X.S.: A new meta heuristic bat-inspired algorithm. *Comput. Knowl. Technol.* **6**(23), 6569–6572 (2010)
3. Pathak, V.K., Singh, A.K.: A modified algorithm of particle swarm optimization for form error evaluation. *TM Tech. Mess.* **84**, 272–292 (2017)
4. Liu, W., Sun, R.B., Wang, H.R.: Escape from the immune particle swarm algorithm embedded mechanism of simulated annea. *Jilin Normal Univ. J. (Nat. Sci. Ed.)* **39**(1), 85–90 (2018)
5. Bouyer, A.: An optimized K-harmonic means algorithm combined with modified particle swarm optimization and cuckoo search algorithm. *Found. Comput. Decis. Sci.* **41**(2), 99–121 (2016)
6. Shi, Y., Eberhart, R.C.: A modified particle swarm optimizer. In: Proceedings of IEEE International Conference on Evolutionary Computation, pp. 69–73. Anchorage (1998)

7. Shi, Y., Eberhart, R.C.: Empirical study of particle swarm optimization. In: Proceedings of the World Multiconference on Systemic, Cybernetic and Informatics, pp. 1945–1950. Orlando, FL (2000)
8. Liu, W., Zhou, Y.R.: Modified inertia weight particle swarm optimizer. *Comput. Eng. Appl.* **45**(7), 46–48 (2009)
9. Chen, G.M., Jia, J.Y., Han, Q.: Study on the strategy of decreasing inertia weight in particle swarm optimization algorithm. *J. Xi'an Jiaotong Univ.* **40**(1), 53–61 (2006)
10. Fu, G.J., Wang, S.M., Liu, S.Y.: A PSO with dimension mutation operator. *J. Wuhan Univ. Hydraul. Electr. Eng.* **38**(4), 79–83 (2005)
11. Wang, W.B., Lin, C., Zheng, Y.K.: Experiment and analysis of parameters in particle swarm optimization. *J. Xihua Univ. (Nat. Sci. Ed.)* **27**(1), 76–80 (2008)
12. Clerc, M., Kennedy, J.: The particle swarm: explosion, stability, and convergence in multi-dimension complex space. *IEEE Trans. Evol. Comput.* **20**(1), 1671–1676 (2002)
13. Li, N., Liu, F., Sun, D.B.: A study on the particle swarm optimization with mutation operator constrained layout optimization. *Chin. J. Comput.* **27**(7), 899–902 (2004)
14. Eberhart, R.C., Shi, Y.: Comparing inertia weights and constriction factors in particle swarm optimization. In: Proceedings of the Congress on Evolutionary Computing, San Diego, USA, pp. 84–89. IEEE Service Center, Piscataway (2000)
15. Ge, R.P., Qin, Y.F.: The globally convexized filled functions for global optimization. *Appl. Math. Comput.* **35**(2), 131–158 (1990)

Part IV
Machine Learning

Chapter 22

Performance of Probabilistic Approach and Artificial Neural Network on Questionnaire Data Concerning Taiwanese Ecotourism



Vladislav Bína, Václav Kratochvíl, Lucie Váchová, Radim Jiroušek, and Tzong-Ru Lee

Abstract This paper aims to perform modeling of Taiwanese farm and ecotourism data using compositional models as a probabilistic approach and to compare its results with the performance of an artificial neural network approach. Authors use probabilistic compositional models together with the artificial neural network as a classifier and compare the accuracy of both approaches. The probabilistic model structure is learned using hill climbing algorithm, and the weights of multilayer feedforward artificial neural network are learned using an R implementation of H2O library for deep learning. In case of both approaches, we employ a non-exhaustive cross-validation method and compare the models. The comparison is augmented by the structure of the compositional model and basic characterization of artificial neural network. As expected, the compositional models show significant advantages in interpretability of results and (probabilistic) relations between variables, whereas the artificial neural network provides more accurate yet “black-box” model.

V. Bína (✉) · V. Kratochvíl · L. Váchová · R. Jiroušek
Faculty of Management, University of Economics, Prague, Czech Republic

Institute of Information Theory and Automation, Czech Academy of Sciences,
Prague, Czech Republic
e-mail: bina@vse.cz

V. Kratochvíl
e-mail: velorex@utia.cas.cz

L. Váchová
e-mail: vachova@vse.cz

R. Jiroušek
e-mail: radim@utia.cas.cz

T.-R. Lee
Department of Marketing, National Chung Hsing University, Taichung, Taiwan
e-mail: trlee@dragon.nchu.edu.tw

22.1 Introduction

When analyzing a sample data set, we cannot infer any results concerning some population with certainty. Therefore, during last more than one hundred years, many tools were developed which are able in a way to handle data having the uncertain nature. They start from the basic test methods of mathematical statistics and so far arrived at the probabilistic and other alternative algorithms and systems more or less capable of serving as a basis for artificial intelligence approaches.

The probabilistic graphical models were considered as a standard tool for support of decision making under uncertainty and become popular as a tool interconnecting probabilistic description and graphical presentation. The approaches for decision-support initially started from the simple managerial tools assessing causes of problematical status. Then, it continued as qualitative schemes of influence diagrams and evolved into the tools of flowcharts, causal loop diagrams, stock and flow diagrams and in the last three decades developed influence diagrams based on Bayesian networks. However, the usability and local computations of Bayesian networks (see, e.g., Jensen and Nielsen [1]) bring together one important source of confusion, namely its graphical representation using directed acyclic graphs. The arrows used in such graphical tools resembles less experienced users direction from causes to effects which is rather intuitive and usually accepted but not correct. Moreover, this graphical presentation is of only little use in case of large diagrams having higher tens or even hundreds of nodes (as we show below). In these situations, one may acknowledge an alternative, yet equivalent algebraic approach based on compositional models Jiroušek [2] and Jiroušek and Kratochvíl [3].

In recent year, based on better algorithms of deep learning and higher performance of computers, the approach of artificial neural networks become successful in a wide variety of tasks including computer vision, text and opinion mining, machine translation, image and video processing, etc. It is sometimes referred as a “deep learning revolution” (see, e.g., Sejnowski [4]). Similar to the comparison of probabilistic and neural network approaches by Tavana et al. [5] or by Simfukwe et al. [6], the aim of this paper is to compare the accuracy of the powerful and well-developed methodology of artificial neural networks with a (yet developing) probabilistic approach of compositional models. Of course, the main disadvantage of compositional models is rather the current state of the art which (unlike the case of neural networks) lack a professional and user-friendly implementation of algorithms. But still, compositional models represent an easy way to interpret “white-box” approach and have thus a great advantage in comparison with neural networks. On a simple (but not toy) example data set concerning Taiwanese farm and ecotourism, we present that under certain circumstances, the performance of compositional models is comparable to the performance of artificial neural networks.

22.2 Brief Summary of Theoretical Background

Throughout the paper, we use two modeling approaches. Compositional models as a probabilistic model structure a multilayer feedforward artificial neural network. In this section, we will briefly characterize both approaches and set a basis for their comparison using a Taiwanese farm and ecotourism data.

22.2.1 Compositional Models

The theory of compositional models was summarized in Jiroušek [2] with the important structural properties summarized in Jiroušek and Kratochvíl [3]. Similarly, we will adopt the notation in the following sense. Throughout this paper, we analyze a set of n finite valued variables $\{X_1, X_2, \dots, X_n\}$. Subsets of variables are denoted by lower-case Roman alphabets (e.g., x, y , and z). $\langle X_i \rangle$ denotes the set of values (states) of variable X_i . Analogously, for sets of variables x, y , the respective Cartesian products of all combinations of their values are denoted by $\langle x \rangle, \langle y \rangle$, respectively. Elements of these sets, i.e., the (combinations of) values of variables will be denoted by lower-case boldface Roman characters (e.g., $\mathbf{a} \in \langle x \rangle$).

Conditional distributions will be denoted using a standard notation, e.g., $\pi(y | X)$. In case that we consider conditioning by a specific value of variable X by $\pi(y | X = \mathbf{a})$, let us stress that since we deal with finite valued variables, the conditional distribution $\pi(y | X)$ is represented by a table where $\pi(y | X = \mathbf{a})$ is a probability distribution for each $\mathbf{a} \in \langle x \rangle$.

For a probability distribution $\pi(x)$ its *marginal* distribution for $y \subset x$ is denoted either by $\pi(y)$ or by $\pi^{\downarrow y}$. Under a notion of extension (in a way opposite to marginalization), we understand any distribution κ defined for a superset of variables, i.e., $\kappa(z)$ for $z \supset x$, such that $\kappa(x) = \pi(x)$. The set of all extensions of distribution $\pi(x)$ for variables $z \supset x$ will be denoted by $\Psi[\pi; z]$. The symbol $\pi(x \cap y) \ll \kappa(x \cap y)$ denotes that $\kappa(x \cap y)$ *dominates* $\pi(x \cap y)$. This holds (in the considered finite setting) when

$$\kappa^{\downarrow x \cap y}(\mathbf{b}) = 0 \implies \pi^{\downarrow x \cap y}(\mathbf{b}) = 0$$

for all $\mathbf{b} \in \langle x \cap y \rangle$.

Now, consider two distributions $\pi(x)$ and $\kappa(y)$. Obviously, there exists their joint extension if and only if they are *consistent*, i.e., if $\pi(x \cap y) = \kappa(x \cap y)$. In case that they are not consistent then one can be interested in getting an extension of π containing from κ *as much information as possible*. Speaking more precisely, one can look for a distribution $\mu(x \cup y)$ that is a *projection* of κ into the set of all extensions $\Psi[\pi; x \cup y]$:

$$\mu(x \cup y) = \arg \min_{\lambda \in \Psi[\pi; x \cup y]} \text{Div}(\lambda(y); \kappa(y)).$$

If we consider a Kullback–Leibler divergence, Theorem 6.2 in [2] states that this type of projection can be got as a composition of π and κ . The composition is defined only if $\pi(x \cap y) \ll \kappa(x \cap y)$ by the formula

$$\mu(x \cup y) = \pi(x) \triangleright \kappa(y) = \frac{\pi(x) \cdot \kappa(y)}{\kappa(x \cap y)}.$$

The use of the operator of composition can be iterated. The result of the repeated application to the sequence of low-dimensional distributions is (if defined) a multi-dimensional distribution which can be written in the following way:

$$\kappa_1 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \dots \triangleright \kappa_n := (\dots ((\kappa_1 \triangleright \kappa_2) \triangleright \kappa_3) \triangleright \dots) \triangleright \kappa_n.$$

Throughout this paper, we will focus on the models composed from the marginals of one data distribution, thus there are no inconsistent distributions.

22.2.2 Artificial Neural Networks

An approach of artificial neural networks (ANN) had roots in 1940s when the first computational model for neural networks was developed (see McCulloch and Pitts [7]). The methodology of ANN became useful and extensively employed after the development of backpropagation algorithms (see Schmidhuber [8]). But the case of networks with higher number of hidden layers tended to give worse results than the shallow networks (see, e.g., Alom et al. [9]). The success of deep learning approaches together with improved computational capacities (use of GPUs and employment of its vector computation features) resulted in last ten years in a so-called deep learning revolution, i.e., a radical change of artificial intelligence industry and massive use of ANNs (this breakthrough resulted in awarding of Turing Award in March 2019).

Let us clarify that under the notion of deep network we understand an artificial neural network with at least three hidden layers, oppositely, the ANN with one or two hidden layers is called shallow (this is quite frequent classification (see, e.g., [9])).

One of the modern and most successful open-source systems for artificial intelligence is a H2O platform capable of analyzing (using in-memory compression) huge data samples. Moreover, it has a linear scalability and is able to interconnect with, e.g., R, Python, and Hadoop (see [10]). We used it as an implementation of artificial neural network and its learning under R software (we used a version 3.6.1 [11]) augmented by a H2O package version 3.26.0.2 [12].

In our case, we use the artificial neural network as a model of dependencies among the set of categorical variables (most of them binomial, two multinomial variables). Because of the architecture of ANN layers, we need to choose one of the variables as a variable in an output layer. The H2O implementation does not provide a possibility

to handle multiple output variables. The nonlinear character of activation functions provides a possibility to perform a classification task.

The considered artificial neural network has a structure of a multilayer feedforward (nodes do not form a cycle) neural network (perceptron). The network is trained with stochastic gradient descent algorithm based on backpropagation, and it is necessary to specify the number of network layers and number of neurons in each layer, see [10]. The binary character of variables allowed to choose among different activation functions (the Tanh activation function performed with the analyzed data set better than ReLU and Maxout). For the classification into multiple classes, i.e., for the multinomial output layer, a softmax activation function can be used. See, e.g., Glorot et al. [13].

Though the H2O package is capable to efficiently handle huge data sets thanks to the parallelization of its procedures, this was not our case. However, relatively small Taiwanese data were divided one hundred times in order to perform multiple training and validation cycles.

22.2.3 Measures Based on Confusion Matrix

The paper employs a basic set of measures derived from the confusion matrix (see, e.g., Fawcett [14]). The confusion matrix visualizes the correspondence of predicted class based on model and class observed in data in a contingency table with the setting according to Table 22.1.

Sensitivity (or true positive rate) is defined as

$$TPR = \frac{TP}{TP + FN},$$

specificity (or true negative rate) is given by

$$TNR = \frac{TN}{TN + FP},$$

precision (or positive predictive value) is defined as

$$PPV = \frac{TP}{TP + FP},$$

Table 22.1 Confusion matrix

	Observed class	
Predicted	True positive (TP)	False positive (FP)
Class	False negative (FN)	True negative (TN)

under *accuracy* we understand

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

and *F1 score* is a harmonic mean of precision and sensitivity which result into

$$F1 = \frac{2TP}{2TP + FP + FN}.$$

22.3 Data Set and Preprocessing

The Taiwanese farm and ecotourism data set contains answers of 1235 respondents who filled the questionnaire in the period from 2015 to 2017. The answers to the questionnaire in case of the first ten respondents are shown in Table 22.2 as an illustration of sample data set.

The answers of six main multiple-choice questions were converted into 51 binary variables containing answers accompanied by respondent’s gender and age category (0–18; 19–25; 26–35; 36–45; 46–55; 56–65 and 66–). The six main questions are:

- Q1** Reasons why you would like to focus on agricultural information.
- Q2** What kinds of instant message you would like to see.
- Q3** What kinds of products or stories that you are interested in.
- Q4** Reasons why you are interested in participating in work exchange.
- Q5** What kinds of workshop you are interested in.
- Q6** What did you experience from the ecotourism.

Table 22.2 Record of answers of the first ten respondents in the Taiwanese farm and ecotourism data set

Que. 1	Que. 2	Que. 3	Que. 4	Que. 5	Que. 6	Age	Gender
3, 4	1, 2, 8	F, I, L	1, 3	1, 2	1, 2, 3	36–45	Female
3, 4, 5	1, 2, 3, 4, 5, 6, 7, 8	4, 8, 9, A, G, H, K, L	1, 2	1, 2	5	19–25	Female
3, 4	3, 4, 8	4	1	1, 2	5	36–45	Male
1	1, 6, 7, 8	4, A	5	1	1, 2	19–25	Female
5	8	3, 8, D, K, L	3	1	5	36–45	Female
4	7	6, A, B, F, I	1, 2, 3, 4	3	5	36–45	Male
2, 3	1, 7	C, D	1, 3, 4	1, 2	1, 2, 4	46–55	Female
3, 4, 5	4, 5, 7, 8	2, D, I, O	1, 3	2, 3	1, 3	26–35	Male
3, 5	1, 8	3, 4, A, B, K	3	1	1, 2	36–45	Female
3, 4	6, 7, 8	4, 7, B, L	3, 4	1	2	26–35	Male

Since all the particular questions are not significant for the sake of comparison of two modeling approaches, we will not describe in detail all possible answers to the six main questions. Let us only mention that there were 5 possible answers for the question Q1, 8 answers for the Q2, 24 answers for the Q3, 6 answers for the Q4, and 3 answers for Q5. Now, let us focus on the answer to the 6th question which will be analyzed as predicted variables. The possible answers to a multiple-choice question 6 were:

- Q6.1** Agricultural experience and understand planting methods.
- Q6.2** Enjoy local natural food.
- Q6.3** Special festivals participation.
- Q6.4** Local culture exchange.
- Q6.5** Not yet experienced.

22.4 Models

The main result of the presented paper includes two types of different modeling approaches describing the Taiwanese farm and ecotourism data. The first approach uses probabilistic compositional models, whereas the second approach employs the artificial neural network methodology.

22.4.1 Resulting Compositional Model

The structure of a compositional model is learned from the data set using a structural EM algorithm (see Friedman [15]) where a maximization step is performed using a tabu search generalization of hill climbing greedy approach in the space of models structures (see Russell and Norvig [16]). The resulting compositional model can be (obviously) written as a model formula. However, because of its rather complex structure and long formula let us only take a taste on the shortened expression, i.e.,

$$\hat{\mu} = \pi(x_{3,4}) \triangleright \pi(x_{3,O}) \triangleright \pi(x_{3,4}, x_{3,A}) \triangleright \pi(x_{3,A}, x_{3,B}) \triangleright \pi(x_{3,4}, x_{3,5}, x_{3,B}) \triangleright \dots$$

As the kind author can imagine, because of the big number of variables the complete compositional model formula would be very long and not transparent. Instead of this, we use an equivalent representation, i.e., a structure visualization using a graphical tool of persegram (introduced by Jiroušek [17], for an application, see, e.g., Kratochvíl [18]). This tool describes a dependency structure of considered variables¹ and is capable to clearly present both all the particular distributions composed

¹In Jiroušek [17] the following assertion is formulated: “Every independence statement read from the structure (or its persegram) of a compositional model corresponds to probability independence statement valid for every multidimensional probability distribution represented by a compositional model with this structure.”

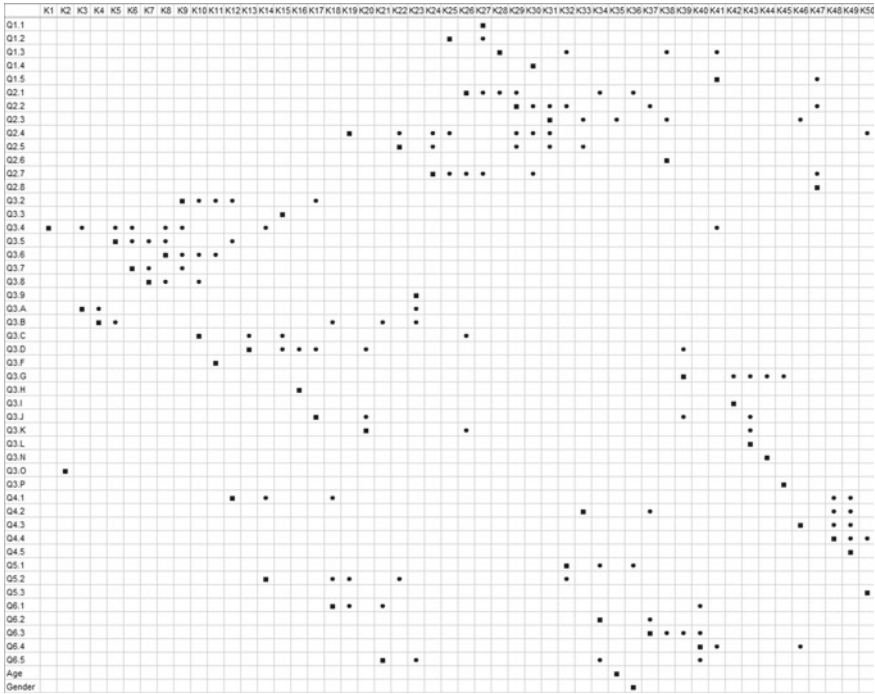


Fig. 22.1 Persegram of a compositional model structure learned from the Taiwanese farm and ecotourism data

in the model (as columns in the table) and occurrence of a variable throughout the distributions within a model (as rows of the table) where the first occurrence is marked by a square and all other occurrences are marked using a bullet. The model learned from an example data set is depicted in Fig. 22.1 and provides an easy insight into the dependence structure of considered variables. Namely, the answers to question 4.1–4.5 shows an apparent interdependency, or similarly, the answers to questions 3.2, 3.4–3.8.

Probably, the readers are more accustomed to the expression of probabilistic models in the form of a directed acyclic graph (DAG). Figure 22.2 shows the above-described dependence structure in the form of DAG, which is obviously rather hard to read and to search the particular (conditional) dependencies. Moreover, the arrows in the graph might be misleading and can lead to an incorrect interpretation as causal relations.

Particular distributions and their conditional variants usable for the process of composition can be easily computed from the data, e.g., conditional distribution $\pi(x_{6.1} \mid x_{3.B}, x_{4.1}, x_{5.2})$ can be summarized in a form of Table 22.3.

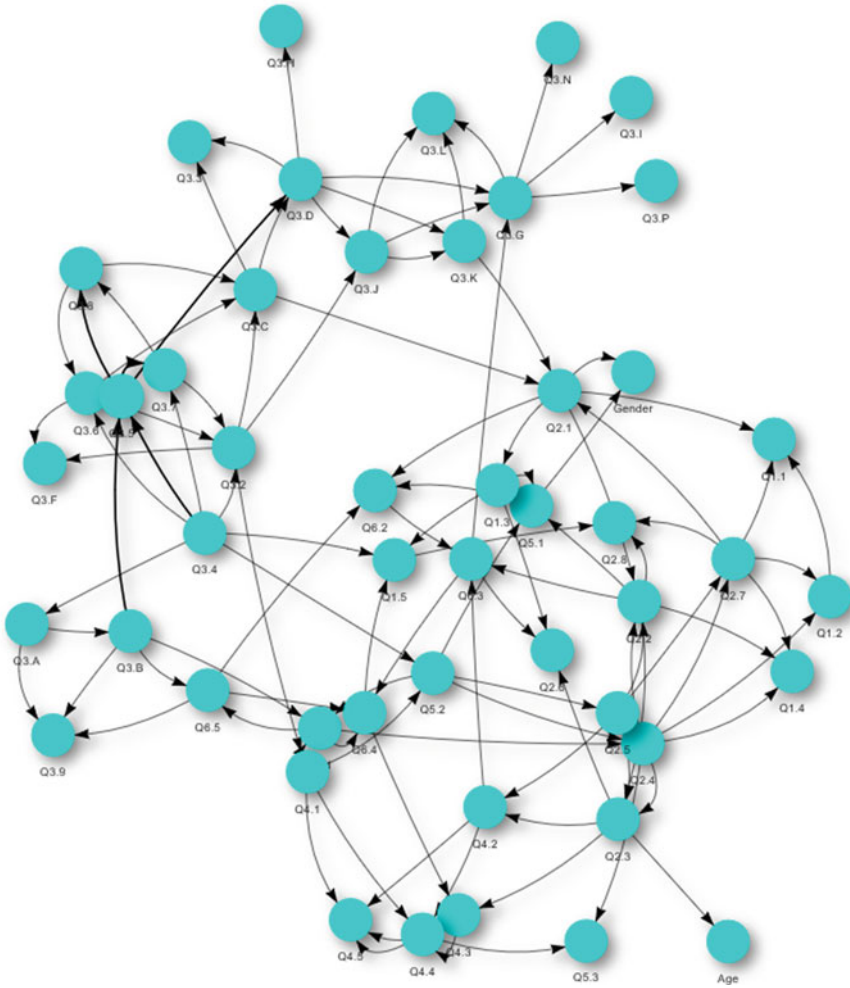


Fig. 22.2 A directed acyclic graph corresponding to a compositional model structure learned from the Taiwanese farm and ecotourism data

Table 22.3 An example of one probability distribution from compositional model $\hat{\mu}$ in a form of conditional probability table, namely $\pi(x_{6,1} | x_{3,B}, x_{4,1}, x_{5,2})$

		Q5.2 = yes				Q5.2 = no			
		Q4.1 = yes		Q4.1 = no		Q4.1 = yes		Q4.1 = no	
		Q3.B		Q3.B		Q3.B		Q3.B	
		Yes	No	Yes	No	Yes	No	Yes	No
Q6.1	Yes	0.639	0.909	0.500	0.634	0.231	0.671	0.258	0.365
	No	0.361	0.091	0.500	0.366	0.769	0.329	0.742	0.635

22.4.2 Resulting Artificial Neural Network

Deep learning algorithm was used five times to build five models. Each of the models has one of the questions 6.1–6.5 as an output node. The method in each case learned the weights of five layers multilayer perceptron with 162 neurons in an input layer, 100, 40, and 6 neurons in hidden layers with tanh activation function and an output layer with two nodes uses a softmax activation function. The algorithm in each case learned values of 20,600 weights and biases.

The creation of five models was necessary since the current implementation of H2O does not support multiple response columns. The authors suggest to train a new model for each response variable. An example of model metrics achieved for the model with an output being a question 6.1 summarized in Table 22.4. Metrics of all particular models and their comparison is performed in the next subsection.

22.4.3 Comparison of Both Types of Models

In each case, the data set was divided into a training frame and a validation frame in the ratio of 80–20%. In other words, the training process was conducted on a sample of 1121, and validation was conducted on 114 statistical units. For both approaches, a non-exhaustive cross-validation method was used which do not compute all possible ways of the splitting of the original sample being thus an approximation of leave-p-out cross-validation approach. In each case, we used 100 iterations of the cross-validation.

The results of the comparison are summarized in Table 22.5 where the metrics of sensitivity, specificity, precision, accuracy, and F1 score are provided for both types of models and for each of the five questions from 6.1 to 6.5. This numerical comparison is augmented by a graphical presentation in Fig. 22.3 of ROC space graphs for each of the five questions.

Table 22.4 An example of basic maximum metrics and their indices in case of model for Q6.1

Metric	Value	Index
Max F1	0.867	95
Max accuracy	0.798	79
Max precision	0.966	28
Max sensitivity	1.000	101
Max specificity	0.973	0

Table 22.5 Set of measures based on confusion matrix of artificial neural network (ANN) and compositional model (CM) for particular questions from 6.1 to 6.5

Question	Model	Sensitivity	Specificity	Precision	Accuracy	F1 score
6.1	ANN	0.835	0.887	0.969	0.841	0.897
	CM	0.921	0.358	0.785	0.759	0.844
6.2	ANN	0.801	0.867	0.959	0.810	0.872
	CM	0.959	0.413	0.776	0.782	0.856
6.3	ANN	0.646	0.852	0.759	0.758	0.692
	CM	0.585	0.771	0.598	0.704	0.581
6.4	ANN	0.692	0.821	0.851	0.736	0.760
	CM	0.646	0.852	0.759	0.758	0.692
6.5	ANN	0.969	0.993	0.947	0.990	0.957
	CM	0.145	0.983	0.538	0.886	0.159

22.5 Conclusion

The paper presented two model approaches for modeling of categorical data. The compositional model approach was applied to build and to use one model approximating the whole data set. The approach of artificial neural networks was employed in order to create five particular models with an output variable of each of five questions from 6.1 to 6.5.

The comparison of both approaches showed that in the case of questions 6.1 and 6.2, both approaches provided more or less similar quality of models. In case of questions 6.3 and 6.4, the approach of artificial neural networks provided a model of higher quality, and finally, in case of question 6.5, the approach of compositional models failed to provide reasonable predictions, whereas the artificial neural network approach was very successful. This was caused by an answer collecting in a way the rest of respondents and having an unbalanced ratio of answers to question 6.5 (142 positive answers and 1093 negative answers) in contrary to more or less comparable frequencies of both answers in case of other answers.

This documents the most serious limitation of the approach of compositional models. But let us mention an important advantage that the compositional models represent a white-box approach, i.e., the possibility to analyze and interpret its building blocks (the low-dimensional distributions to be composed) as probabilities usual for description of uncertainty. Moreover, similar to the approach of compositional models, the user is able to insert evidence into the model and to analyze interesting marginal distributions which can be calculated more or less easily from the compositional model. The artificial neural network approach comprises in a way a black box. It is theoretically possible to look at the weights of each neuron, but its possible interpretation is very limited.

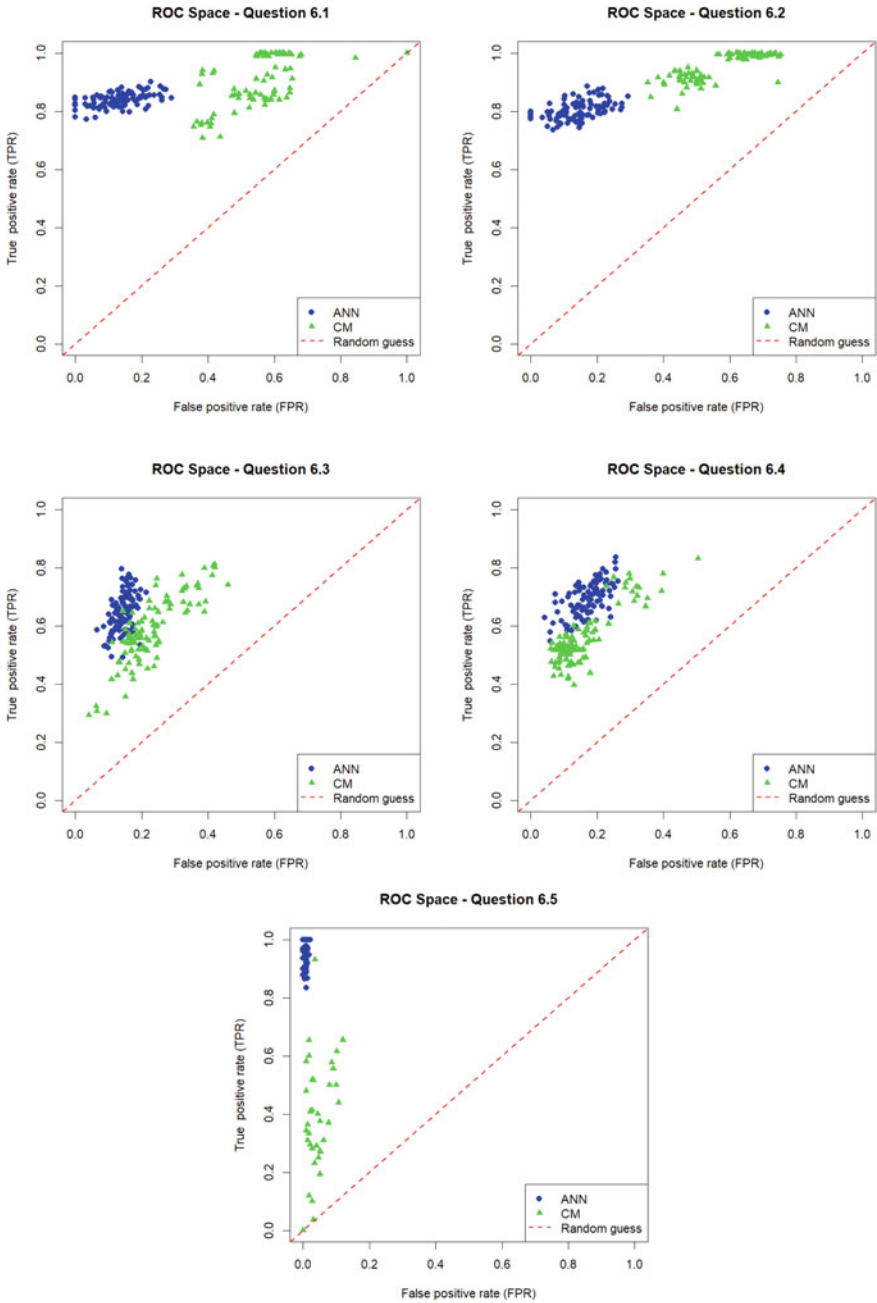


Fig. 22.3 ROC space of artificial neural network and compositional model for non-exhaustive cross-validation of prediction for particular questions from 6.1 to 6.5

Acknowledgements The research was financially supported by grants GAČR no. 19-06569S (first, second and fourth author) and AVČR no. MOST-04-18 (remaining authors).

References

1. Jensen, F.V., Nielsen, T.D.: Bayesian Networks and Decision Graphs. Information Science and Statistics Series. Springer, New York (2007)
2. Jiroušek, R.: Foundations of compositional model theory. *Int. J. Gen. Syst.* **40**(6), 623–678 (2011). <https://doi.org/10.1080/03081079.2011.562627>
3. Jiroušek, R., Kratochvíl, V.: Foundations of compositional models: structural properties. *Int. J. Gen. Syst.* **44**(1), 2–25 (2015). <https://doi.org/10.1080/03081079.2014.934370>
4. Sejnowski, T.J.: *The Deep Learning Revolution*. MIT Press (2018)
5. Tavana, M., Abtahi, A.R., Di Caprio, D., Poortarigh, M.: An artificial neural network and Bayesian network model for liquidity risk assessment in banking. *Neurocomputing* **275**, 2525–2554 (2018). <https://doi.org/10.1016/j.neucom.2017.11.034>
6. Simfukwe, M., Kunda, D., Chembe, C.: Comparing naive bayes method and artificial neural network for semen quality categorization. *Int. J. Innov. Sci. Eng. Technol.* **2**(7), 689–694 (2015)
7. McCulloch, W., Pitts, W.: A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**(4), 115–133 (1943). <https://doi.org/10.1007/BF02478259>
8. Schmidhuber, J.: Learning complex, extended sequences using the principle of history compression. *Neural Comput.* **4**, 234–242 (1992)
9. Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Hasan, M., Van Essen, B.C., Awwal, A.A., Asari, V.K.: A state-of-the-art survey on deep learning theory and architectures. *Electronics* **8**(3), 292 (2019)
10. Candel, A., Parmar, V., LeDell, E., Arora, A.: *Deep Learning with H2O*, 6th edn. H2O.ai, Inc. (2019). <http://h2o.ai/resources>
11. R Core Team: *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria (2019). <https://www.R-project.org/>
12. LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., Malohlava M.: *H2O: R interface for ‘H2O’*. R package version 3.26.0.2 (2019). <https://CRAN.R-project.org/package=h2o>
13. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323 (2011)
14. Fawcett, T.: An introduction to ROC analysis. *Pattern Recogn. Lett.* **27**(8), 861–874 (2006). <https://doi.org/10.1016/j.patrec.2005.10.010>
15. Friedman, N.: Learning belief networks in the presence of missing values and hidden variables. In: *Proceedings of the 14th International Conference on Machine Learning*, pp. 125–133 (1997)
16. Russell, S.J., Norvig, P.: *Artificial Intelligence: A Modern Approach*, 3rd edn. Prentice Hall (2009)
17. Jiroušek, R.: Persegrams of compositional models revisited: conditional independence. In: *Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 915–922 (2008)
18. Kratochvíl, V.: Probabilistic compositional models: solution of an equivalence problem. *Int. J. Approx. Reason.* **54**(5), 590–601 (2013). <https://doi.org/10.1016/j.ijar.2013.01.002>

Chapter 23

AC Operation Hardware Learning Neural Circuit Using V-F Converter System



Masashi Kawaguchi, Naohiro Ishii, and Masayoshi Umeno

Abstract In the artificial intelligence, machine learning and neural network field, many practical use models have been proposed. However, these models are based on a digital Von Neumann model's computer. There are minority studies in the field of analog learning neural networks. Early analog hardware neural network models were configured with the operational amplifier and the resistance element. It is difficult to change the value of solid resistance for the learning process. In the present paper, proposed is a learning neural network using electronic alternating current (AC) circuits including a voltage-frequency (V-F) converter. These circuits are composed of an amplifier, additional circuit, inverter, subtract circuit, rectifier circuit and V-F converter. The input voltage value was described as the input signal and the input frequency considered the connecting weights. It is easy to change the connecting weights with a V-F converter. Finally, the input frequency converges to a constant value after only several learning process. The learning count time is extremely small. The learning time is quite fast in this AC transmission circuit. The model works using pure analog electronic circuits. The learning time is quite short compare to with a digital process computer.

23.1 Introduction

Recently, multilayer neural network models, in particular, a deep learning model have been researched very sprightly. The performance has been extremely improved in the specialty of image/sound recognition. The internal mechanism of recognition system is revealed more clearly; self-learning integrated circuit (IC) chips have also been realized. However, these models are operating on a general-purpose Von

M. Kawaguchi (✉)

National Institute of Technology, Suzuka College, Shiroko, Suzuka, Mie 510-0294, Japan
e-mail: masashi@elec.suzuka-ct.ac.jp

N. Ishii

Aichi Institute of Technology, Yachigusa Yagusa-cho, Toyota, Aichi 470-0392, Japan

M. Umeno

Chubu University, 1200 Matsumoto-cho, Kasugai, Aichi 487-8501, Japan

© Springer Nature Singapore Pte Ltd. 2021

S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation, Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_23

Neumann model's computer with the application system. There are minority studies that construct an analog parallel hardware structure using a biological information processing mechanism. We proposed the neural network machine learning model with a pure analog network electronic circuit. Such a model will develop original recognition or prediction system using the analog neural electronic circuit. In the field of multilayer learning models, many useful models such as image recognition or phenomenon prediction have been proposed. And there are many hardware realization models such as an image/sound sensor or computer by parallel circuit have been enlargement.

23.1.1 Analog Hardware Neural Network

The main strong point of an analog machine learning network, its operation is by the real-time linear system, not due to the clock frequency behavior. On another side, a digital system behavior is due to the clock operation by central processing unit (CPU) based on a Neumann computer. As another researchers' study, innovative analog neural models were proposed [1, 2]. In the complete analog circuit, one element is the implementation of analog data saving unit, keeping the analog numerical value for a while time using analog memory [3]. The dynamic random access memory (DRAM) can be memorized in the condenser memory in short period of time, because it was performed in the generic complementary metal oxide semiconductor (CMOS) [4]. However, when the charge is maintained for a long time in the capacitor, it needs the processing system to keep the numerical value data in the memory. It is required the refresh process. Capacitors reduce the electric charge over time. It is not difficult to get back the electric charge of a condenser by refresh process utilization in the general digital binary memory. Nevertheless, in the situation using analog memory, the refresh process is not easy because memorizing data is linear analog data, it needs the time analysis system of electronic charge reducing curve. Other researchers proposed the memorize methods of connecting weights, the floatage gate type device [5] and magnetic substance memories [6].

23.1.2 Pulsed Neural Network

Pulsed neural networks receive great number pulses as learning data and change the connecting weights due to the number of pulses. Such networks can maintain their connecting weights after learning by the number of pulses and outputs the signal depending on the input value [7]. Nevertheless, it is necessary the long time for learning because great number pulses are required before complete learning. For example, the time interval average of the pulse is 10 μ s and 100 pulses are needed before finished the learning procedure, and it needs 1 mS for finished learning approximately.

23.1.3 How to Realize the Variable Connecting Weights?

In our former study, proposed model was a movement detection biomedical vision model using analog general electronic circuits. The suggestion model is composed of four layers. There are differentiation circuit, difference circuit and multiple circuits in each layer for detecting pure motion output. From a technical standpoint, the proposed model makes possible to elucidation of the artificial vision system mechanism, which can detect the target object, motion and velocity by the design and simulation using an analog network electric circuit [8, 9].

On the other hand, attempted to be realize was a multi-layered hardware neural network using an analog electronic circuit. In the machine learning and neural network field, many practical use models have been proposed. But, these models are based on a digital general Neumann computer. There are few studies in the practical study about analog learning neural networks. Early, analog circuit network models were configured of the difference circuits, multiple circuits and solid resistance. It is not easy to change the value of solid resistance for the learning process.

The first constructed model was a hardware neural network using variable resistance elements as the connecting weights. In the learning operation process, each resistance value needs to be adjusted by hand. Next, multiple circuits were used as the connecting weights. Multiple circuits can calculate the products of a two input signals. One is an input signal value. Another input as the connecting value is considered.

In the former study, three-layered neural network analog electronic circuits were designed. The model used multiple circuits by opamp and metal-oxide-semiconductor field-effect transistor (MOSFET) as the connecting weights. The connecting weights vary easily by controlling the input signal. The model has two input units and one-output unit with three layers. After the learning process, the model worked Exclusive OR (EX-OR) logic as the simulation program with integrated circuit emphasis (SPICE) simulation, this is a linear inseparable problem [10, 11].

23.2 Neural Network by Analog Electronic Circuit

23.2.1 Neural Network Using Solid or Variable Resistance Elements

Early analog neural circuits' models were configured of the difference circuits, multiple circuits and solid resistance. It is not easy to change the value of solid resistance for the learning process. In past research, a hardware neural network was constructed using variable resistance. This variable resistance means the connecting weights of network. The network has nine units in the input layer, three units in the middle layer and three units in the output layer. The system was able to recognized simple

patterns by pure analog circuits. However, in the learning process, each resistance value needs to be adjusted by hand.

23.2.2 Neural Network Using Multiple Circuits

In a former study, multiple circuits were utilized as the connecting weights. The connecting weights could be easily changed by controlling the input signal. Figure 23.1 shows the 2-input and 1-output neural circuit. It means the structure of one neuron. There are three input units, two input signals and one threshold value. The input unit calculates the product of two voltages, input signal value and connecting weights. The connecting weights can be easily changed by operating the voltage of the MOSFET gate signal in Fig. 23.1.

23.3 Perceptron Network by Analog Circuits

Next, a learning neural network was constructed. It is a two-input unit and one-output unit basic perceptron model with a feedback circuit. The diagram of this model is indicated in the upper half of Fig. 23.2. In the diagram, multiple circuits are identified as “Mul”. Additional circuits are indicated as “Add” and subtraction circuits as “Sub”. Figure 23.2 also shows the perceptron network of analog electronic

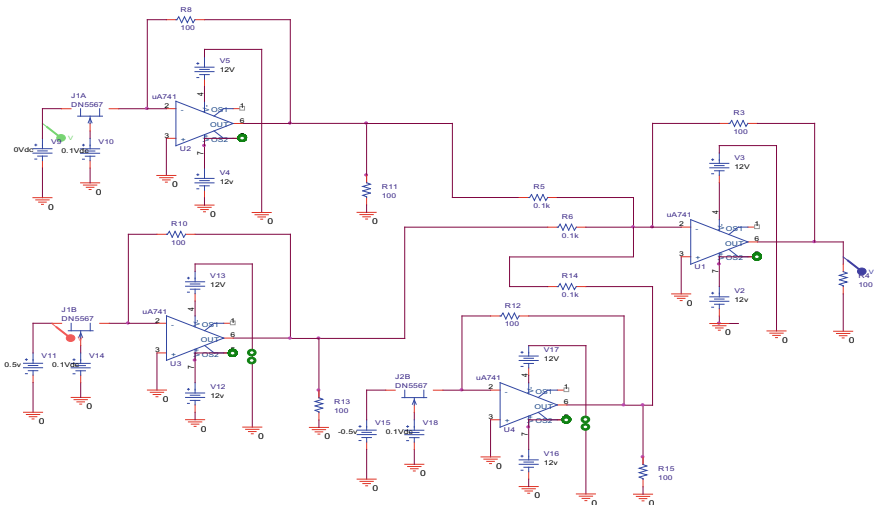


Fig. 23.1 Neuron model by multiple circuits

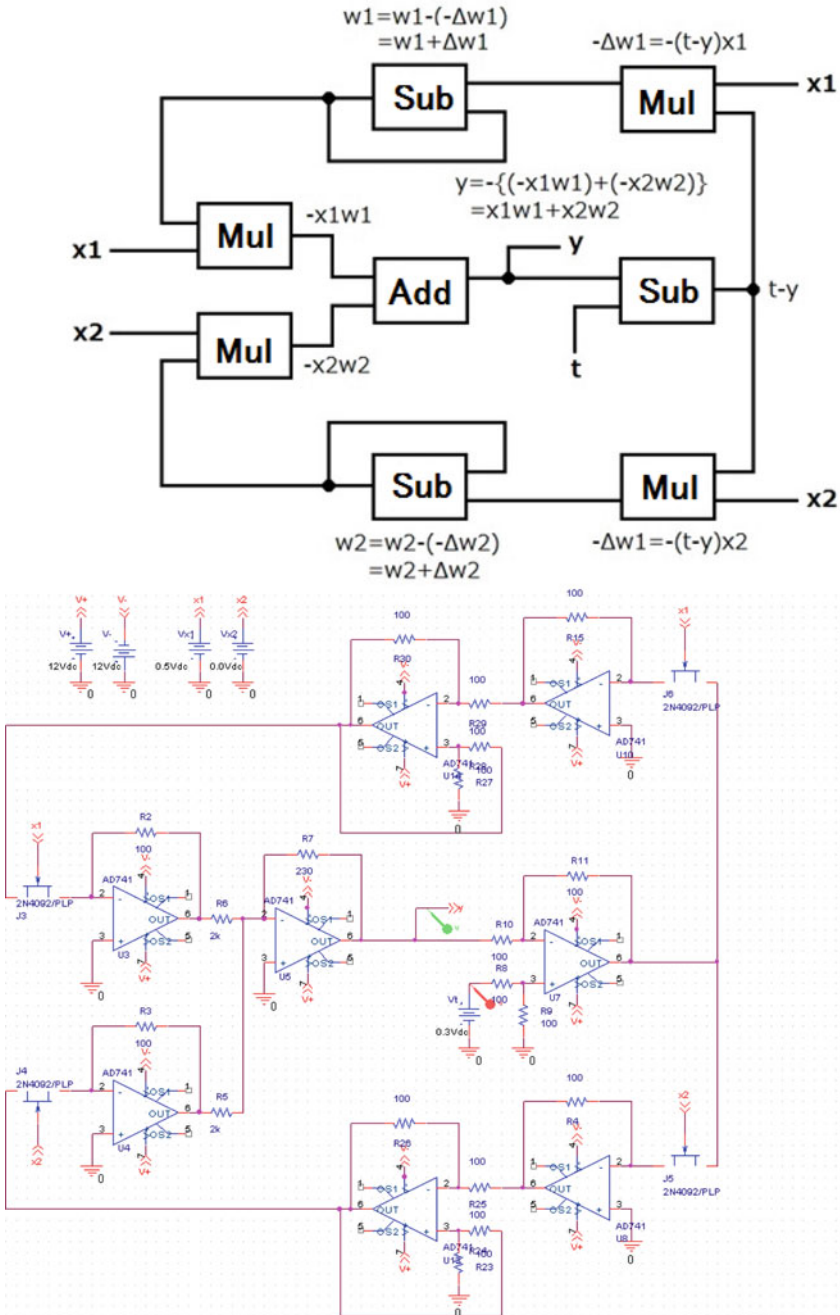


Fig. 23.2 Block diagram and learning circuit of perceptron

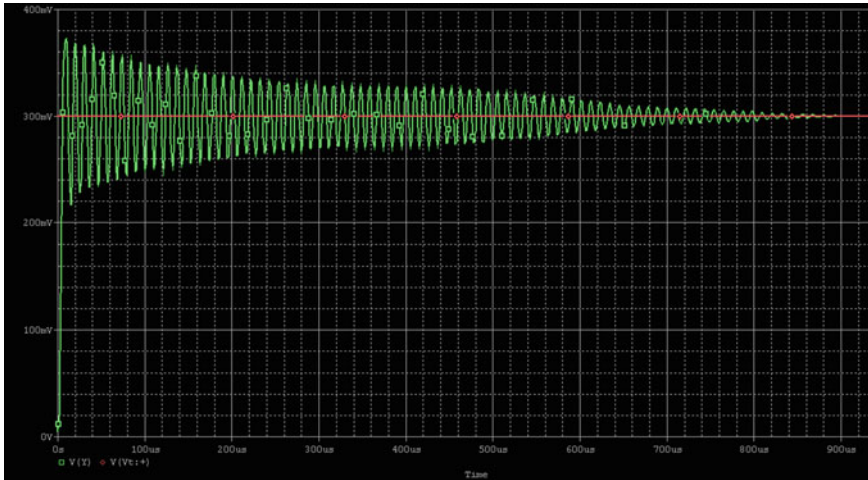


Fig. 23.3 Convergence output of perceptron by SPICE simulation

circuits. The SPICE simulation result is shown in Fig. 23.3 [12]. It is shown that the convergence time is about 900 μ S, when the learning process is finished.

23.4 AC Operation Neural Circuit

In the previous chapter, the hardware learning network is explained. Although, in the situation of setting up a network, there is one problem. The operating range of input and output-voltage level is limited. Moreover, the circuit behavior is sometimes instability because of the multiple circuit features using a semiconductor. It is said “circuit limitations.” One reason is a semiconductor specificity. Not all semiconductors are manufacturing equally. Another reason is the output-voltage limitation of semiconductor element.

The alternative current (AC), not direct current (DC) was used as a transmission signal. Shown is the AC operation of one neuron unit in Fig. 23.4. There are two components in the alternative current. On the alternative currents, current flows with two elements, voltage and frequency. The input signal of the neural circuit is the input voltage of the alternative current. The connecting weights of the neural network are the frequency of the alternative current. The parameter of the circuit is decided as the capacitance and resistance (CR) coefficient. CR circuit has good stability compared to a semiconductor analog circuit using the frequency characteristics of capacitor and resistance. As the result, this circuits outputs the approximately products of voltage and frequency.

Figure 23.5 is the graph of input frequency and output voltage by AC neural circuits. Output voltage is the root mean square (RMS) value. In the network, it is

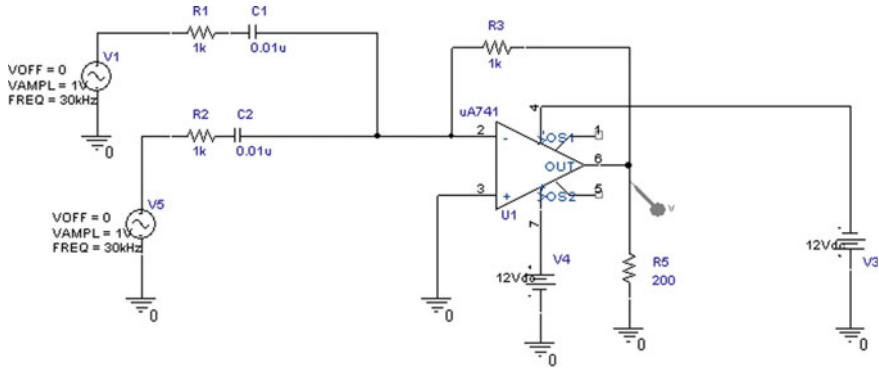


Fig. 23.4 AC operating neural model

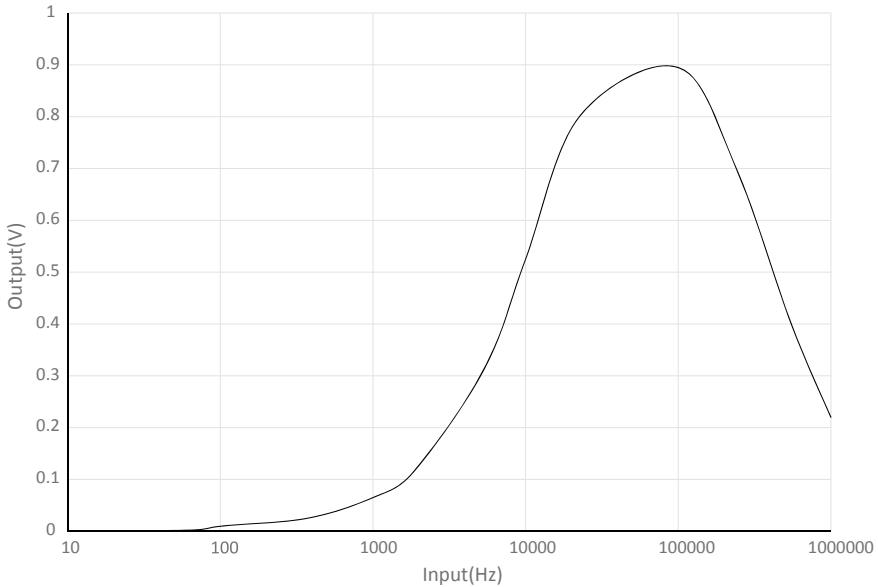
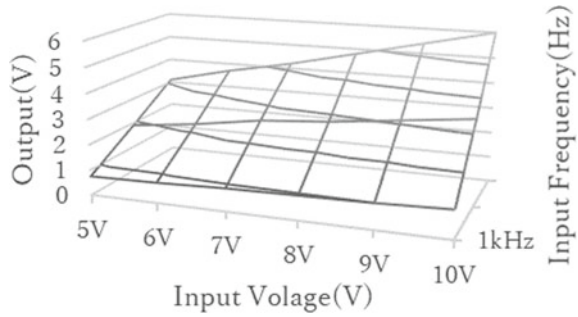


Fig. 23.5 Output AC operation neural circuits

added the two input alternative currents by an additional circuit. Additional circuit outputs a wave modulation. This additional circuits works that the outputs increase satisfactory in the general-purpose frequency range, 3–30 kHz. Figure 23.6 is the output value of the same neural circuit by a two-dimensional graph. The frequency range is from 3 to 30 kHz and the voltage range is 5–10 V. It was confirmed that the output RMS voltage value is also monotonically increasing.

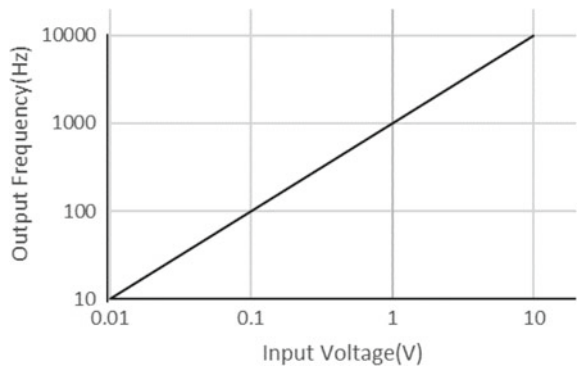
Fig. 23.6 Output behavior of AC operation neural circuit



23.4.1 V-F Converter Circuits

A learning AC operation neural network with a feedback circuit was designed. The error value feedback signal needs to be converted into a connecting weight. Because, these circuits use the frequency data as the connecting weights. The initial input signal is DC current. A voltage-frequency converter circuits was used when generating the connecting weight. V-F converter circuits were used to calculate the frequency data from the error value. The unit generates alternative current as the connecting weight. It is used the backpropagation (BP) learning process using AC feedback circuit. It has to convert from DC voltage to AC current with frequency, after receive the DC current by a rectifier circuit. Figure 23.7 shows V-F converter circuit characteristics, NJM4151, New Japan Radio Co., Ltd. (JRC). The input voltage and output frequency are monotonically increasing in the logarithmic scale.

Fig. 23.7 V-F converter circuit characteristics NJM4151, New JRC



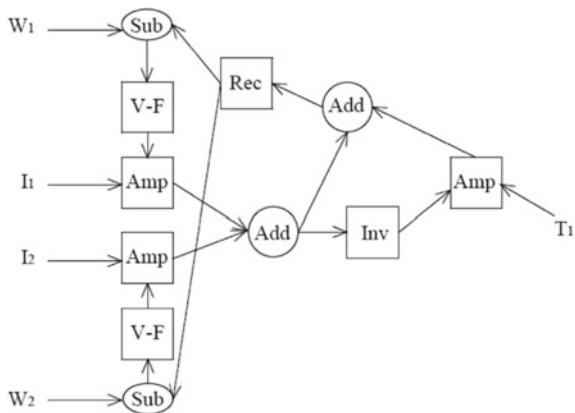
23.4.2 The Behavior of the Neural Circuit by AC Transmission

In the AC operation neural network, first amplifier circuit generates the AC current. Input signal is AC current and the gain of the amplifier is input DC current. Two AC currents after each amplifier circuits are added by an additional circuit. The modulated wave is generated in the output of the additional circuit. The first layer's output of this neural network is modulated wave. In the learning process of the AC neural circuit, the feedback error signal needs to be converted to a connecting weight by AC current. The products difference error signal and input signal means the correction error signal. The difference between the output signal and the teaching signal means the difference error signal [13, 14].

Figure 23.8 shows the diagram of AC operation learning neural network model unit. These circuits are composed of an amplifier, additional circuit, inverter, subtract circuit, rectifier circuit and V-F converter. The initial input signal and connecting weights are both DC current. The V-F converter units generate the frequency from the input DC current. The output of the amplifier circuit is the AC current. Input signal is AC current and the gain of the amplifier is input DC current. It is shown the AC operation learning neural circuits except the V-F converter in Fig. 23.9.

Two AC currents after each amplifier circuits calculate the sum by an additional circuit shown in Fig. 23.10. Each signal is different in voltage and frequency. The modulated wave is generated in the output of the additional circuit. The output of inverse circuit is the phase inversion modulated wave. The input of the second amplifier circuits is phase inversion modulated wave. The teaching signal means the gain of second amplifier circuit. The second adder circuit calculates the sum of the second amplifier circuits output and the first adder circuits output. The input of second amplifier circuits is the output of phase inversion circuit. It means the phase inverted teaching signal. The output of the second adder circuit is the difference of the first adder circuits' output and teaching signal, it is the error value. Thus, the subtract

Fig. 23.8. Diagram of AC operation learning neural network model unit



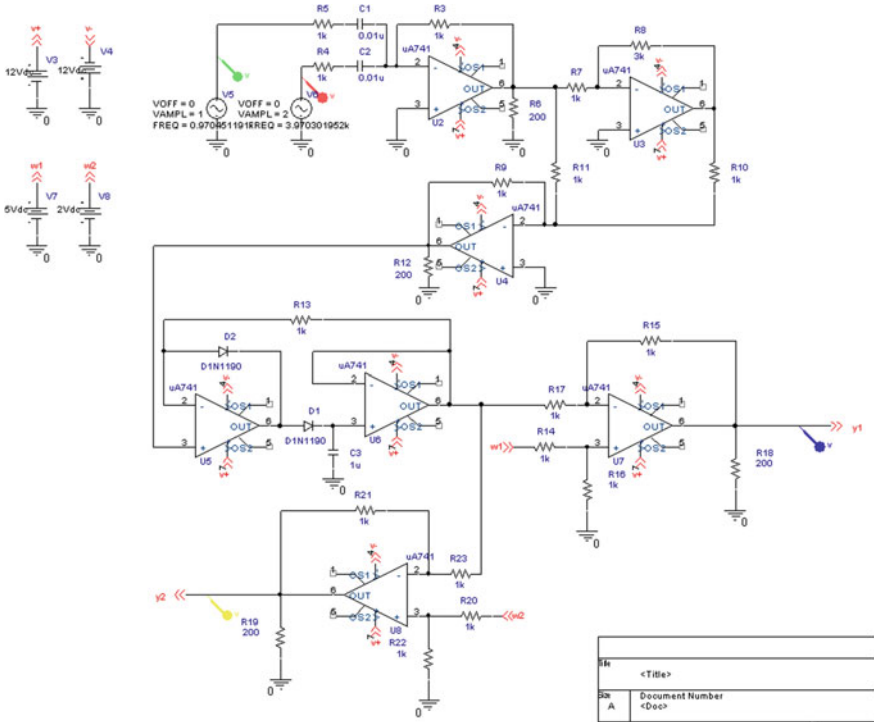
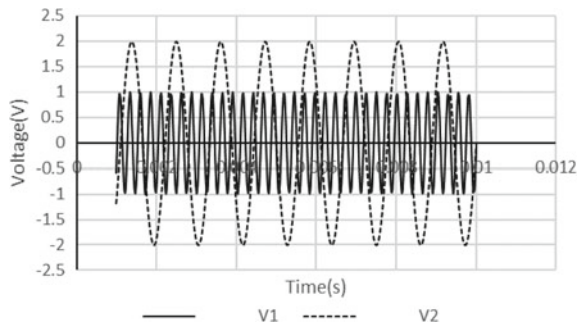


Fig. 23.9 Whole circuit of AC operation learning neural network

Fig. 23.10 Two input signals of AC feedback neural model



circuit does not need to be used to calculate the error value. The output of rectifier circuit is the DC current, converted from the modulated wave of the second adder circuits' output. Figure 23.11 shows the simulation results after rectifier circuits. There is some distortion in the rectifier signal, but it is not a big problem.

The correction quantity of the connection weights is the DC current, output of the rectifier circuits. The subtract circuits calculate the difference of initial connecting weights and correction quantity. The output of subtract circuits is new connecting

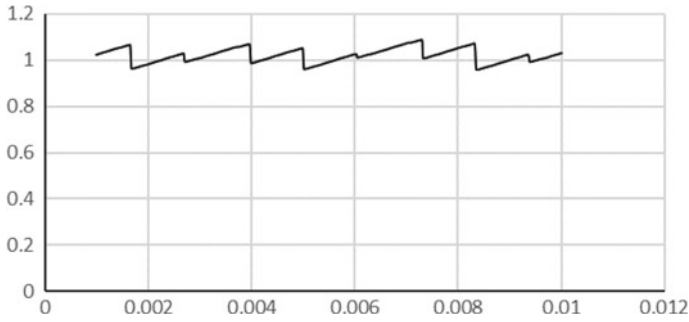


Fig. 23.11 Simulation results after rectifier circuits of AC feedback neural model

weights. Figure 23.12 shows the new connecting weights by an AC feedback neural model. It is shown that the new connecting weight's value is almost constant.

The input of V-F converter circuits is the subtract circuits' output and the output of V-F converter is the AC current by a frequency signal. The input of first amplifier circuits is AC current by a frequency signal and input DC current. At last, first amplifier circuits generated the AC current. The input value means the RMS voltage of AC current. The connecting weights of the neural network are the frequency of the AC current. The series of behaviors means BP learning operation, feedforward and feedback process.

The graph showing the relationship between learning count time and the frequency of output is shown in Fig. 23.13. Frequency f1 and f2 means the each connecting

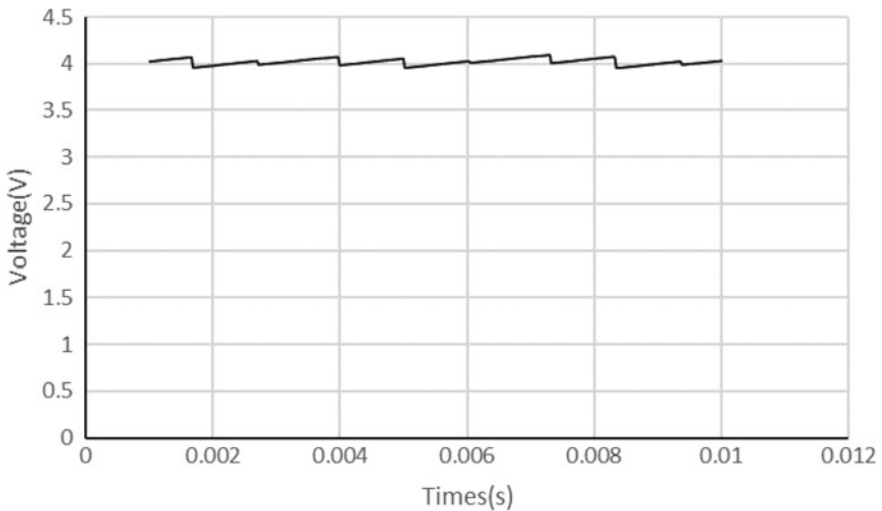


Fig. 23.12 Simulation results of new connecting weights by AC feedback neural model

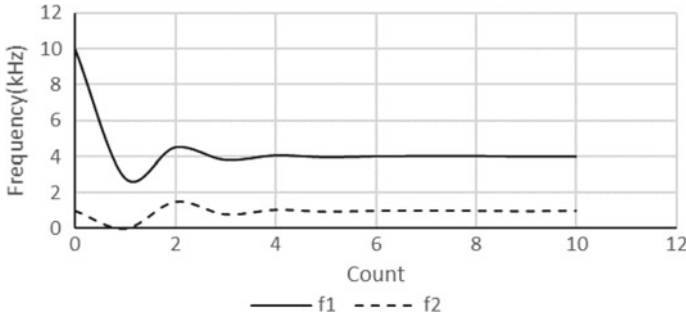


Fig. 23.13 Learning count time and the frequency of output

weights wiring from each input unit. It shows each frequency convergences to constant value approximately. The learning count is only 5 or 6 times; it is quite small compared to other learning model. Another merit is fast learning speed; however, the performance of network is inferior compared to another digital learning network model. But this model refers to biomedical neuron model. Therefore, this proposed method outputs soft and smooth nonlinear signal.

23.5 Conclusion and Future Work

23.5.1 Conclusion

An analog neuron model by multiple circuits was designed using operational amplifier and MOSFET instead of using variable resistance. The operation of the network was confirmed by electronic circuit simulation. Next, one unit of analog neural network was constructed by an AC current operation circuit. The input of first amplifier circuits is AC current by a frequency signal and input DC current. There are two components in the alternative current. The input value means the RMS voltage of AC current. The connecting weights of the neural network are the frequency of the AC current. Two AC currents after each amplifier circuits calculate the sum by an additional circuit.

In the final learning process of the network model, the input of V-F converter circuits is the subtract circuits' DC output and the output of V-F converter is the AC current by a frequency signal. In the SPICE simulation results, the frequency converges to a constant frequency and finishes only several learning processes. In the experimental result, frequency converges to 4 kHz and 1 kHz, respectively. The learning count is only 5 or 6 times; it is quite small compared to other learning model. Another merit is fast learning speed; this model refers to biomedical neuron model. Especially, a deep learning model has been developed very rapidly in the research area of image/sound recognition. Recently, multilayer neural network models, in

particular, a deep learning model have been researched very sprightly. The performance has been extremely improved in the specialty of image/sound recognition. It is awaited high performance artificial intelligence model using the learning system using analog circuit in the near future.

23.5.2 Future Scope Using Deep Learning Model

Deep learning model is one kind of machine learning model. The performance of the recognition is getting better more and more. This model is used for practical purposes in the field of image detection, video/sound recognition. This model will be developed in the field of advanced technology, self-driving, robotics and artificial intelligence. In the original BP, learning neural network is three-layer structure. However, the structure of in the general deep learning model, there are nine layers. And there are also three layered sub-networks, like a convolution network [15]. Furthermore, learning algorithm in the field of deep learning model uses a stacked auto-encoder. This algorithm can detect the feature data and abstract expression data from the input image using large quantity learning data. The proposed AC operation circuits prompt the possibility for flexible structure neural network like a deep learning model [16].

If AC operation learning network system developed in the direction of multi-layered network model like a deep learning, many applications will be expected. It is one of the kinds of soft computing. The output data is soft and smooth, difficult to realize in the digital processing. In the field of image recognition and speech recognition, the proposed AC operation model suggests the chance of making the electrical circuit model of deep learning [17]. The model will develop the artificial intelligence unit under the environment of automated operation with tough and beneficial for fault tolerance network. Increase the number of units, large-scale system development and creating integrated circuit are future problems.

References

1. Mead, C.: Analog VLSI and Neural Systems. Addison Wesley Publishing Company, Inc. (1989)
2. Chong, C.P., Salama, C.A.T., Smith, K.C.: Image-motion detection using analog VLSI. *IEEE J. Solid-State Circuits* **27**(1), 93–96 (1992)
3. Lu, Z., Shi, B.E.: Subpixel resolution binocular visual tracking using analog VLSI vision sensors. *IEEE Trans. Circ. Syst. II Analog Digital Sig. Process.* **47**(12), 1468–1475 (2000)
4. Saito, T., Inamura, H.: Analysis of a simple A/D converter with a trapping window. *IEEE Int. Symp. Circ. Syst.*, 1293–1305 (2003)
5. Luthon, F., Dragomirescu, D.: A cellular analog network for MRF-based video motion detection. *IEEE Trans. Circ. Syst. I Fundam. Theory Appl.* **46**(2), 281–293 (1999)
6. Yamada, H., Miyashita, T., Ohtani, M., Yonezu, H.: An analog MOS circuit in-spired by an inner retina for producing signals of moving edges. Technical report of IEICE, NC99-112, 149–155 (2000)

7. Okuda, T., Doki, S., Ishida, M.: Realization of back propagation learning for pulsed neural networks based on delta-sigma modulation and its hardware implementation. *ICICE Trans. J88-D-II-4*, 778–788 (2005)
8. Kawaguchi, M., Jimbo, T., Umeno, M.: Motion detecting artificial retina model by two-dimensional multi-layered analog electronic circuits. *IEICE Trans. E86-A-2*, 387–395 (2003)
9. Kawaguchi, M., Jimbo, T., Umeno, M.: Analog VLSI layout design of advanced image processing for artificial vision model. In: *IEEE International Symposium on Industrial Electronics, ISIE2005 Proceeding*, vol. 3, pp. 1239–1244 (2005)
10. Kawaguchi, M., Jimbo, T., Umeno, M.: Analog VLSI layout design and the circuit board manufacturing of advanced image processing for artificial vision model. In: *KES2008, Part II, LNAI*, vol. 5178, pp. 895–902 (2008)
11. Kawaguchi, M., Jimbo, T., Umeno, M.: Dynamic learning of neural network by analog electronic circuits. In: *Intelligent System Symposium, FAN2010, S3-4-3* (2010)
12. Kawaguchi, M., Jimbo T., Ishii, N.: Analog learning neural network using multiple and sample hold circuits. In: *IIAI/ACIS International Symposiums on Innovative E-Service and Information Systems, IEIS 2012*, pp. 243–246 (2012)
13. Kawaguchi, M., Ishii, N., Umeno, M.: Analog Learning neural circuit with switched capacitor and the design of deep learning model. *Comput. Sci. Intell. Appl. Inf. Stud. Comput. Intell.* **726**, 93–107 (2017)
14. Kawaguchi, M., Ishii, N., Umeno, M.: Analog neural circuit by AC operation and the design of deep learning model. In: *DEStech Transactions on Computer Science and Engineering, 3rd International Conference on Artificial Intelligence and Industrial Engineering*, pp. 228–233 (2017)
15. Yoshua, B., Aaron, C., Courville, P.: Vincent: representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(8), 1798–1828 (2013)
16. Kawaguchi, M., Ishii, N., Umeno, M.: Analog neural circuit with switched capacitor and design of deep learning model. In: *3rd International Conference on Applied Computing and Information Technology and 2nd International Conference on Computational Science and Intelligence, ACIT-CSI*, pp. 322–327 (2015)
17. Kawaguchi, M., Ishii, N., Umeno, M.: Analog learning neural circuit with switched capacitor and the design of deep learning model. *Comput. Sci. Intell. Appl. Inf.* **726**, 93–107 (2018)

Chapter 24

An Inequality for Linear Canonical Transform



Mawardi Bahri and Ryuichi Ashino

Abstract In our previous work, we established some basic properties of the linear canonical transform and obtained alternative form of convolution and correlation theorems. In this paper, we study essential properties of the linear canonical transform (LCT). The properties are modifications of the classical Fourier transform properties. They are very need in applying LCT in signal processing. In addition, we formulate an inequality associated with the LCT, which is different from the uncertainty principle in literature.

AMS Subject Classification 11R52 · 42A38 · 15A66 · 83A05 · 35L05

24.1 Introduction

In recent years, the linear canonical transformation (LCT) has been widely discussed in a number of papers (see, e.g., [1–7]). It is a three-parameter class of linear integral transformations which is effectively used in many field of applied mathematics, optics, digital information processing, and so on. The LCT can be looked as a generalization of the classical Fourier transform [8, 9], so most of the properties of the classical Fourier transform (FT) are modified in the LCT domain. In literature, the LCT is also known as the affine Fourier transform, Collins formula, the ABCD transform and Moshinsky-queue-transform [6, 10]. Because of the benefits of LCT, there are many general transformations, which are built using LCT, that is, by merging kernel transformation with LCT kernel (see, e.g., [1, 11–19]).

M. Bahri (✉)
Hasanuddin University, Makassar 90245, Indonesia
e-mail: mawardibahri@gmail.com

R. Ashino
Osaka Kyoiku University, Osaka 582-8582, Japan
e-mail: ashino@cc.osaka-kyoiku.ac.jp

In this paper, we focus on investigating some properties of the LCT, which are missing in the literature. Based on the properties we build an inequality related to the LCT. The inequality gives information about a signal and its LCT. It is shown that the inequality only holds for the specific matrix parameters of the LCT.

The rest of the paper is organizes as follows. In Sect. 24.2, we investigate several properties of the LCT, which are generalizations of the corresponding properties of the classical Fourier transform. Sect. 24.3 is devoted to establish the inequality for the LCT which describes the relation between the signal and its LCT. In Sect. 24.4, we conclude the article.

24.2 Essential Properties of Linear Canonical Transform

In this part, we investigate some essential properties of the linear canonical transform (LCT). They can be regarded as an extension of the classical Fourier transform properties.

For $1 \leq p < +\infty$, the space $L^p(\mathbb{R})$ consists of complex-valued functions $f \rightarrow \mathbb{C}$ such that $\int_{\mathbb{R}} |f(x)|^p dx < \infty$. The L^p -norm $\|f\|_p$ of $f \in L^p(\mathbb{R})$ is defined by

$$\|f\|_p = \left(\int_{\mathbb{R}} |f(x)|^p dx \right)^{1/p} .$$

The space $L^p(\mathbb{R})$ is complete with respected to the L^p -norm, that is, $L^p(\mathbb{R})$ is a Banach space.

In the present section, we investigate several properties of the linear canonical transform (LCT) which missed in the literature. Let us now introduce the LCT definition as follows.

Definition 1 (*LCT Definition*) Let $h \in L^1(\mathbb{R})$ and

$$M = (m, n, p, q) = \begin{bmatrix} m & n \\ p & q \end{bmatrix} \in \mathbb{R}^{2 \times 2}$$

be a matrix parameter such that $\det(M) = mq - np = 1$.

For $n \neq 0$, the linear canonical transform of f is defined by

$$L_M\{h\}(\omega) = \int_{\mathbb{R}} h(x)K_M(\omega, x) dx, \tag{24.1}$$

where $K_M(x, \omega)$ is so-called kernel of the LCT given by

$$K_M(x, \omega) = \frac{1}{\sqrt{2\pi n}} e^{\frac{i}{2}(\frac{m}{n}x^2 - \frac{2}{n}x\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2})}.$$

When $n = 0$, the LCT is reduced to

$$L_M\{h\}(\omega) = \sqrt{q} e^{\frac{iq}{2}\omega^2} h(q\omega).$$

Note that the LCT kernel mentioned above has the following important property

$$K_{M^{-1}}(\omega, x) = \overline{K_M(x, \omega)} = \frac{1}{\sqrt{2\pi n}} e^{-\frac{i}{2}(\frac{m}{n}x^2 - \frac{2}{n}x\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2})}.$$

Definition 1 implies that the LCT of a signal is essentially a chirp multiplication when $n = 0$. Therefore, in this paper, we always assume that $n \neq 0$.

As a special case, when $M = (m, n, p, q) = (0, 1, -1, 0)$, the LCT definition (24.1) reduces to the Fourier transform (FT) definition. It is known that given $L_M\{h\}$ we can obtain h by the inverse formula of the LCT given by

$$L_M^{-1}[L_M\{h\}](x) = h(x) = \int_{\mathbb{R}} L_M\{h\}(\omega) K_M^{-1}(\omega, x) d\omega, \quad n \neq 0. \quad (24.2)$$

$$h(x) = \int_{\mathbb{R}} L_M\{h\}(\omega) \frac{1}{\sqrt{2\pi n}} e^{-\frac{i}{2}(\frac{m}{n}x^2 - \frac{2}{n}x\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2})} d\omega, \quad n \neq 0. \quad (24.3)$$

We can obtain the LCT of $h \in L^1(\mathbb{R})$ via associated FT by

$$L_M\{h\}(\omega) = \frac{e^{-i\frac{\pi}{4}}}{\sqrt{n}} e^{\frac{iq}{2n}\omega^2} \mathcal{F}\{e^{\frac{im}{2n}x^2} h(x)\} \left(\frac{\omega}{n}\right), \quad (24.4)$$

where $\mathcal{F}\{h\}(\omega) = \hat{h}(\omega)$ is the Fourier transform of $h \in L^1(\mathbb{R})$ defined by

$$\mathcal{F}\{h\}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} h(x) e^{-i\omega x} dx. \quad (24.5)$$

Let $r(x) = \frac{e^{-i\frac{\pi}{4}}}{\sqrt{n}} e^{\frac{im}{2n}x^2} h(x)$. Then, (24.4) can be written in the following form:

$$e^{-\frac{iq}{2n}\omega^2} L_M\{h\}(\omega) = \mathcal{F}\{r\} \left(\frac{\omega}{n}\right). \quad (24.6)$$

The space $L^2(\mathbb{R})$ is a Hilbert space with the inner product

$$(h, g) = \int_{\mathbb{R}} h(x)\overline{g(x)} \, dx. \tag{24.7}$$

As a consequence of the inner product (24.7), we easily obtain Parseval’s formula as

$$(h, g) = (L_M\{h\}, L_M\{g\}). \tag{24.8}$$

Some important properties of the LCT are demonstrated in the following theorems. We shall see that most properties are corresponding to generalizations of the FT properties (compare to [8, 9]).

Theorem 24.1 *Let $f \in L^1(\mathbb{R})$. Then, we have*

$$\overline{L_M\{f\}(\omega)} = L_{\hat{M}}\{f\}(\omega), \quad \hat{M} = \begin{bmatrix} m & -n \\ p & q \end{bmatrix}. \tag{24.9}$$

Proof Indeed, we have

$$\begin{aligned} \overline{L_M\{f\}(\omega)} &= \int_{\mathbb{R}} \overline{f(x) \frac{1}{\sqrt{2\pi n}} e^{\frac{i}{2}(\frac{m}{n}\omega^2 - \frac{2}{n}\omega x + \frac{q}{n}x^2 - \frac{\pi}{2})}} \, dx \\ &= \int_{\mathbb{R}} \bar{f}(x) \frac{1}{\sqrt{2\pi n}} e^{-\frac{i}{2}(\frac{m}{n}\omega^2 - \frac{2}{n}\omega x + \frac{q}{n}x^2 - \frac{\pi}{2})} \, dx \\ &= \int_{\mathbb{R}} f(x) \frac{1}{\sqrt{-2\pi n}} e^{\frac{i}{2}(\frac{m}{-n}\omega^2 + \frac{2}{n}\omega x - \frac{q}{n}x^2 - \frac{\pi}{2})} \, dx \\ &= L_{\hat{M}}\{f\}(\omega), \end{aligned} \tag{24.10}$$

which proves the theorem. □

Theorem 24.2 *For $f, g \in L^1(\mathbb{R})$, it holds that*

$$\overline{(L_M\{h\}, \bar{g})} = (\bar{h}, L_M\{\bar{g}\}). \tag{24.11}$$

Proof Simple computations show that

$$\begin{aligned}
 (\overline{L_M\{h\}}, \bar{g}) &= \int_{\mathbb{R}} \overline{L_M\{h\}(t)} g(t) dt \\
 &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x) \frac{1}{\sqrt{2\pi n}} e^{\frac{i}{2}(\frac{m}{n}x^2 - \frac{2}{n}xt + \frac{q}{n}t^2 - \frac{\pi}{2})} dx \right) g(t) dt \\
 &= \int_{\mathbb{R}} \int_{\mathbb{R}} \bar{h}(x) \frac{1}{\sqrt{2\pi n}} e^{-\frac{i}{2}(\frac{m}{n}x^2 - \frac{2}{n}xt + \frac{q}{n}t^2 - \frac{\pi}{2})} g(t) dt dx \\
 &= \int_{\mathbb{R}} \bar{h}(x) \left(\int_{\mathbb{R}} \bar{g}(t) \frac{1}{\sqrt{2\pi n}} e^{\frac{i}{2}(\frac{m}{n}x^2 - \frac{2}{n}xt + \frac{q}{n}t^2 - \frac{\pi}{2})} dt \right) dx \\
 &= \int_{\mathbb{R}} \bar{h}(x) \overline{L_M\{\bar{g}\}(x)} dx \\
 &= (\bar{h}, L_M\{\bar{g}\}),
 \end{aligned} \tag{24.12}$$

which completes the proof. \square

Theorem 24.3 For $h, g \in L^1(\mathbb{R})$, we have

$$(h, \overline{L_M\{g\}}) = (L_M\{h\}, \bar{g}). \tag{24.13}$$

Proof By the LCT definition, we obtain

$$\begin{aligned}
 (h, \overline{L_M\{g\}}) &= \int_{\mathbb{R}} h(t) L_M\{g\}(t) dt \\
 &= \int_{\mathbb{R}} h(t) \left(\int_{\mathbb{R}} g(x) \frac{1}{\sqrt{2\pi n}} e^{\frac{i}{2}(\frac{m}{n}x^2 - \frac{2}{n}xt + \frac{q}{n}t^2 - \frac{\pi}{2})} dx \right) dt \\
 &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(t) \frac{1}{\sqrt{2\pi n}} e^{\frac{i}{2}(\frac{m}{n}x^2 - \frac{2}{n}xt + \frac{q}{n}t^2 - \frac{\pi}{2})} dt \right) g(x) dx \\
 &= \int_{\mathbb{R}} L_M\{h\}(x) g(x) dx \\
 &= (L_M\{h\}, \bar{g}),
 \end{aligned} \tag{24.14}$$

which completes the proof. \square

Theorem 24.4 *If $h \in L^1(\mathbb{R})$ and $g = \overline{L_M\{h\}}$, then $h = \overline{L_M\{g\}}$.*

Proof Applying (24.13), we have

$$(h, \overline{L_M\{g\}}) = (L_M\{h\}, \bar{g}) = (L_M\{h\}, L_M\{h\}) = \|L_M\{h\}\|_2^2 = \|h\|_2^2. \quad (24.15)$$

In view of (24.11), we have

$$(\overline{h}, L_M\{g\}) = \overline{(L_M\{h\}, g)} = \overline{(L_M\{h\}, L_M\{h\})} = \|L_M\{h\}\|_2^2. \quad (24.16)$$

With the help of Parseval's formula, we obtain

$$\|L_M\{g\}\|_2^2 = \|g\|_2^2 = \|L_M\{h\}\|_2^2 = \|h\|_2^2. \quad (24.17)$$

By the hypothesis of the theorem, we have

$$\begin{aligned} \|h - \overline{L_M\{g\}}\|_2^2 &= (h - \overline{L_M\{g\}}, h - \overline{L_M\{g\}}) \\ &= \|h\|_2^2 - (h, \overline{L_M\{g\}}) - (\overline{h}, L_M\{g\}) + \|L_M\{g\}\|_2^2 \\ &= \|h\|_2^2 - \|L_M\{h\}\|_2^2 - \|h\|_2^2 + \|L_M\{g\}\|_2^2 \\ &= 0. \end{aligned} \quad (24.18)$$

This proves the theorem. \square

Definition 2 Let L_M be the linear canonical transform. The transform L_M^* given by

$$(L_M\{h\}, g) = (h, L_M^*\{g\}). \quad (24.19)$$

is called adjoint of L_M .

The definition mentioned above leads to the relationship between adjoint of LCT and its inverse.

Theorem 24.5 *Let $h, g \in L^2(\mathbb{R})$. The adjoint of the LCT equals to its inverse, that is,*

$$(L_M\{h\}, g) = (h, L_M^{-1}\{g\}). \quad (24.20)$$

Proof It follows from (24.7) that

$$\begin{aligned}
 (L_M\{h\}, g) &= \int_{\mathbb{R}} L_M\{h\}(\omega) \overline{g(\omega)} \, d\omega \\
 &= \int_{\mathbb{R}} \int_{\mathbb{R}} h(x) \frac{1}{\sqrt{2\pi n}} e^{i\left(\frac{m}{n}x^2 - \frac{2}{n}x\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2}\right)} \overline{g(\omega)} \, d\omega \, dx \\
 &= \int_{\mathbb{R}} h(x) \left(\int_{\mathbb{R}} \overline{g(\omega) \frac{1}{\sqrt{2\pi n}} e^{-i\left(\frac{m}{n}x^2 - \frac{2}{n}x\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2}\right)}} \, d\omega \right) dx \\
 &= \int_{\mathbb{R}} h(x) \overline{L_M^{-1}\{g\}} \, dx \\
 &= (h, L_M^{-1}\{g\}).
 \end{aligned} \tag{24.21}$$

Combining (24.19) with (24.21) completes the proof. \square

Theorem 24.6 (Parseval's formula for L_M^*) *If $h, g \in L^1(\mathbb{R})$, then, we have*

$$(L_M^*\{h\}, L_M^*\{g\}) = (h, g). \tag{24.22}$$

Proof Applying (24.2) and (24.20), we have

$$\begin{aligned}
 (L_M^*\{h\}, L_M^*\{g\}) &= (L_M^{-1}\{h\}, L_M^{-1}\{g\}) \\
 &= \int_{\mathbb{R}} L_M^{-1}\{h\}(x) \overline{L_M^{-1}\{g\}(x)} \, dx \\
 &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} L_M\{h\}(t) \frac{1}{\sqrt{2\pi n}} e^{-i\left(\frac{m}{n}x^2 - \frac{2}{n}xt + \frac{q}{n}t^2 - \frac{\pi}{2}\right)} \, dt \right) \overline{L_M^{-1}\{g\}(x)} \, dx \\
 &= \int_{\mathbb{R}} L_M\{h\}(t) \left(\int_{\mathbb{R}} \overline{g(x) \frac{1}{\sqrt{2\pi n}} e^{i\left(\frac{m}{n}x^2 - \frac{2}{n}xt + \frac{q}{n}t^2 - \frac{\pi}{2}\right)}} \, dx \right) dt \\
 &= \int_{\mathbb{R}} L_M\{h\}(t) \left(\int_{\mathbb{R}} g(x) \frac{1}{\sqrt{2\pi n}} e^{i\left(\frac{m}{n}x^2 - \frac{2}{n}xt + \frac{q}{n}t^2 - \frac{\pi}{2}\right)} \, dx \right) dt \\
 &= \int_{\mathbb{R}} L_M\{h\}(t) \overline{L_M\{g\}(t)} \, dt \\
 &= (L_M\{h\}, L_M\{g\}) \\
 &= (h, g),
 \end{aligned} \tag{24.23}$$

which completes the proof. \square

The following theorem is important because it is a useful tool for investigating the inequality related to the LCT.

Theorem 24.7 *Let $f(t) \in L^1(\mathbb{R})$. Suppose that $L_M\{\frac{d^l}{dt^l}f(t)\}(\omega)$ exists. Then, we have*

$$L_M\left\{\frac{d^l}{dt^l}f(t)\right\}(\omega) = \left(-p\omega i + m\frac{d}{d\omega}\right)^l L_M\{f(t)\}(\omega), \quad l \in \mathbb{N}. \tag{24.24}$$

Proof We will show (24.24) by induction. When $l = 0$, it is trivial. For $l = 1$, we easily get

$$\begin{aligned} L_M\left\{\frac{d}{dt}f(t)\right\}(\omega) &= \int_{\mathbb{R}} \left(\frac{d}{dt}f(t)\right) K_M(t, \omega) dt \\ &= \frac{1}{\sqrt{2\pi n}} \int_{\mathbb{R}} e^{i\frac{1}{2}\left(\frac{m}{n}t^2 - \frac{2}{n}t\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2}\right)} \left(\frac{d}{dt}f(t)\right) dt. \end{aligned}$$

By the integration by parts, we get

$$\begin{aligned} L_M\left\{\frac{d}{dt}f(t)\right\}(\omega) &= \frac{1}{\sqrt{2\pi n}} \left[e^{i\frac{1}{2}\left(\frac{m}{n}t^2 - \frac{2}{n}t\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2}\right)} f(t) \Big|_{-\infty}^{\infty} \right. \\ &\quad \left. - \int_{\mathbb{R}} f(t) e^{i\frac{1}{2}\left(\frac{m}{n}t^2 - \frac{2}{n}t\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2}\right)} i \left(\frac{m}{n}t - \frac{\omega}{n}\right) dt \right] \\ &= \frac{1}{\sqrt{2\pi n}} \left[0 - \int_{\mathbb{R}} f(t) e^{i\frac{1}{2}\left(\frac{m}{n}t^2 - \frac{2}{n}t\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2}\right)} i \left(\frac{m}{n}t - \frac{\omega}{n}\right) dt \right] \\ &= -\frac{1}{\sqrt{2\pi n}} \left[\int_{\mathbb{R}} f(t) e^{i\frac{1}{2}\left(\frac{m}{n}t^2 - \frac{2}{n}t\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2}\right)} i \left(\frac{m}{n}t - \frac{mq - np}{n}\omega\right) dt \right] \\ &= -\frac{1}{\sqrt{2\pi n}} \left[\int_{\mathbb{R}} f(t) e^{i\frac{1}{2}\left(\frac{m}{n}t^2 - \frac{2}{n}t\omega + \frac{q}{n}\omega^2\right)} i \left(\frac{m}{n}t - \frac{mq}{n}\omega + p\omega\right) dt \right] \\ &= \frac{1}{\sqrt{2\pi n}} \left[\int_{\mathbb{R}} f(t) e^{i\frac{1}{2}\left(\frac{m}{n}t^2 - \frac{2}{n}t\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2}\right)} i \left(-\frac{m}{n}t + \frac{mq}{n}\omega - p\omega\right) dt \right] \\ &= \frac{1}{\sqrt{2\pi n}} \left[\int_{\mathbb{R}} f(t) e^{i\frac{1}{2}\left(\frac{m}{n}t^2 - \frac{2}{n}t\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2}\right)} i \left(m\left(-\frac{1}{n}t + \frac{q}{n}\omega\right) - p\omega\right) dt \right] \\ &= \frac{1}{\sqrt{2\pi n}} \int_{\mathbb{R}} f(t) e^{i\frac{1}{2}\left(\frac{m}{n}t^2 - \frac{2}{n}t\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2}\right)} im \left(-\frac{1}{n}t + \frac{q}{n}\omega\right) dt \\ &\quad + \frac{1}{\sqrt{2\pi n}} \int_{\mathbb{R}} f(t) e^{i\frac{1}{2}\left(\frac{m}{n}t^2 - \frac{2}{n}t\omega + \frac{q}{n}\omega^2 - \frac{\pi}{2}\right)} i(-p\omega) dt. \end{aligned}$$

Hence, we get

$$\begin{aligned}
L_M \left\{ \frac{d}{dt} f(t) \right\}(\omega) &= m \frac{1}{\sqrt{2\pi n}} \int_{\mathbb{R}} f(t) \left(\frac{d}{d\omega} e^{\frac{i}{2} \left(\frac{m}{n} t^2 - \frac{2}{n} t\omega + \frac{q}{n} \omega^2 - \frac{\pi}{2} \right)} \right) dt \\
&\quad - p\omega i \frac{1}{\sqrt{2\pi n}} \int_{\mathbb{R}} f(t) e^{\frac{i}{2} \left(\frac{m}{n} t^2 - \frac{2}{n} t\omega + \frac{q}{n} \omega^2 - \frac{\pi}{2} \right)} dt \\
&= m \frac{d}{d\omega} \left(\frac{1}{\sqrt{2\pi n}} \int_{\mathbb{R}} f(t) e^{\frac{i}{2} \left(\frac{m}{n} t^2 - \frac{2}{n} t\omega + \frac{q}{n} \omega^2 - \frac{\pi}{2} \right)} dt \right) \\
&\quad - p\omega i \frac{1}{\sqrt{2\pi n}} \int_{\mathbb{R}} f(t) e^{\frac{i}{2} \left(\frac{m}{n} t^2 - \frac{2}{n} t\omega + \frac{q}{n} \omega^2 - \frac{\pi}{2} \right)} dt \\
&= \left(-p\omega i + m \frac{d}{d\omega} \right) \left(\frac{1}{\sqrt{2\pi n}} \int_{\mathbb{R}} f(t) e^{\frac{i}{2} \left(\frac{m}{n} t^2 - \frac{2}{n} t\omega + \frac{q}{n} \omega^2 - \frac{\pi}{2} \right)} dt \right) \\
&= \left(-p\omega i + m \frac{d}{d\omega} \right) L_M \{f(t)\}(\omega).
\end{aligned}$$

It means that

$$L_M \left\{ \frac{d}{dt} f(t) \right\}(\omega) = \left(-p\omega i + m \frac{d}{d\omega} \right) L_M \{f(t)\}(\omega). \quad (24.25)$$

Assume that it is true for $l-1$, that is,

$$L_M \left\{ \frac{d^{l-1}}{dt^{l-1}} f(t) \right\}(\omega) = \left(-p\omega i + m \frac{d}{d\omega} \right)^{l-1} L_M \{f(t)\}(\omega). \quad (24.26)$$

It follows that

$$\begin{aligned}
L_M \left\{ \frac{d^l}{dt^l} f(t) \right\}(\omega) &= L_M \left\{ \frac{d}{d\omega} \left(\frac{d^{l-1}}{dt^{l-1}} f(t) \right) \right\}(\omega) \\
&= \left(-p\omega i + m \frac{d}{d\omega} \right) L_M \left\{ \frac{d^{l-1}}{dt^{l-1}} f(t) \right\}(\omega) \\
&= \left(-p\omega i + m \frac{d}{d\omega} \right) \left(-p\omega i + m \frac{d}{d\omega} \right)^{l-1} L_M \{f(t)\}(\omega) \\
&= \left(-p\omega i + m \frac{d}{d\omega} \right)^l L_M \{f(t)\}(\omega).
\end{aligned}$$

The proof is complete. \square

24.3 Inequality for LCT

As we know, the uncertainty principle is one of the fundamental results in the LCT. It describes how a signal relates to its LCT. In the following, we introduce an uncertainty associated with the LCT.

Theorem 24.8 (LCT uncertainty principle [5, 20]) *Let $f \in L^2(\mathbb{R})$ be a complex signal. If $L_M\{f\} \in L^2(\mathbb{R})$, then the following inequality holds*

$$\int_{\mathbb{R}} x^2 |f(x)|^2 dx \int_{\mathbb{R}} \omega^2 |L_M\{f\}(\omega)|^2 d\omega \geq \frac{n^2}{4} \left(\int_{\mathbb{R}} |f(x)|^2 dx \right)^2. \tag{24.27}$$

The other form of the uncertainty is described below. We shall see that the inequality only holds for the specific matrix parameters of the LCT.

Theorem 24.9 *Assume that $f \in C^1(\mathbb{R}) \cap L^2(\mathbb{R})$, $f' \in L^2(\mathbb{R})$, and that $L_{\check{M}}\{f\}(\omega)$, $\omega L_{\check{M}}\{f\}(\omega) \in L^2(\mathbb{R})$. Then, the following inequality holds:*

$$\int_{\mathbb{R}} x^2 |f(x)|^2 dx \int_{\mathbb{R}} \omega^2 |L_{\check{M}}\{f\}(\omega)|^2 d\omega \geq \frac{\|f\|_2^4}{4p^2}, \tag{24.28}$$

where $\check{M} = \begin{bmatrix} 0 & n \\ p & q \end{bmatrix}$.

Proof Simple computation shows that

$$\frac{d}{dx} |f(x)|^2 = 2\text{Re}f(x)\overline{f'(x)}. \tag{24.29}$$

The right side of (24.29) is equal to

$$2\text{Re} \int_{\mathbb{R}} f(x)\overline{f'(x)} dx = - \int_{\mathbb{R}} |f(x)|^2 dx.$$

By Schwarz’s inequality, we have

$$\begin{aligned} \left(\int_{\mathbb{R}} |f(x)|^2 dx \right)^2 &= \left(-2\text{Re} \int_{\mathbb{R}} xf(x)\overline{f'(x)} dx \right)^2 \\ &\leq 4 \left(\left| \int_{\mathbb{R}} xf(x)\overline{f'(x)} dx \right| \right)^2 \\ &\leq 4 \int_{\mathbb{R}} x^2 |f(x)|^2 dx \int_{\mathbb{R}} |f'(x)|^2 dx. \end{aligned}$$

Applying Parseval's formula (24.8) and Theorem 24.7, we have

$$\begin{aligned} \left(\int_{\mathbb{R}} |f(x)|^2 dx \right)^2 &\leq 4 \int_{\mathbb{R}} x^2 |f(x)|^2 dx \int_{\mathbb{R}} |L_M \{f'\}(\omega)|^2 d\omega \\ &= 4 \int_{\mathbb{R}} x^2 |f(x)|^2 dx \int_{\mathbb{R}} \left| \left(-p\omega i + m \frac{d}{d\omega} \right) L_M \{f\}(\omega) \right|^2 d\omega. \end{aligned}$$

By substituting the matrix parameters $M = \begin{bmatrix} m & n \\ p & q \end{bmatrix}$ by $\check{M} = \begin{bmatrix} 0 & n \\ p & q \end{bmatrix}$, the above identity can be written in the form

$$\begin{aligned} \left(\int_{\mathbb{R}} |f(x)|^2 dx \right)^2 &\leq 4 \int_{\mathbb{R}} x^2 |f(x)|^2 dx \int_{\mathbb{R}} |(-p\omega i) L_{\check{M}} \{f\}(\omega)|^2 d\omega \\ &= 4p^2 \int_{\mathbb{R}} x^2 |f(x)|^2 dx \int_{\mathbb{R}} \omega^2 |L_{\check{M}} \{f\}(\omega)|^2 d\omega. \end{aligned}$$

This proof is complete. □

Remark 1 It is not difficult to see that if $\|f\| = 1$, $M = \begin{bmatrix} m & 2 \\ p & q \end{bmatrix}$, $\check{M} = \begin{bmatrix} 0 & \frac{1}{2} \\ p & q \end{bmatrix}$ Eqs. (24.27) and (24.28) will be reduced to

$$\int_{\mathbb{R}} x^2 |f(x)|^2 dx \int_{\mathbb{R}} \omega^2 |L_M \{f\}(\omega)|^2 d\omega \geq 1, \tag{24.30}$$

and

$$\int_{\mathbb{R}} x^2 |f(x)|^2 dx \int_{\mathbb{R}} \omega^2 |L_{\check{M}} \{f\}(\omega)|^2 d\omega \geq 1, \tag{24.31}$$

For example, the Gaussian function may be defined in the form

$$f(x) = e^{-\alpha x^2}, \quad \alpha > 0. \tag{24.32}$$

Its LCT is given by

$$L_{\check{M}} \{e^{-\alpha x^2}\}(\omega) = \frac{1}{\sqrt{2\alpha ni}} e^{\frac{p+2\alpha qi}{4\alpha n} \omega^2}. \tag{24.33}$$

24.4 Conclusion and Future Work

We have presented new properties of the LCT. By using those properties, we have established the inequality for the LCT, which describes the relationship between a signal and its LCT. The future work will be focused on the generalization of properties and uncertainty principles in the quaternion linear canonical transform which is a generalization of the LCT in the framework of quaternion algebra and develop the LCT applications in partial differential equations.

Acknowledgements This work was supported by JSPS KAKENHI Grant Numbers JP16K05216, JP17K05298, JP17K05363 of Japan.

References

1. Urynbassarova, D., Li, B.-Z., Tao, R.: Convolution and correlation theorems for Wigner-Ville distribution associated with the offset linear canonical transform. *Optik-Int. J. Light Electron Opt.* **157**, 455–466 (2018)
2. Bahri, M., Ashino, R.: A simplified proof of uncertainty principle for quaternion linear canonical transform. *Abst. Appl. Anal.* **2016**(Article ID 5874930) (2016), 11 p
3. Bahri, M., Ashino, R.: Convolution and correlation theorem for linear canonical transform and properties. *Information* **17**(6B), 2509–2521 (2014)
4. Bahri, M., Amir, A.K., Ashino, R.: Formulation, correlation, using relationship between convolution and correlation in linear canonical transform domain. In: *International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, pp. 177–182. Ningbo, China (2017)
5. Tao, R., Li, Y.-L., Wang, Y.: Uncertainty principles for linear canonical transforms. *IEEE Trans. Sig. Process.* **57**(7), 2856–2858 (2009)
6. Wei, D., Ren, Q., Li, Y.: A convolution and correlation theorem for the linear canonical transform and its application. *Circ. Syst. Sig. Process.* **31**(1), 301–312 (2012)
7. Wei, D., Ran, Q., Li, Y.: New convolution theorem for the linear canonical transform and its translation invariance property. *Optik-Int. J. Light Electron. Opt.* **123**(16), 1478–1481 (2012)
8. Debnath, L., Shah, F.A.: *Wavelet Transforms and their Applications*. Birkhäuser, New York (2010)
9. Gröchenig, K.: *Foundation of Time-Frequency Analysis*. Birkhäuser, Boston (2001)
10. Moshinsky, M., Quesnee, C.: Linear canonical transform and their unitary representations. *Journal of Mathematical Physics* **12**, 1772–1783 (1971)
11. Bahri, M., Ashino, R.: Convolution and correlation theorems for Wigner-Ville distribution associated with linear canonical transform. In: *Proceedings of the 12th International Conference on Information Technology: New Generations (ITNG)*. Las Vegas, NV, USA (2015)
12. Bahri, M., Ashino, R.: Some properties of windowed linear canonical transform and its logarithmic uncertainty principle. *Int. J. Wavelets Multiresolut Inf. Process.* **14**, 21 p
13. Bai, R.-F., Li, B.-Z., Cheng, Q.-Y.: Wigner-Ville distribution associated with the linear canonical transform. *J. Appl. Math.* **2012**(Article ID 740161), 14 p
14. Urynbassarova, D., Li, B.Z., Tao, R.: The Wigner-Ville distribution in the Linear canonical transform domain. *IAENG Int. J. Appl. Math.* **46**(4), 559–563 (2016)
15. Urynbassarova, D., Li, B.Z., Tao, R.: Convolution and correlation theorems for Wigner-Ville distribution associated with the offset linear canonical transform. *Optik* **157**, 455–466 (2018)
16. Bahri, M., Haddade, A., Toaha, S.: Some useful properties of ambiguity function associated with linear canonical transform. *Far East J. Electron. Commun.* **17**(2), 455–473 (2017)

17. Bahri, M.: Relationships among various definitions of two-dimensional quaternion linear canonical transform. *IAENG In. J. Appl. Math.* **49**(1), 61–68 (2019)
18. Kou, K.I., Xu, R.H.: Windowed linear canonical transform and its applications. *Sig. Process.* **92**(1), 179–188 (2012)
19. Li, Y.-G., Li, B.-Z., Sun, H.-F.: Uncertainty principle for Wigner-Ville distribution associated with the linear canonical transform. *Abstr. Appl. Anal.* **2014**(Article ID 470459), 9 p
20. Guanlei, X., Xiaotong, W., Xiaogang, X.: Uncertainty inequities for linear canonical transform. *IET Sig. Process.* **3**(5), 392–402 (2009)

Chapter 25

A General Approach to Probabilistic Data Mining



Radim Jiroušek and Václav Kratochvíl

Abstract The basic idea of compositional models is very simple: It is beyond human capabilities to describe global knowledge from an application area—one always works only with pieces of local knowledge. Such local knowledge can be, within probability theory, easily represented by low-dimensional distributions. This idea is employed in this paper in the reversed order. Multidimensional probability distribution, which generated the data, is approximated by a compositional model, and pieces of local knowledge, i.e., results of the data mining process, are read from the low-dimensional distribution constituting the compositional model, and from the way, how these low-dimensional distributions are composed together (more precisely, in what order they are composed).

25.1 Introduction

There is abundant literature on data mining and, quite naturally, a great number of different definitions explaining what the authors understand by this notion. All the authors agree that data mining is a process discovering interesting relationships that are to be found in large databases, the process uncovering useful information that can be expressed in the form of knowledge. And this is the point in which the individual data mining processes differ from each other. Some of the authors consider data mining to be a part of machine learning, as e.g., [2, 3], and therefore they represent the discovered knowledge in a form of specific models. Some others look for the information supporting optimum decision-making processes [4, 5]. The present paper

This survey lecture is patterned on the manuscript of the book [1], and on preceding papers of the authors.

R. Jiroušek (✉) · V. Kratochvíl
Institute of Information Theory and Automation, Czech Academy of Sciences,
Prague, Czech Republic
e-mail: velorex@utia.cas.cz

Faculty of Management, University of Economics, Prague, Czech Republic
e-mail: radim@utia.cas.cz

Table 25.1 An example of a direct proportion

Age	Proportion of patients with disease D (%)
Less than 40	2.1
40–49	7.3
50–59	14.6
60–69	31.1
70+	44.9

goes back to principles of data mining technology, showing that this technique can produce new knowledge expressible in a plain language legible to human experts. One of the oldest papers of this type is [6] describing the GUHA method. A modern version of these ideas can be found in [7]. These papers describe methods for discovering unknown relations between (among) variables, the relations that can easily be expressed verbally (e.g., in a form of IF-THEN rules loaded with some uncertainty). As we are now going to support with arguments, in the approach described in this paper, we look for knowledge representable with the tools of probability theory.

In the beginning, the researchers in artificial intelligence were reluctant to accept probability theory for knowledge representation and inference because of several reasons. Among them, the rigidity of statistical methodology and the complexity of the respective computational procedures played important roles. It was in the middle of the eighties of the last century when the probability theory started penetrating into the field of artificial intelligence thanks to the papers like [8], and the tools based on probabilistic graphical models [9, 10]. Naturally, we do not claim that the probability theory is an approach capable to represent all the forms of knowledge, but it is general enough that it can serve for the purpose of this paper. It can represent a logical implication (IF-THEN rule) by a two-dimensional distribution (fourfold table) with one zero value. If this rule is loaded with uncertainty, then it contains instead of the zero a small probability. In probability logic [11], the validity of implication can be expressed by a conditional probability. Moreover, probability table (distribution), like in Table 25.1, can represent a type of dependence we express in words “the older, the greater the risk of suffering from disease D.”

There are many other types of dependence that can be read from a respective low-dimensional probability table. One of the most important is *independence*, or more generally, *conditional independence*. The concept of independence plays an important role in human thinking and therefore we will speak about it later in more details. If the reader cannot find a more general description of dependence, it is always possible to express it as a series of conditional probabilities, i.e., a series of IF-THEN rules, the validity of which corresponds to the value of the corresponding conditional probability. But keep in mind that the type of dependence can be read only from the low-dimensional probability tables. As a rule, the dimension should usually be lower than five. Otherwise, the tables starts to be labyrinthine, and, what is even more important, the estimates of the respective probabilities are unreliable.

Before proceeding to a more formal exposition let us admit that the probability theory has also its limits: for example, it cannot model *ambiguity*. It has been known since the middle of the last century [12] that humans do not like ambiguity. They prefer situations when they know probabilities of the alternatives influencing their decision, and hate situations when the probabilities are fully unknown. This phenomenon is connected with an Ellsberg paradox [13], and it is known that classical probability theory cannot treat such situations easily. The only way how to overcome this problem is to employ some generalization designed for treating uncertain probabilities, as, e.g., theory of imprecise probabilities, or belief function theory [14].

25.2 Basic Notions and Notation

Let us assume the records from the available data file represent observations of random variables, which are, in this paper, denoted by upper-case characters of Latin alphabet (like X, Y, \dots). Thus, for example, the *gender* and *age* of a responder may be two variables. The first one is binary, it can have only two values (female, male), for the latter one, several intervals of age may be considered. In this paper, finite sets of values of variables will be considered. The set of values of variable X will be denoted by \mathbb{X}_X , for variable Y the set of its value will be \mathbb{X}_Y , and so on. Since most of the time we will deal with sets of variables, we will use bold-face characters $\mathbf{K}, \mathbf{L}, \mathbf{M}, \mathbf{N}$ to denote them. Thus, \mathbf{K} may be $\{X, Y, W\}$. By a *state* of variables \mathbf{K} we understand any combination of values of the respective variables, i.e., in the considered case $\mathbf{K} = \{X, Y, W\}$, a state is an element of a Cartesian product $\mathbb{X}_X \times \mathbb{X}_Y \times \mathbb{X}_W$. For the sake of simplicity, this Cartesian product is denoted $\mathbb{X}_{\mathbf{K}}$. For a state $y \in \mathbb{X}_{\mathbf{K}}$ and $\mathbf{L} \subset \mathbf{K}$, $y^{\downarrow \mathbf{L}}$ denote a *projection* of $y \in \mathbb{X}_{\mathbf{K}}$ into $\mathbb{X}_{\mathbf{L}}$, i.e., $y^{\downarrow \mathbf{L}}$ is the state from $\mathbb{X}_{\mathbf{L}}$ that is got from y by dropping out all the values of variables from $\mathbf{K} \setminus \mathbf{L}$.

Probability tables (distributions) are denoted by characters of Greek alphabet ($\kappa, \lambda, \mu, \pi$). Recall that it means that $\kappa(\mathbf{K}) : \mathbb{X}_{\mathbf{K}} \rightarrow [0, 1]$, for which $\sum_{x \in \mathbb{X}_{\mathbf{K}}} \kappa(x) = 1$. For an example of a two-dimensional distribution see Table 25.2.

Having a probability distribution $\kappa(\mathbf{K})$, and a subset of variables $\mathbf{L} \subset \mathbf{K}$, $\kappa^{\downarrow \mathbf{L}}$ denote a *marginal distribution* of κ defined for each $x \in \mathbb{X}_{\mathbf{L}}$ by the formula

Table 25.2 Two-dimensional distribution for $X = \text{age}$, and $Y = \text{disease D}$

Age	With disease D	Without disease D
Less than 40	0.00105	0.18895
40–49	0.00365	0.17635
50–59	0.00730	0.19270
60–69	0.01555	0.21445
70+	0.02245	0.12755

$$\kappa^{\downarrow \mathbf{L}}(x) = \sum_{y \in \mathbb{X}_{\mathbf{K}}: y^{\downarrow \mathbf{L}}=x} \kappa(y). \quad (25.1)$$

Consider two distributions $\kappa(\mathbf{K})$ and $\lambda(\mathbf{L})$. We say that κ and λ are *consistent* if $\pi^{\downarrow \mathbf{K} \cap \mathbf{L}} = \kappa^{\downarrow \mathbf{K} \cap \mathbf{L}}$. For two probability distributions defined for the same group of variables, say $\pi(\mathbf{K}), \kappa(\mathbf{K})$, we say that κ *dominates* π (in symbol $\pi \ll \kappa$) if

$$\forall x \in \mathbb{X}_{\mathbf{K}} \quad (\kappa(x) = 0 \implies \pi(x) = 0).$$

Consider a probability distribution $\pi(\mathbf{N})$, and three disjoint subsets of variables $\mathbf{K}, \mathbf{L}, \mathbf{M}$ ($\mathbf{K} \cup \mathbf{L} \cup \mathbf{M} \subseteq \mathbf{N}$). Let \mathbf{K} and \mathbf{L} be nonempty. We say that groups of variables \mathbf{K} and \mathbf{L} are *conditionally independent given \mathbf{M} for distribution π* if

$$\pi^{\downarrow \mathbf{K} \cup \mathbf{L} \cup \mathbf{M}} \cdot \pi^{\downarrow \mathbf{M}} = \pi^{\downarrow \mathbf{K} \cup \mathbf{M}} \cdot \pi^{\downarrow \mathbf{L} \cup \mathbf{M}}. \quad (25.2)$$

In symbol, this property is expressed by $\mathbf{K} \perp\!\!\!\perp \mathbf{L} \mid \mathbf{M} [\pi]$. In case of $\mathbf{M} = \emptyset$, we use only $\mathbf{K} \perp\!\!\!\perp \mathbf{L} [\pi]$ and speak about an unconditional independence.

We have already mentioned that the conditional independence is considered in this paper as a property expressing knowledge about the reality described by the considered probability distribution. Perhaps, it is not visible just from the definition, but it is an easy exercise to show that if $\mathbf{K} \perp\!\!\!\perp \mathbf{L} \mid \mathbf{M} [\pi]$, then

$$\pi^{\downarrow \mathbf{K} \mid \mathbf{L} \cup \mathbf{M}} = \pi^{\downarrow \mathbf{K} \mid \mathbf{M}}. \quad (25.3)$$

In words: If we know a state from $\mathbb{X}_{\mathbf{M}}$, then learning values of variables from \mathbf{L} does not bring us any new information about variables from \mathbf{K} . If we know a state from $\mathbb{X}_{\mathbf{M}}$, then groups of variables \mathbf{K} and \mathbf{L} become independent. For example, the intensity of training and the placing of a runner in a race are dependent. But conditionally, given the time in which the runner accomplished the race, these two events become independent. Namely, when knowing the time she achieved in the race, the probability that she wins the race does not change when learning how much time she spent in training.

For a probability distribution π , by its *independence structure* we understand the list of all conditional independence relations holding for π . It explains us, which of the dependence relations between variables are direct, and which are mediated through other variables.

25.3 Decomposition of Probability Distributions

As said in introduction, humans can read knowledge from low-dimensional probability tables. It means that when considering a multidimensional distribution one has to decompose it into low-dimensional ones, and the decomposition should be done in the way that it gives evidence about the data [15]. Moreover, any decomposition is required to meet the following properties

- the result of the decomposition are two objects of the same type as the decomposed object;
- both these sub-objects are simpler (smaller) than the original object;
- not all objects can be decomposed;
- there exists an inverse operation (we call it a composition) yielding the original object from its decomposed parts.

The following definition directly follows from the above-stated properties.

Definition 1 We say that a probability distribution $\pi(\mathbf{M})$ is *decomposed* into its marginals $\pi^{\downarrow\mathbf{K}}$ and $\pi^{\downarrow\mathbf{L}}$ if

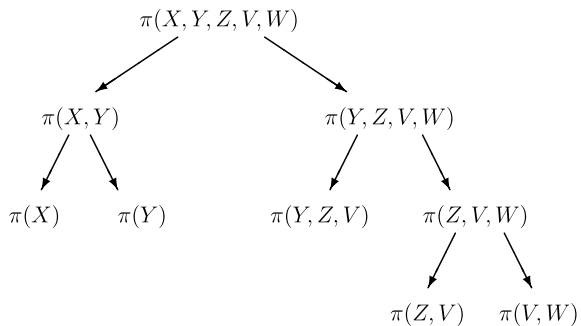
1. $\mathbf{K} \cup \mathbf{L} = \mathbf{M}$;
2. $\mathbf{K} \neq \mathbf{M}, \mathbf{L} \neq \mathbf{M}$;
3. $\pi(\mathbf{M}) \cdot \pi^{\downarrow\mathbf{K} \cap \mathbf{L}} = \pi^{\downarrow\mathbf{K}} \cdot \pi^{\downarrow\mathbf{L}}$.

Notice that the third condition is nothing else than $\mathbf{K} \perp\!\!\!\perp \mathbf{L} \mid \mathbf{K} \cap \mathbf{L} [\pi]$, and that the original distribution $\pi(\mathbf{M})$ can be uniquely reconstructed from the marginals $\pi^{\downarrow\mathbf{K}}$ and $\pi^{\downarrow\mathbf{L}}$.

Probability distributions can hierarchically be decomposed into a system of low-dimensional distributions that cannot be further decomposed. An example of such a hierarchical process represented by a corresponding tree structure can be seen in Fig. 25.1, where distribution $\pi(X, Y, Z, V, W)$ is decomposed into a system of its marginal distributions: $\pi^{\downarrow X}, \pi^{\downarrow Y}, \pi^{\downarrow\{Y,Z,V\}}, \pi^{\downarrow\{Z,V\}}, \pi^{\downarrow\{V,W\}}$. Each decomposition is made possible by the fact that the respective conditional independence relation holds for distribution π . It means that the decomposition process from Fig. 25.1 is made possible by the assumption that the following system of conditional independence relations holds for distribution π :

- $X \perp\!\!\!\perp \{Z, V, W\} \mid Y [\pi]$;
- $X \perp\!\!\!\perp Y [\pi]$;
- $Y \perp\!\!\!\perp W \mid \{Z, V\} [\pi]$;
- $Z \perp\!\!\!\perp W \mid V [\pi]$.

Fig. 25.1 Hierarchical decomposition of $\pi(X, Y, Z, V, W)$



Therefore, having a multidimensional probability distribution (a generator of the considered data) decomposed into its “primes” (low-dimensional distributions that cannot be further decomposed), we are able to express all the knowledge contained in the distribution by

- the list of conditional independence relations enabling the decomposition of the considered multidimensional distribution;
- all the knowledge that can be read from the low-dimensional “prime” distributions.

The only problem is that the process of decomposition of a multidimensional distribution (represented by the considered data file) is, as a rule, computationally intractable. Therefore, we have to find an indirect process that yields a decomposition of such a multidimensional distribution, or, which is more realistic, to find an approximation of such a hierarchical decomposition process. And the description of such a process forms the content of the rest of the paper.

25.4 Compositional Models

The basic idea is as follows: If we cannot decompose a multidimensional distribution because of the great computational complexity of the respective process, we find the approximation of the considered distribution, which is in a form of a compositional model. It means that its decomposition is “visible” from the structure of the model. Compositional models are multidimensional distributions *composed* from a system of low-dimensional distributions by an operator of a composition realizing an inverse process to the decomposition defined in the previous section.

Recall that $\pi(\mathbf{N})$ can be decomposed into its marginals $\pi(\mathbf{K})$ and $\pi(\mathbf{L})$ if $\mathbf{K} \cup \mathbf{L} = \mathbf{N}$ and $\pi(\mathbf{N}) \cdot \pi^{\downarrow \mathbf{K} \cap \mathbf{L}} = \pi^{\downarrow \mathbf{K}} \cdot \pi^{\downarrow \mathbf{L}}$. From this, one immediately gets that an inverse operation, the operation of composition is

$$\pi(\mathbf{N}) = \frac{\pi^{\downarrow \mathbf{K}} \cdot \pi^{\downarrow \mathbf{L}}}{\pi^{\downarrow \mathbf{K} \cap \mathbf{L}}}. \quad (25.4)$$

This is trivial if we compose distributions $\kappa(\mathbf{K})$ and $\lambda(\mathbf{L})$ that are consistent. The question is whether one can compose also inconsistent distributions, i.e., distributions $\kappa(\mathbf{K})$ and $\lambda(\mathbf{L})$, for which $\kappa^{\downarrow \mathbf{K} \cap \mathbf{L}} \neq \lambda^{\downarrow \mathbf{K} \cap \mathbf{L}}$. We need it because the estimates got from a data file with missing values are rarely consistent. Therefore, we advocate the following definition that was first introduced in [16].

Definition 2 For arbitrary two distributions $\kappa(\mathbf{K})$ and $\lambda(\mathbf{L})$, for which $\lambda^{\downarrow \mathbf{K} \cap \mathbf{L}}$ dominates $\kappa^{\downarrow \mathbf{K} \cap \mathbf{L}}$, their *composition* is for each $x \in \mathbb{X}_{\mathbf{K} \cup \mathbf{L}}$ given by the following formula¹

¹Define $\frac{0 \cdot 0}{0} = 0$.

$$(\kappa \triangleright \lambda)(x) = \frac{\kappa(x \downarrow \mathbf{K}) \lambda(x \downarrow \mathbf{L})}{\lambda \downarrow \mathbf{K} \cap \mathbf{L} (x \downarrow \mathbf{K} \cap \mathbf{L})}. \quad (25.5)$$

In case that $\kappa \downarrow \mathbf{K} \cap \mathbf{L} \not\ll \lambda \downarrow \mathbf{K} \cap \mathbf{L}$, the composition remains undefined.

By a *multidimensional compositional model*, we understand a multidimensional probability distribution assembled from a sequences of low-dimensional distributions with the help of the introduced operator of composition, i.e.,

$$\kappa_1 \triangleright \kappa_2 \triangleright \cdots \triangleright \kappa_n.$$

Unfortunately, the operator of composition is not associative, and therefore the above expression is ambiguous. Therefore, let us make a convention that we will omit the parentheses if the operators are to be performed from left to right:

$$\kappa_1 \triangleright \kappa_2 \triangleright \cdots \triangleright \kappa_n = (\cdots ((\kappa_1 \triangleright \kappa_2) \triangleright \kappa_3) \triangleright \cdots \triangleright \kappa_{n-1}) \triangleright \kappa_n.$$

On the other hand, it is important that for these models, similarly to Bayesian networks, efficient computational algorithms were designed that make the application of these models for the inference possible. There are algorithms for the marginalization of compositional models and for computation of conditional distributions. In this paper, we are not interested in this type of applications, and therefore we do not need to present all the properties of the operator of composition, which make the theoretical basis for the design of these computational procedures. At this place, let us highlight just that this operator is generally neither commutative nor associative. The reader interested in other theoretical issues concerning the operator of composition is referred to [17] and the papers cited there.

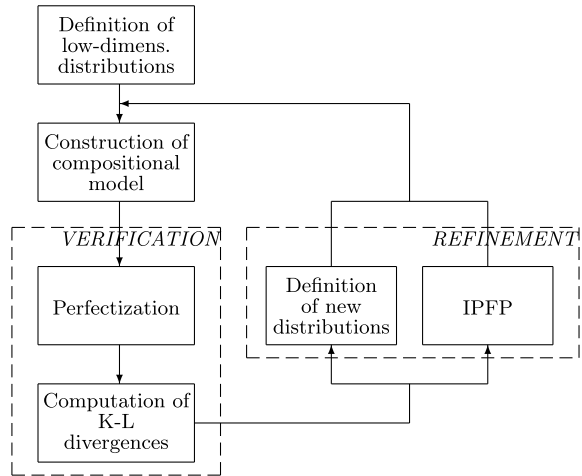
25.5 Heuristics for Model Construction

As said at the beginning of Sect. 25.4, having data, the goal is to construct a compositional model approximating the distribution that generated the data. This model is a bearer of the knowledge from the data. Recall the two types of knowledge mentioned at the beginning of Sect. 25.3 that can be directly read from a compositional model: First, it is the list of conditional independence relations holding for the distribution represented by the model,² second, each of the low-dimensional distributions, from which the model is composed, can be interpreted in the way mentioned in the introductory section.

Similarly to Bayesian network construction, there is no generally accepted “best” approach to data-based compositional model construction. For the purpose of data

²Instructions for reading all the conditional independence relation from the structure of the model can be found in [18].

Fig. 25.2 Process of model construction



mining, one possibility is to use a heuristic procedure³ schematically depicted in Fig. 25.2. Notice that the described process is fully controlled by an expert, who has a possibility to influence the constructed model, and thus consequently also the type of the received knowledge.

As can be seen from the diagram in Fig. 25.2, the process is initiated with the definition of a system of low-dimensional distributions. Regarding the fact that the process cyclically employs steps of *verification* and *refinement*, during which this initial system is gradually changed, the result is fairly independent of the initial selection. For example, starting with all two-dimensional distributions may be quite reasonable (for application to small data files with a limited number of variables one can consider a possibility to start with three-dimensional marginal distributions). In other situations, an expert can select the initial marginal distributions from which the model should be constructed. Generally, we propose to select distributions carrying a greater amount of information. This idea is supported by the assertion, proved in [20] (Corollary 1). It claims that the higher information content of a compositional model, the better approximation of the unknown distribution. This means that it is necessary to compute the information content of a multidimensional distribution, which is for a distribution $\pi(\mathbf{K})$ defined⁴

³For a more detailed description of this process, as well as for the survey of the necessary theoretical background, the reader is referred to [19].

⁴ $\text{Div}(\pi \parallel \mu)$ denotes the famous Kullback–Leibler divergence defined (for $\pi(\mathbf{K})$ and $\mu(\mathbf{K})$)

$$\text{Div}(\pi \parallel \mu) = \sum_{x \in \mathbb{X}_{\mathbf{K}}} \pi(x) \log \frac{\pi(x)}{\mu(x)}.$$

$$\text{IC}(\pi) = \text{Div}(\pi \parallel \prod_{X \in \mathbf{K}} \pi^{\downarrow\{X\}}) = \sum_{x \in \mathbb{X}_{\mathbf{K}}} \pi(x) \log \frac{\pi(x)}{\prod_{X \in \mathbf{K}} \pi^{\downarrow\{X\}}(x^{\downarrow\{X\}})}. \quad (25.6)$$

Notice, the information content is a generalization of the famous *Shannon mutual information*

$$\text{MI}_{\pi}(X, Y) = \sum_{(x, y) \in \mathbb{X} \times \mathbb{Y}} \pi(x, y) \log \frac{\pi(x, y)}{\pi(x) \cdot \pi(y)}. \quad (25.7)$$

In the process of model construction, the following conditional version of Shannon mutual information is often utilized

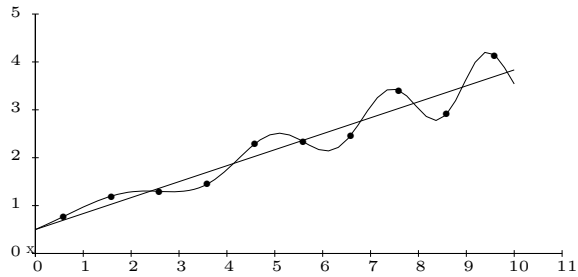
$$\text{MI}_{\pi}(X, Y|Z) = \sum_{(x, y, z) \in \mathbb{X} \times \mathbb{Y} \times \mathbb{Z}} \pi(x, y, z) \log \frac{\pi(x, y, z) \cdot \pi(z)}{\pi(x, z) \cdot \pi(y, z)}. \quad (25.8)$$

It is important to highlight here, that computation of this value is computationally cheap for, so-called, *perfect models*. Therefore, the process of “perfectization” of the model is included into the process of model construction. For perfect models, the information content of the whole multidimensional distribution can be computed from the information content of the individual low-dimensional distributions. This is why the computation of this value for the model is cheap, and also why we want to get low-dimensional distributions with high information content.

Realization of the box “computation of Kullback–Leibler divergence” in Fig. 25.2 means computing the divergence between the distribution given by data and the distribution represented by a model. It can easily be done for perfect models. It helps the user to decide whether the model reasonably approximates the distribution that generated data. Naturally, one cannot expect that the first choice of the low-dimensional distributions would yield a satisfactory model. This is true even more in situations when starting just with two-dimensional distributions. Comparing the data distribution and the model locally (i.e., computing the Kullback–Leibler divergence for the marginals corresponding to the low-dimensional distributions, from which the model is set up), the user can see, which parts of the model do not reflect the data properly. As the reader will see in the next section, it usually means that it is necessary to increase the dimension of some low-dimensional distributions. It can be done, as one can see in Fig. 25.2, in two ways. Having enough data, one can get just new estimates from the data. However, quite often it may be better to get a maximum entropy estimate from several low-dimensional distributions of the original model by the *iterative proportional fitting procedure* [21].

Let us stress once more that the process in Fig. 25.2 is fully controlled by the expert. The more cycles of the process are performed, the higher dimensions of the input distributions are usually considered. If the expert had continued ad absurdum, the process would have finished with an application of IPFP to all of the initial low-dimensional distributions (which is, computationally intractable in practical situations). Therefore, it is obvious that one has to avoid the overfitting of the model.

Fig. 25.3 Overfitted linear dependence



Model overfitting is a well-known phenomenon both in statistics [22] and machine learning (artificial intelligence) [23]. It is used to describe a situation when a constructed model reflects the noninformative properties of the source data file (like noise and other misleading properties that each randomly generated data file possesses). Let us illustrate it on two stochastically dependent variables, the dependence of which is known to be linear. Because the dependence is stochastic, if randomly generated data are plotted in a graph, the respective dots are concentrated along a straight line describing the dependence. Naturally, only a part of dots lies directly on the line. If one tries to find a curve that connects all the dots in the plot (see Fig. 25.3), the model becomes for knowledge discovery useless. Realize that such a complex curve is described (defined) by a much larger number of parameters than the straight line, which can be determined just by two points.

In agreement with other parts of this section, we propose a solution to this problem from the perspective of information theory. Namely, the process of model construction can be viewed as a transformation process; a process transforming the information contained in data into the information represented by a model. Thus, using one of the basic laws of information theory saying that any transformation cannot increase the amount of information, we get the basic restriction laid on models constructed from data: A model is *acceptable* if it does not contain more information than the input data file.

However, the application of this idea hits the problem, how to measure information in a data file, and how to measure information contained in a model. For this, we go back more than a half a century to seminal papers by von Mises [24] and Kolmogorov [25], who explored relations interconnecting randomness, complexity and information. They came with the idea that the amount of information in a sequence of 0's and 1's is increasing with the complexity of the sequence, and that the complexity of such a sequence can be measured by the length of the shortest program⁵ generating the sequence. We accept here this idea but instead of the length of an (abstract) program we consider the length of a lossless encoding (one can always generate the sequence from its lossless encoding).

⁵An abstract program for a universal Turing machine.

Therefore, in agreement with results of Kolmogorov and von Mises, before accepting a final compositional model, we look for the shortest possible encoding of both, data file and the resulting model. In case we get a model, the encoding of which requires more bits than the encoding of data, we are sure, that some undesirable information has been added into the model. Regardless of the way the data were collected, they always contain some specific part of the information, employment of which results in the overfitting of the model. It should not be included in the model. Thus, we enforce a principle under which we accept only models, the encoding of which is substantially shorter than the encoding of the data file. The meaning of the word substantially is usually left to the user's discretion.

25.6 Example

The goal of this section is to show how the theoretical ideas presented above can be utilized during the process of supervised model construction, and what type of knowledge can be gained during this process. A supervised process is used because of several reasons. First, no generally accepted method for optimum model construction is known. Second, the user usually has some prior knowledge about the area of application, and this knowledge should be utilized during the process. Further, the user can have some knowledge about data, based on which the model is constructed. They may know that the data are not well stratified and some properties should be suppressed some others highlighted. Quite often, they want to adapt the constructed model to the purpose for which the discovered knowledge will be used. Therefore, it is natural that we cannot give general instructions how to proceed when constructing a model.

The example taken from [1] considers six random variables $\mathbf{M} = \{B, D, N, R, T, W\}$ with $\mathbb{X}_B = \{1, 2, 3\}$ and $\mathbb{X}_D = \mathbb{X}_N = \mathbb{X}_R = \mathbb{X}_T = \mathbb{X}_W = \{1, 2\}$. From the point of view of model construction, we are interested in couples of variables, which are closely (strongly) connected, and in couples of independent variables. Therefore, after computing values of mutual information for all pairs of variables, we sort the couples according to the value of mutual information. In the present example, we get

$$\left\{ \begin{array}{l} \text{MI}(D; N) = 0.4356, \\ \text{MI}(B; R) = 0.2871, \\ \text{MI}(R; W) = 0.2578, \\ \text{MI}(N; R) = 0.2070, \\ \text{MI}(T; W) = 0.1813, \\ \text{MI}(N; W) = 0.1546 \\ \text{MI}(D; R) = 0.0958 \\ \text{MI}(B; W) = 0.0814 \\ \text{MI}(N; T) = 0.0709 \end{array} \right.$$

$$\begin{aligned}
& \text{MI}(D; W) = 0.0627 \\
& \text{MI}(B; N) = 0.0619 \\
& \text{MI}(D; T) = 0.0421 \\
& \text{MI}(B; D) = 0.0342 \\
& \left\{ \begin{array}{l} \text{MI}(R; T) = 0.0019, \\ \text{MI}(B; T) = 0.0007. \end{array} \right.
\end{aligned}$$

The head of this sequence contains the couples of closely connected variables, the tail of this sequence may point at independent variables. The first five couples are grouped together because these first five couples covers the whole \mathbf{M} . Therefore, let us start building compositional models from two-dimensional distributions defined for these couples of variables. To get their best ordering in a model, the information content of the whole model should be taken into account. The higher information content, the better the model because it incorporates more information from data.

Consider estimates of the first five two-dimensional distributions got from data and denote them, respectively: $\kappa_1(D, N)$, $\kappa_2(B, R)$, $\kappa_3(R, W)$, $\kappa_4(N, R)$, $\kappa_5(T, W)$. If considering model $\pi_1 = \kappa_1 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_4 \triangleright \kappa_5$ we get $\text{IC}(\pi_1) = 1.1618$. For $\pi_2 = \kappa_1 \triangleright \kappa_4 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_5$ we get $\text{IC}(\pi_2) = 1.3687$. In fact, this model is the best possible among those assembled from distributions $\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5$, if the amount of information content is taken as the only criterion of optimality. However, in the step of model verification, it appears that this model does not reflect the other information obtained from computing the mutual information for all couples of variables: The two smallest values of mutual information suggest that variables T and R , and variables T and B are independent.

To incorporate this knowledge into the model, one has to proceed to the REFINEMENT step, and define a new model, e.g., $\pi_3 = \kappa_5 \downarrow^T \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_4 \triangleright \kappa_5 \triangleright \kappa_1$. However, $\text{IC}(\pi_3) = 1.1874$. Thus, it may seem that one can incorporate the knowledge about the two independence relations into the model only at the cost of a decrease of information content, i.e., at the cost of the loss of information. To get out of this trap, let us start studying the way, how variable T is connected with all others. Let us compute (from the data) the conditional mutual information of T and B given the remaining variables, and similarly, the conditional mutual information of T and R given the remaining variables. We get

$$\begin{aligned}
& \text{MI}(T; B|D) = 0.002, \quad \text{MI}(T; R|D) = 0.001, \\
& \text{MI}(T; B|N) = 0.006, \quad \text{MI}(T; R|N) = 0.013, \\
& \text{MI}(T; B|W) = 0.024, \quad \text{MI}(T; R|W) = 0.084.
\end{aligned}$$

How to explain the fact that variables T and R can be considered independent ($\text{MI}(R; T) = 0.0019$) but not conditionally independent ($\text{MI}(T; R|W) = 0.084$)? A straightforward explanation is that T and R are independent and jointly influence other variables. In case we know the meaning of the variables, we should choose the one, which is, in our knowledge, directly influenced by T and R (or T and B). Otherwise, we choose the one indicated by the highest value of conditional mutual

information: $MI(T; R|W)$. It makes us believe that two independent variables T and R influence W , and the only way how to incorporate this knowledge into the model is to start considering a three-dimensional distribution: Let $\kappa_6(R, T, W)$ be the corresponding estimate got from data. Naturally, this three-dimensional distribution is a bearer of all the information expressed by both κ_3 and κ_5 , which can be now dropped from the further consideration. Naturally, κ_6 contains more information than κ_3 and κ_5 . It describes the combined influence of T and R on W , which cannot not be expressed by two two-dimensional distributions. To illustrate the fact that a three-dimensional distribution may bear more information than a collection of its two-dimensional marginals, consider the following simple example. Children have usually more fun if the weather is warm. Similarly, they prefer sunny days to days with precipitation. However, in winter, a precipitation in very cold days usually means snowing, which is a great fun for children. And this type of knowledge cannot be expressed just by describing two separate relations: Day temperature and children fun, and precipitation and children fun.

Let us turn back to our example. After adding κ_6 and deleting κ_3 and κ_5 , the remaining distributions $\kappa_1, \kappa_2, \kappa_4, \kappa_6$ provide the best model $\pi_4 = \kappa_6 \triangleright \kappa_2 \triangleright \kappa_4 \triangleright \kappa_1$. Nevertheless, the above discussed independence of variables is not visible from the model, it is only encoded in the distribution κ_6 . Therefore, we can prefer model $\pi_5 = \kappa_6^{\downarrow B} \triangleright \kappa_6^{\downarrow T} \triangleright \kappa_6 \triangleright \kappa_2 \triangleright \kappa_4 \triangleright \kappa_1$, from which the considered independence relations $R \perp\!\!\!\perp T [\pi_5], B \perp\!\!\!\perp T [\pi_5]$ are obvious.

What are the differences between the models π_4 and π_5 ? When computing the information content of these models we get $IC(\pi_4) = 1.4531$, and $IC(\pi_5) = 1.4512$. The imperceptible decrease of the value of information content when transforming π_4 into π_5 is due to small changes necessary for introducing the independence of T and R .

Model π_5 seems to meet all the requirements made for data-based models. Generally, to accept a model the user should realize the model verification process consisting in the verification that

- the independence relations deduced from the model do not contradict the intuition of the supervising user;
- the independence relations deduced from the model are not in contradiction with the values of (conditional) mutual information values computed from data;
- the marginals from which the prefectedized model is set up do not differ substantially from the corresponding estimates from data.

Realizing these verifications one finds that $N \perp\!\!\!\perp T [\pi_5]$, which is in contradiction with $MI(N; T) = 0.0709$. To set this serious imperfectness right, we substitute $\kappa_4(N, R)$ by $\kappa_7(N, R, T)$, and consider model $\pi_6 = \kappa_6^{\downarrow B} \triangleright \kappa_6^{\downarrow T} \triangleright \kappa_6 \triangleright \kappa_2 \triangleright \kappa_7 \triangleright \kappa_1$. For this model, we have $IC(\pi_6) = 1.5659$, and the repeated verification process does not find any further imperfection.

25.7 Conclusions

In this paper, we have described the main ideas on which probabilistic data mining methods are based. The term “general” in the title refers to the fact that the approach can be applied to any data file that may be assumed to be generated by a random generator—by a multidimensional probability distribution. This is also why the paper is focused on knowledge that is encoded in such generators. The knowledge is either qualitative or quantitative. The former one can be characterized by the independence structure of a multidimensional probability distribution, i.e., by a list of the conditional independence relations holding for the probability distribution. It expresses which variables (or groups of variables) are mutually independent and which are interconnected. Moreover, it describes which linkages between the considered variables (features, characteristics) are direct, and which are mediated by other variables. The latter type of knowledge, the quantitative one, is based on the values of probabilities and/or conditional probabilities of states of the considered variables. The instructions on how to formulate this quantitative knowledge in a common language are provided by probabilistic logic [11, 26] and, in some situations, even Bayesian (probabilistic) justification logic [27]. Both these theories explain the relationship between probability theory and logic formulas. Thus, for example, based on probability logic quite often one can understand a conditional probability as the probability of the validity of implication. The justification logic supports the knowledge of why something is known or believed, and it was Milnikel [28], who started studying this type of reasoning within the framework of probability theory. The description of how these theories are applied to interpret values of (conditional) probabilities in a plain language is beyond the scope of this paper. However, neither the deduction of the conditional independence structure nor the computation of the (conditional) probability of states from multidimensional probability distributions is tractable for distributions, the dimensionality of which corresponds to practical applications.

To cope with the problem of the high dimensionality of distributions in a practical situation, we suggest employing the idea used by humans to solve complex problems: The decomposition of a complex problem into its parts (subproblems) that are easier to solve. For data mining, the multidimensional distribution generated the data is approximated by a compositional model, i.e., the multidimensional probability distributions assembled from sequences of low-dimensional distributions. These low-dimensional parts are then connected with the help of the operator of composition. From the structure of the model, one can read all the (conditional) independence relations that form the qualitative knowledge, and selected probabilities of low-dimensional distributions are interpreted in a form of uncertain (probabilistic) logical expressions. It is important to stress that the described process of model construction is controlled by the user. It means that the user, usually an expert in the field of application, can influence the whole process of data mining, which results in the fact that the resulting knowledge is a combination of both sources: Knowledge mined from data and the knowledge of the expert. Naturally, in many places and in particular, in connection with the process of data-based model construction, we

could only present the main ideas. The interested reader is referred either to original journal papers or to the book [1] published in 2019 as the first summarizing text on probabilistic compositional models.

Let us conclude this paper by mentioning that several papers were written also on the compositional models in other uncertainty theories, like possibility theory, belief function theory [14], and even on compositional models in Shenoy's valuation-based systems [29, 30]. So the process is applicable even in the mentioned alternative theories of uncertainty.

Acknowledgements The research was financially supported by grants GAČR no. 19-06569S, and AV ČR no. MOST-04-18.

References

1. Jiroušek, R., Kratochvíl, V., et al.: *Discrete Probabilistic Models for Data Mining*. Matfyz Press, Praha (2019)
2. Hand, D.J., Mannila, H., Smyth, P.: *Principles of Data Mining (Adaptive Computation and Machine Learning)*. MIT Press, Cambridge, MA (2001)
3. Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E.: Machine learning: a review of classification and combining techniques. *Artif. Intell. Rev.* **26**(3), 159–190 (2006)
4. Golpíra, H.: A novel multiple attribute decision making approach based on interval data using U2P-miner algorithm. *Data Knowl. Eng.* **115**, 116–128 (2018)
5. Liu, Y.-H.: Mining frequent patterns from univariate uncertain data. *Data Knowl. Eng.* **71**(1), 47–68 (2012)
6. Hájek, P., Havránek, T.: On generation of inductive hypotheses. *Int. J. Man-Mach. Stud.* **9**(4), 415–438 (1977)
7. Lin, W., Alvarez, S.A., Ruiz, C.: Efficient adaptive-support association rule mining for recommender systems. *Data Mining Knowl. Discov.* **6**(1), 83–105 (2002)
8. Cheeseman, P.C.: In defense of probability. In: *IJCAI*, vol. 2, pp. 1002–1009. Citeseer (1985)
9. Lauritzen S.L., Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc. Ser. B (Methodol.)* 157–224 (1988)
10. Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (1988)
11. Hailperin, T., et al.: Probability logic. *Notre Dame J. Formal Logic* **25**(3), 198–212 (1984)
12. Ellsberg, D.: Risk, ambiguity, and the savage axioms. *Q. J. Econ.* 643–669 (1961)
13. Camerer, C., Weber, M.: Recent developments in modeling preferences: uncertainty and ambiguity. *J. Risk Uncertain.* **5**(4), 325–370 (1992)
14. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976)
15. Jiroušek, R.: Decomposition of multidimensional distributions represented by perfect sequences. *Ann. Math. Artif. Intell.* **35**(1–4), 215–226 (2002)
16. Jiroušek, R.: Composition of probability measures on finite spaces. In: *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pp. 274–281. Morgan Kaufmann Publishers Inc. (1997)
17. Jiroušek, R.: Foundations of compositional model theory. *Int. J. Gen. Syst.* **40**(6), 623–678 (2011)
18. Jiroušek, R., Kratochvíl, V.: Foundations of compositional models: structural properties. *Int. J. Gen. Syst.* **44**(1), 2–25 (2015)
19. Jiroušek, R.: Data-based construction of multidimensional probabilistic models with mudim. *Logic J. IGPL* **14**(3), 501–520 (2006)

20. Jiroušek, R.: On approximating multidimensional probability distributions by compositional models. In: *ISIPTA*, pp. 305–320 (2003)
21. Deming, W.E., Stephan, F.E.: On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Stat.* **11**(4), 427–444 (1940)
22. Ryan, T.P.: *Modern Regression Methods*, vol. 655. Wiley (2008)
23. Berka, P.: *Dobývání znalostí z databází*. Academia (2003)
24. Von Mises, R.: *Probability, Statistics, and Truth*. Courier Corporation (1981) (Originally published in German by Springer, 1928)
25. Kolmogorov, A.N.: Tri podchoda k kvantitativnomu opredeleniju informacii. *Problemy peredachi informacii* **1**, 4–7 (1965)
26. Pfeifer, N., Kleiter, G.D.: The conditional in mental probability logic. In: *Cognition and Conditionals: Probability and Logic in Human Thinking*, pp. 153–173 (2010)
27. Bucheli, S., Kuznets, R., Studer, T.: Justifications for common knowledge. *J. Appl. Non-Class. Logics* **21**(1), 35–60 (2011)
28. Milnikel, R.S.: The logic of uncertain justifications. *Ann. Pure Appl. Logic* **165**(1), 305–315 (2014)
29. Shenoy, P.P.: A valuation-based language for expert systems. *Int. J. Approx. Reason.* **3**(5), 383–411 (1989)
30. Jiroušek, R., Shenoy, P.P.: Compositional models in valuation-based systems. In: *Belief Functions: Theory and Applications*, pp. 221–228. Springer (2012)

Chapter 26

A Novel Four-Dimensional Chaotic System with Four Cross Terms



Jinmei Liu

Abstract A novel simple four-dimensional chaotic system with four cross terms is discussed. Four linear terms and four nonlinear terms are included in the 4D chaotic system. The system exhibits chaotic behaviors as its parameters vary in a very large range. Its phase portraits, Poincaré maps, equilibrium points, spectra of Lyapunov exponents, bifurcation diagrams, and power spectra are analyzed by mathematical analyses and numerical simulations. Theoretical analysis and simulation test results prove that the proposed 4D system has unstable equilibrium points and strange chaotic or hyper-chaotic attractors when its system parameters belong to a large scope.

26.1 Introduction

Chaotic systems have been one of hot research topics for many years. Some three-dimensional chaotic systems and higher dimensional chaotic systems have been studied and discussed, such as the systems in the literature [1–21]. For many applications, such as chaotic communication, image encryption, steganography, and random bits generation, it is very desirable that a dynamical system will be chaotic or hyper-chaotic even when its system parameters change in a large scope. In the literature [8], we described a 3D chaotic system with wide-range parameters. Further studies on high-dimensional chaotic systems with large-range parameters have also been done by the author. In this paper, a new 4D chaotic system that shows chaotic and hyper-chaotic behaviors as its parameters vary in a large range is proposed, and its dynamic characteristics are analyzed. Besides, the proposed 4D system is easy to construct because it consists of only four linear terms and four cross terms.

The paper is arranged as follows. Mathematic expression and basic characteristics of the novel 4D chaotic system are presented in Sect. 26.2. Dynamical properties of

J. Liu (✉)

School of Information Science and Engineering, University of Jinan, Jinan 250022, Shandong, China

e-mail: jinmei_liu@126.com

© Springer Nature Singapore Pte Ltd. 2021

S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation, Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_26

341

the system, such as equilibria, Poincaré maps, Lyapunov exponent spectra, bifurcation diagrams, and power spectra, are included in Sect. 26.3. The paper is concluded in Sect. 26.4.

26.2 A New 4D Chaotic System

According to the literature [22, 23] written by J. C. Sprott in 2011, it was suggested that a new system should coincide with at least one of the following three standard criteria:

1. The new system should model some important unsolved problem in nature and analyze the problem.
2. The system should exhibit some novel behavior which is not seen in the literature.
3. The system should be simpler than all other available systems exhibiting the observed behavior.

Here, we propose a 4D system that can be chaotic or hyper-chaotic even when its parameters change in a large scope. The system includes only eight terms, where four of them are linear terms and the others are cross terms of two arguments. The proposed 4D system can meet the above second criterion and third criterion.

A new 4D chaotic system is described by the following equations.

$$\begin{cases} \dot{x} = -ax + yz \\ \dot{y} = by - xz \\ \dot{z} = -cz + yw \\ \dot{w} = -dw - xy \end{cases} \quad (26.1)$$

in which all the parameters a , b , c , and d are positive. In Eq. (26.1), four cross products are adopted to construct nonlinear terms. Figures 26.1 and 26.2 show phase portraits of the system (26.1) for the case of $(a, b, c, d) = (4, 1, 2, 0.5)$, with initial value $(x_0, y_0, z_0, w_0) = (0.1, 0.1, 0.1, 0.1)$. From Figs. 26.1 and 26.2, we can see that the system exhibits complex dynamical traces. Figure 26.3 indicates the Poincaré maps.

For system (26.1), we get $\nabla V = \frac{\partial \dot{x}}{\partial x} + \frac{\partial \dot{y}}{\partial y} + \frac{\partial \dot{z}}{\partial z} + \frac{\partial \dot{w}}{\partial w} = -a + b - c - d$.

In order to ensure the dissipative property of the system, $(-a + b - c - d) < 0$ should be satisfied, and the system will converge at a speed of $e^{-(a-b+c+d)t}$. Thus, each volume cell containing the system orbits finally contracts into zero as $t \rightarrow \infty$. In order to make the orbits of the system (26.1) form an attractor, $(a - b + c + d) > 0$ is needed.

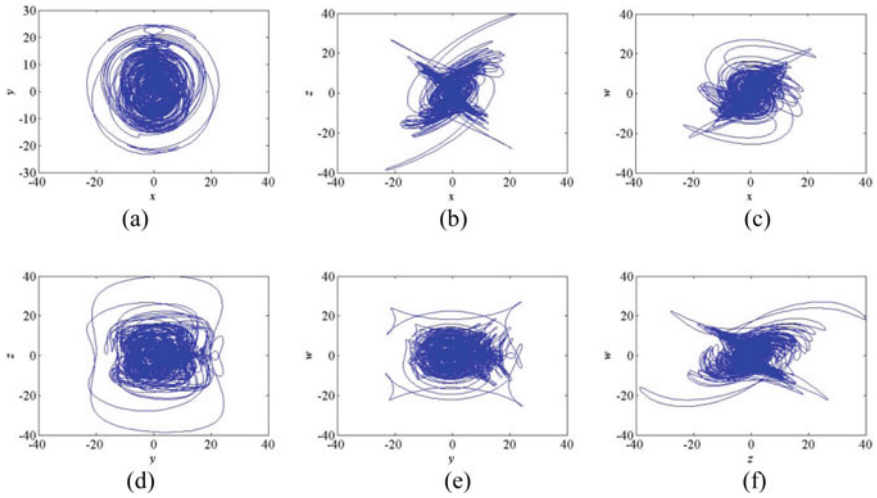


Fig. 26.1 Phase portraits for $(a, b, c, d) = (4, 1, 2, 0.5)$: **a** x, y ; **b** x, z ; **c** x, w ; **d** y, z ; **e** y, w ; **f** z, w

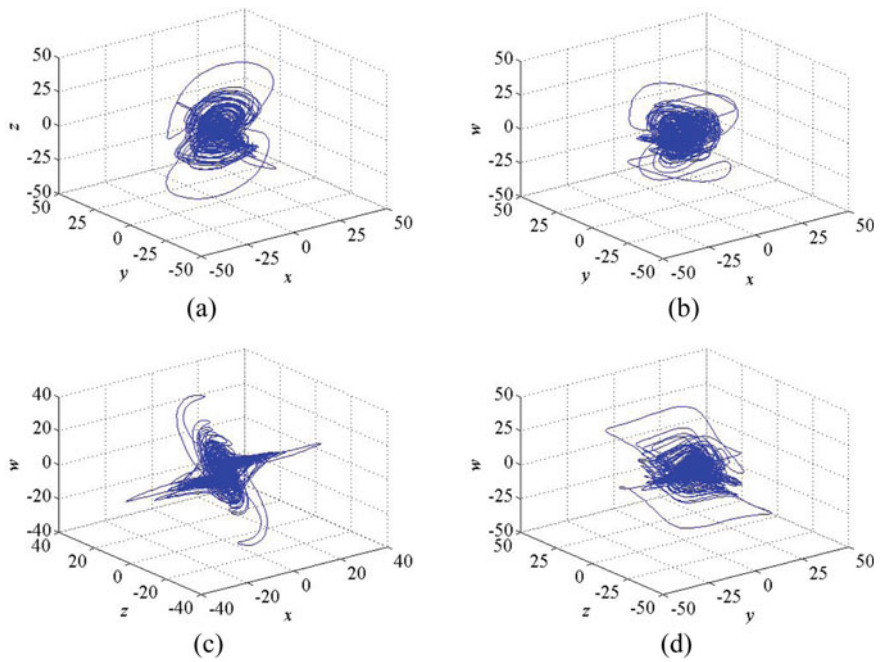


Fig. 26.2 3D phase portraits for $(a, b, c, d) = (4, 1, 2, 0.5)$: **a** x, y, z ; **b** x, y, w ; **c** x, z, w ; **d** y, z, w

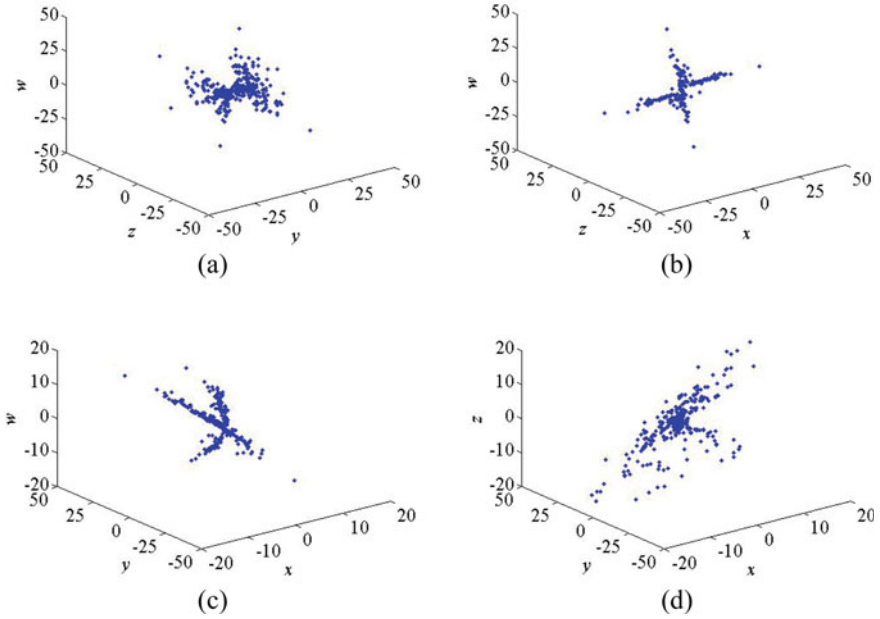


Fig. 26.3 Poincaré maps for $(a, b, c, d) = (4, 1, 2, 0.5)$: **a** $x = 0$; **b** $y = 0$; **c** $z = 0$; **d** $w = 0$

26.3 Some Properties of the New Chaotic System

26.3.1 Symmetry, Equilibrium Points, and Stability

If we use the transformation $(x, y, z, w) \rightarrow (-x, y, -z, -w)$, the system is unchanged. So, the system (26.1) is symmetrical.

To calculate equilibrium points of the system (26.1), let

$$\begin{aligned}
 -ax + yz &= 0 \\
 by - xz &= 0 \\
 -cz + yw &= 0 \\
 -dw - xy &= 0
 \end{aligned}
 \tag{26.2}$$

Obviously, $E_0(0, 0, 0, 0)$ is one of the equilibrium points.

According to Eq. (26.2), we have

$$ax^2 = by^2
 \tag{26.3}$$

$$z = -\frac{ax^3}{bcd}
 \tag{26.4}$$

By solving the equations, we get the following nonzero equilibrium points $E_1 \left(\sqrt[6]{\frac{b^3c^2d^2}{a}}, -\sqrt[3]{acd}, -\sqrt{ab}, \sqrt[6]{\frac{ab^3c^4}{d^2}} \right)$ and $E_2 \left(-\sqrt[6]{\frac{b^3c^2d^2}{a}}, -\sqrt[3]{acd}, \sqrt{ab}, -\sqrt[6]{\frac{ab^3c^4}{d^2}} \right)$.

For non-zero equilibrium points $E(x, y, z, w)$, its corresponding Jacobian matrix is

$$J_E = \begin{bmatrix} -a & z & y & 0 \\ -z & b & -x & 0 \\ 0 & w & -c & y \\ -y & -x & 0 & -d \end{bmatrix} \tag{26.5}$$

Let $|\lambda I - J_E| = 0$, we get

$$\begin{aligned} & [(\lambda + a)(\lambda - b) + z^2](\lambda + c)(\lambda + d) \\ & + (\lambda + 2d)(\lambda + 2a)xw + (\lambda - 2b)y^3 = 0 \end{aligned} \tag{26.6}$$

For E_1 and E_2 , Eq. (26.6) can be transformed into the following equation:

$$\begin{aligned} & \lambda(\lambda + a - b)(\lambda + c)(\lambda + d) + (\lambda + 2d)(\lambda + 2a)bc \\ & - (\lambda - 2b)acd = 0 \end{aligned} \tag{26.7}$$

$$\begin{aligned} & \lambda^4 + (a - b + c + d)\lambda^3 + (ac + ad - bd + cd)\lambda^2 \\ & + (2abc + bcd)\lambda + 6abcd = 0 \end{aligned} \tag{26.8}$$

Both of the non-zero equilibrium points E_1 and E_2 have the identical eigenvalues that can be calculated by Eq. (26.8). When a, b, c , and d equal to 4, 1, 2, and 0.5 separately, we have the eigenvalues: $\lambda_1 = -3, \lambda_2 = -2.5525, \lambda_{3,4} = 0.0263 \pm 1.7702i$. That is to say, the equilibria E_1 and E_2 are unstable.

Theorem 3.1 *If $(a, b, c) = (4, 1, 2)$, the two nonzero equilibrium points of the system (26.1) is stable when $d \in (0, 0.4670)$.*

Proof Let $u_1 = a - b + c + d, u_2 = ac + ad - bd + cd, u_3 = 2abc + bcd, u_4 = 6abcd, H_1 = \begin{vmatrix} u_1 & 1 \\ u_3 & u_2 \end{vmatrix}, H_2 = \begin{vmatrix} u_1 & 1 & 0 \\ u_3 & u_2 & u_1 \\ 0 & u_4 & u_3 \end{vmatrix}$.

According to the Routh-Hurwitz criteria, we know that all Eq. (26.8) roots are of negative real parts if and only if $H_1 > 0$ and $H_2 > 0$. Therefore, the nonzero equilibria of system (26.1) are stable if and only if $H_1 > 0$ and $H_2 > 0$.

If $(a, b, c) = (4, 1, 2)$, we have $u_1 = 5 + d, u_2 = 8 + 5d, u_3 = 16 + 2d$, and $u_4 = 48d$. When $0 < d < 0.4670$, we have $H_1 > 0$ and $H_2 > 0$. Therefore, when (a, b, c) is $(4, 1, 2)$ and $d \in (0, 0.4670)$, the system (26.1) possesses two stable nonzero equilibrium points.

Similarly, according to mathematical analyses, we can get the following three important remarks.

Remark 3.1 When $(b, c, d) = (1, 2, 0.5)$ and $a > 4.433$, the nonzero equilibria of the system (26.1) are stable.

Remark 3.2 When $(a, c, d) = (4, 2, 0.5)$ and $0 < b < 0.8660$, the nonzero equilibria of the system (26.1) are stable.

Remark 3.3 When $(a, b, d) = (4, 1, 0.5)$ and $c > 2.167$, the nonzero equilibria of the system (26.1) are stable.

For the equilibrium point $E_0(0, 0, 0, 0)$, its corresponding Jacobian matrix is

$$J_{E_0} = \begin{bmatrix} -a & 0 & 0 & 0 \\ 0 & b & 0 & 0 \\ 0 & 0 & -c & 0 \\ 0 & 0 & 0 & -d \end{bmatrix} \tag{26.9}$$

Let $|\lambda I - J_{E_0}| = 0$, we get the following eigenvalues: $\lambda_1 = -a$, $\lambda_2 = b$, $\lambda_3 = -c$, $\lambda_4 = -d$. As long as all of the parameters a, b, c , and d are positive, $E_0(0, 0, 0, 0)$ is always unstable.

26.3.2 Lyapunov Exponent Spectra and Bifurcation Diagrams

Lyapunov exponent shows the average exponential rates of divergence or convergence of adjacent orbits in phase space of a dynamic system. A system is chaotic if it possesses at least one positive Lyapunov exponent.

The author calculated the Lyapunov exponents by the use of the Wolf algorithm [24]. When (a, b, c, d) is equal to $(4, 1, 2, 0.5)$, the Lyapunov exponents are $\lambda_{L1} = 0.3689$, $\lambda_{L2} = 0.0002$, $\lambda_{L3} = -0.6792$, $\lambda_{L4} = -5.190$, and its Lyapunov dimension is 2.543. When (a, b, c, d) is $(1, 0.8, 2, 0.5)$, the Lyapunov exponents are $\lambda_{L1} = 1.117$, $\lambda_{L2} = 0.3126$, $\lambda_{L3} = -0.02965$, $\lambda_{L4} = -4.099$, and its Lyapunov dimension is 3.3414. The attractor of the system (26.1) is of fractional dimension.

Figures 26.4, 26.5, 26.6, and 26.7 show curves of two largest Lyapunov exponents of the system (26.1) versus a, b, c, d separately. From Fig. 26.7a–d, we notice that, when $(a, b, c) = (4, 1, 2)$, the system exhibits chaotic behaviors in a large field of $d \in [0.47, 160]$.

Figure 26.8 indicates the bifurcation diagram of the system (26.1) changing with a , when $(b, c, d) = (1, 2, 0.5)$. By comparing Fig. 26.4 with Fig. 26.8, we can see that there are three periodic windows $[2.72, 2.80]$, $[3.32, 3.41]$, and $[4.14, 4.17]$ in the field of $a \in [3, 5]$. When $a > 4.43$, the system is not chaotic, that is consistent with Remark 3.1 in Sect. 3.1.

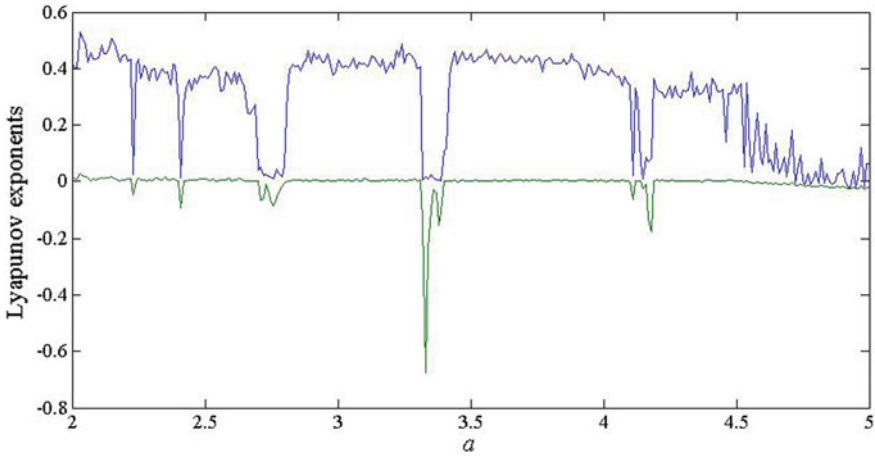


Fig. 26.4 Two Lyapunov exponents varying with a , when $(b, c, d) = (1, 2, 0.5)$

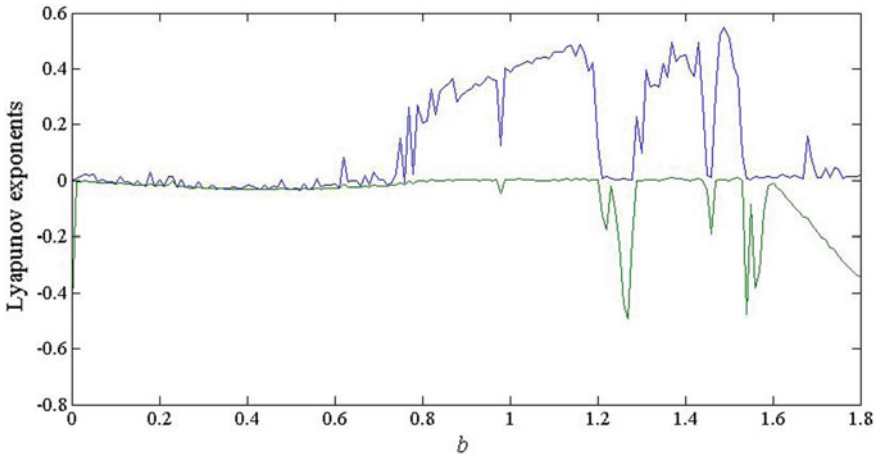


Fig. 26.5 Two Lyapunov exponents varying with b , when $(a, c, d) = (4, 2, 0.5)$

Figure 26.9 shows the bifurcation diagram of system (26.1) changing b , when (a, c, d) equals to $(4, 2, 0.5)$. According to Figs. 26.5 and 26.9, there exists one periodic window $[1.21, 1.28]$ in the field of $b \in [0.9, 1.5]$. When $b < 0.86$, the system is not chaotic that coincides with Remark 3.2 in Sect. 3.1.

Figure 26.10 is the bifurcation diagram changing with c , when $(a, b, d) = (4, 1, 0.5)$.

Figure 26.11a–d shows the bifurcation diagrams when d varies in a wide range of $[0, 160]$. The system is chaotic when d changes in a large field. Besides, according to Theorem 3.1, the system is not chaotic when $d < 0.467$, that is consistent with Fig. 26.11a.

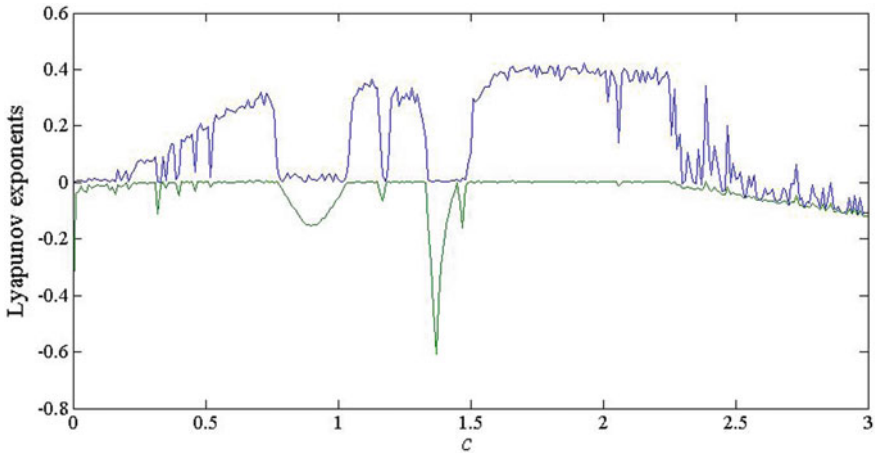


Fig. 26.6 Two Lyapunov exponents varying with c , when $(a, b, d) = (4, 1, 0.5)$

26.3.3 Time Domain Waveform and Frequency Spectrum

Let the parameters (a, b, c, d) be $(4, 1, 2, 0.5)$, the author calculated $x(t)$ of the system (1) with initial state $(x_0, y_0, z_0, w_0) = (10, 10, 10, 10)$ and $x(t)$ with initial state $(x_0, y_0, z_0, w_0) = (10.1, 10, 10, 10)$, shown in Fig. 26.12. We can see from Fig. 26.12 that the time waveform is highly sensitive to minor difference of initial states.

Frequency spectrum $\text{Log}|x| \sim f$ of the system is demonstrated in Fig. 26.13, where the bandwidth is about 0–60 Hz.

26.4 Conclusions

A novel 4D autonomous chaotic system with four linear terms and four cross products is studied. The system is hyper-chaotic within some range of parameters. For example, if the parameters (a, b, c, d) equal to $(10, 4, 5, 3)$, the system is hyper-chaotic because its Lyapunov exponents are 1.1997, 0.028110, -0.59293 , -14.635 . Besides, the system is also hyper-chaotic if $(a, b, c, d) = (10, 4, 5, 30)$. And, it is still chaotic when $d = 92$. Since the 4D system can keep chaotic when one of its parameters varies in a large field, it is expected to be adopted in a variety of engineering applications.

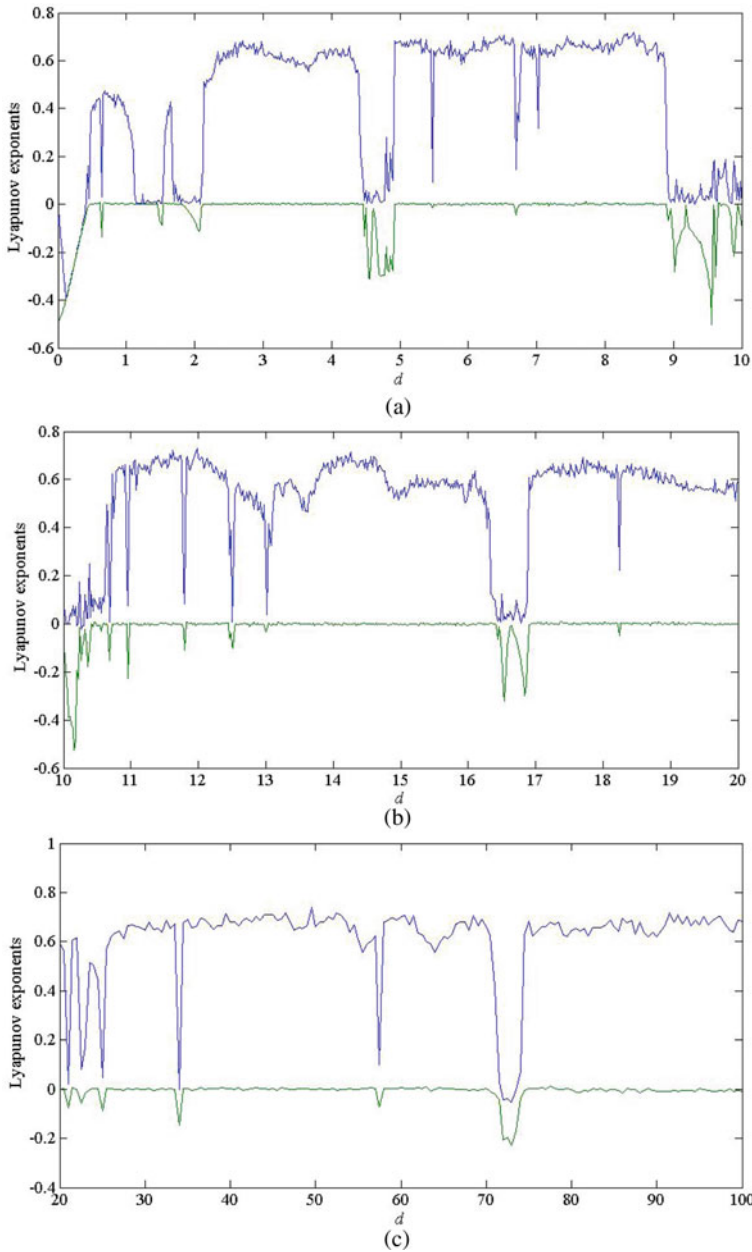


Fig. 26.7 Two Lyapunov exponents varying with d , when $(a, b, c) = (4, 1, 2)$: **a** $d \in [0, 10]$; **b** $d \in [10, 20]$; **c** $d \in [20, 100]$; **d** $d \in [100, 160]$

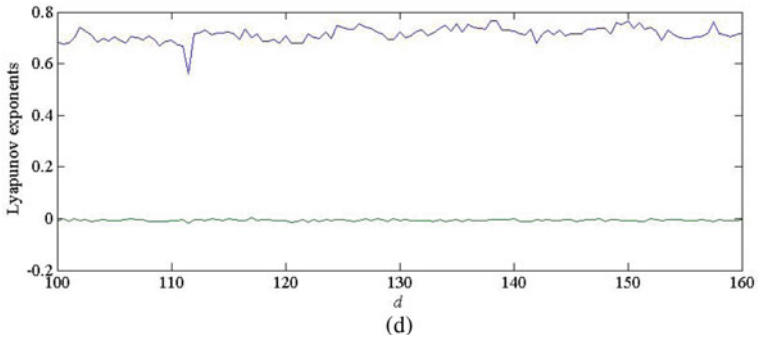


Fig. 26.7 (continued)

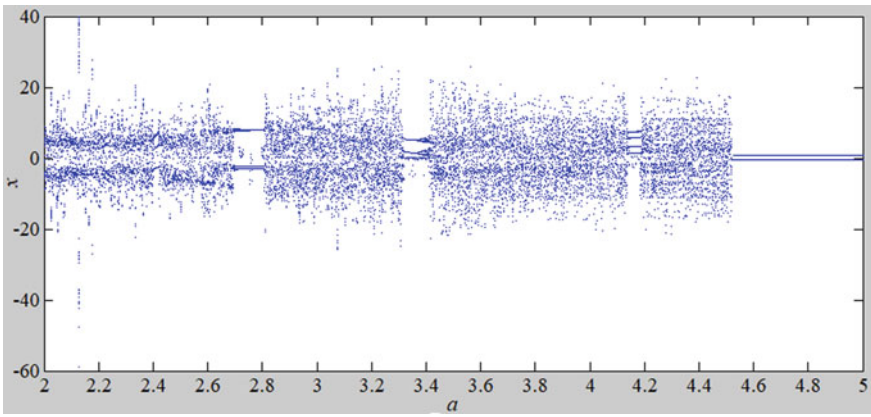


Fig. 26.8 Bifurcation diagram varying with a , when $(b, c, d) = (1, 2, 0.5)$

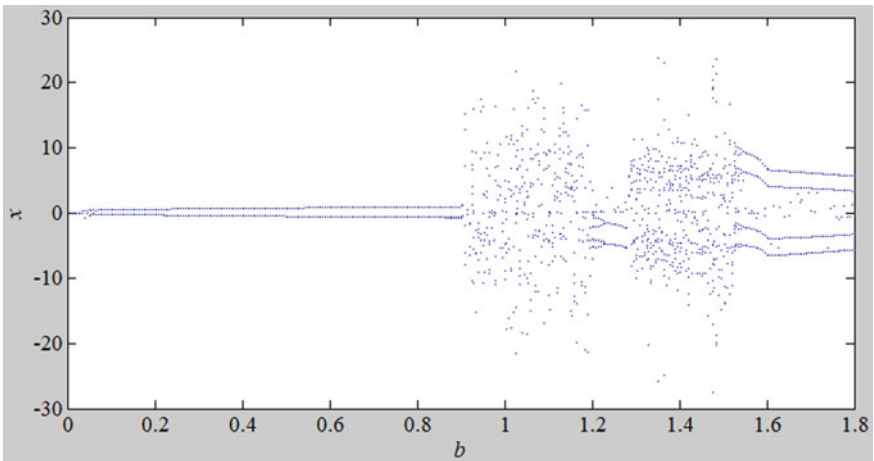


Fig. 26.9 Bifurcation diagram varying with b , when $(a, c, d) = (4, 2, 0.5)$

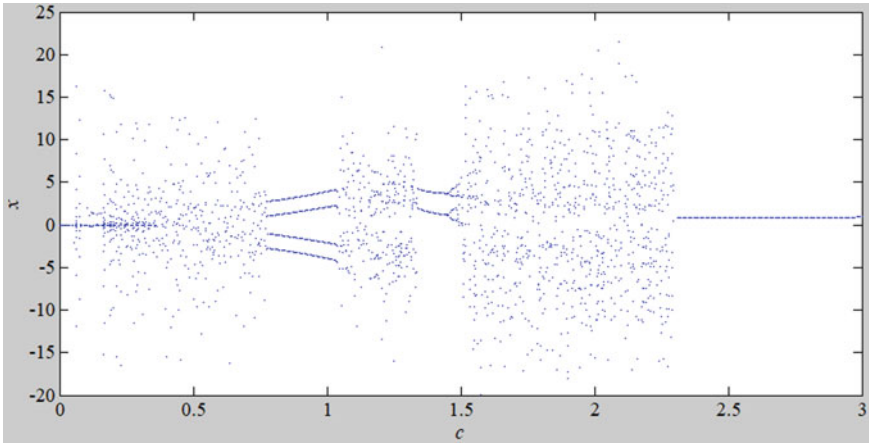
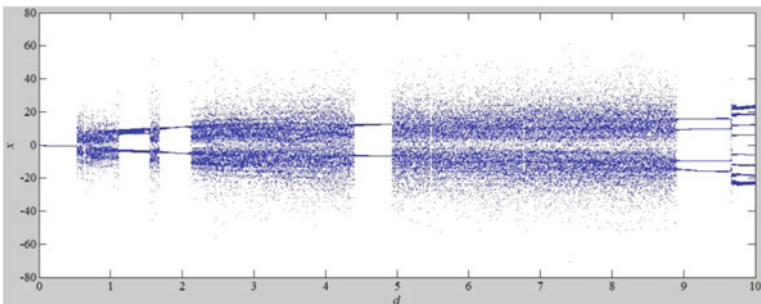
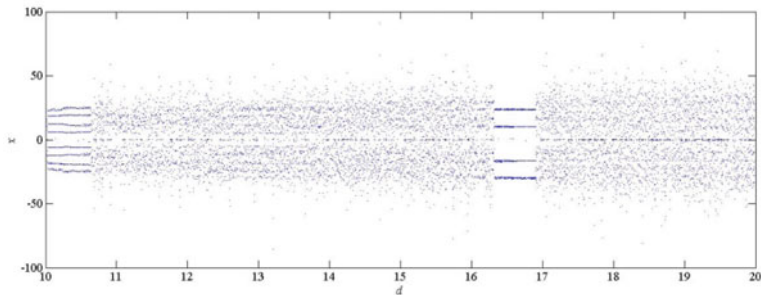


Fig. 26.10 Bifurcation diagram varying with c , when $(a, b, d) = (4, 1, 0.5)$

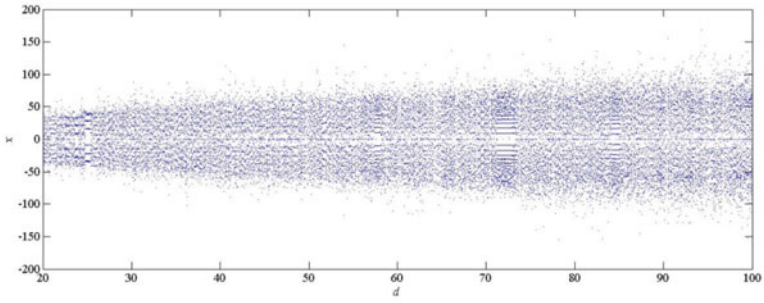


(a)

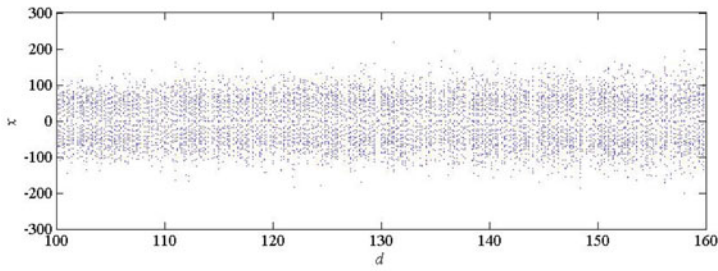


(b)

Fig. 26.11 Bifurcation diagrams varying with d , when $(a, b, c) = (4, 2, 1)$: **a** $d \in [0, 10]$; **b** $d \in [10, 20]$; **c** $d \in [20, 100]$; **d** $d \in [100, 160]$



(c)



(d)

Fig. 26.11 (continued)

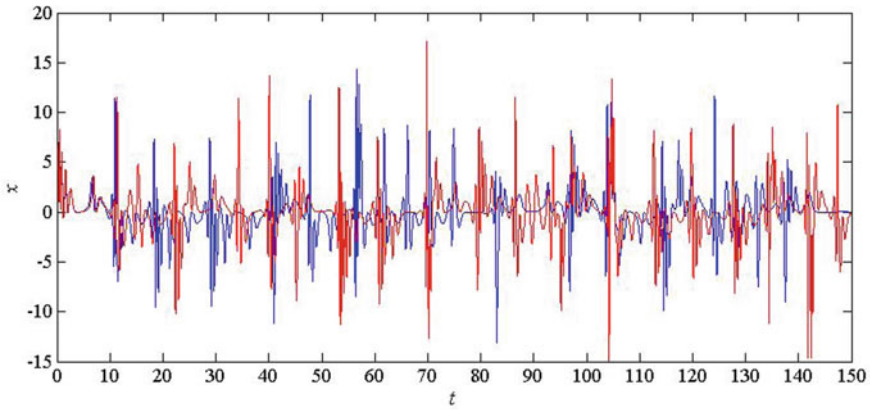


Fig. 26.12 $x(t)$ waveform: red line for initial state (10, 10, 10, 10); blue line for initial state (10.1, 10, 10, 10)

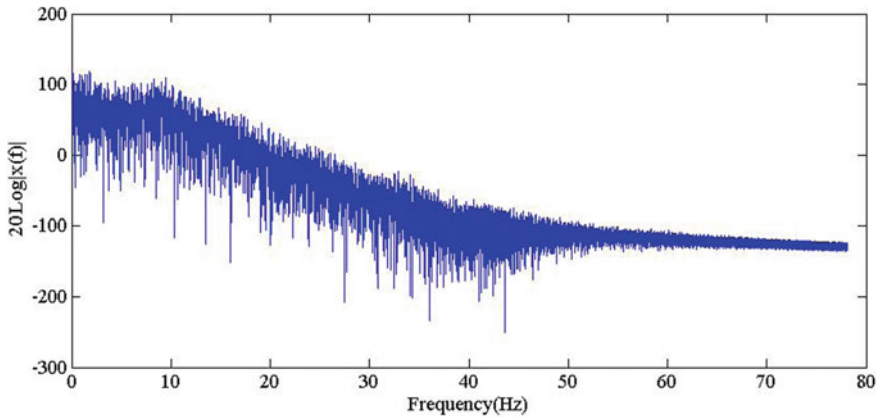


Fig. 26.13 Frequency spectrum of the system

Acknowledgements This work is funded by the National Natural Science Foundation of China (No. 61501206).

The author is grateful to the paper reviewers for their worthy comments and suggestions.

References

1. Lorenz, E.: Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141 (1963)
2. Rossler, O.: An equation for continuous chaos. *Phys. Lett. A* **57**, 397–398 (1976)
3. Chen, G., Ueta, T.: Yet another chaotic attractor. *Int. J. Bifurcat. Chaos* **9**, 1465–1466 (1999)
4. Edward, O.: *Chaos in Dynamical Systems*, 2nd edn. Cambridge University Press, Cambridge (2002)
5. Lü, J., Chen, G.: A new chaotic attractor coined. *Int. J. Bifurc. Chaos* **3**, 659–661 (2002)
6. Liu, C., Liu, T., Liu, L., Liu, K.: A new chaotic attractor. *Chaos Soliton. Fract.* **22**, 1031–1038 (2004)
7. Dadras, S., Reza, M., Qi, G.: Analysis of a new 3D smooth autonomous system with different wing chaotic attractors and transient chaos. *Nonlinear Dyn.* **62**, 391–405 (2010)
8. Liu, J., Zhang, W.: A new three-dimensional chaotic system with wide range of parameters. *Optik* **124**, 5528–5532 (2013)
9. Evangelista, J.V., Artiles, J.A., Chaves, D.P., Pimentel, C.: Emitter-coupled pair chaotic generator circuit. *AEU Int J Electron Commun* **77**, 112–117 (2017)
10. Feng, C., Cai, L., Kang, Q., Wang, S., Zhang, H.: Novel hyperchaotic system and its circuit implementation. *J. Comput. Nonlinear Dyn.* **10**, 061012–061019 (2015)
11. Yu, S., Tang, W.K.S., Lü, J., Chen, G.: Generation of $n \times m$ -Wing Lorenz-like attractors from a modified Shimizu-Morioka model. *IEEE Trans. Circ. Syst.-II* **55**, 1168–1172 (2008)
12. Peng, D., Sun, K., He, S., Zhang, L., Abdulaziz, O.A.: Numerical analysis of a simplest fractional-order hyperchaotic system. *Theor. Appl. Mech. Lett.* **9**, 220–228 (2019)
13. Kim, D., Chang, P.H.: A new butterfly-shaped chaotic attractor. *Results Phys.* **3**, 14–19 (2013)
14. Li, Q., Zeng, H., Li, J.: Hyperchaos in a 4D memristive circuit with infinitely many stable equilibria. *Nonlinear Dyn.* **79**, 2295–2308 (2014)
15. Zhang, C., Yu, S.: Generation of grid multi-scroll chaotic attractors via switching piecewise linear controller. *Phys. Lett. A* **374**(30), 3029–3037 (2010)

16. Li, C., Li, H., Li, W., Tong, Y., Zhang, J., Wei, D., Li, F.: Dynamics, implementation and stability of a chaotic system with coexistence of hyperbolic and non-hyperbolic equilibria. *AEU Int. J. Electron Commun* **84**, 199–205 (2018)
17. Zhou, P., Yang, F.: Hyperchaos, chaos, and horseshoe in a 4D nonlinear system with an infinite number of equilibrium points. *Nonlinear Dyn.* **76**, 473–480 (2014)
18. Liu, J.: A four-wing and double-wing 3D chaotic system based on sign function. *Optik* **125**, 7089–7095 (2014)
19. Liu, J., Qu, Q., Li, G.: A new six-term 3-D chaotic system with fan-shaped Poincaré maps. *Nonlinear Dyn.* **82**, 2069–2079 (2015)
20. Singh, J., Roy, B.: Analysis of a one equilibrium novel hyperchaotic system and its circuit validation. *Int. J. Control Theory Appl* **8**, 1015–1023 (2015)
21. Zhang, S., Zeng, Y., Li, Z., Wang, M., Xiong, L.: FPGA-based implementation of different families of fractional-order chaotic oscillators applying Grünwald-Letnikov method. *Commun. Nonlinear Sci. Numer Simulat.* **72**, 516–527 (2019)
22. Sprott, J.: A proposed standard for the publication of new chaotic systems. *Int. J. Bifurc. Chaos* **21**, 2391–2394 (2011)
23. Sajad, J., Sprott, J., Molaie, M.: A simple chaotic flow with a plane of equilibria. *Int. J. Bifurc. Chaos* **26**, 1650098–1650104 (2016)
24. Wolf, A., Swift, J., Swinney, H., Vastano, J.: Determining Lyapunov exponents from a time series. *Phys D* **16**, 285–317 (1985)

Chapter 27

Influence of the Optical Aspects of Photographic Composition on the User Experience in the Issues Related to Decision Making, Choices and Level of Visual Comfort



Marcia Campos and Fabio Campos

Abstract This article investigates the level of influence of the optical aspects of image composition on the issues and consequent problems of the user experience. This work is part of the research project n. 461710-JOR-013-2019/1 (registered in the Unicap Research Database and approved by the CCP at an ordinary meeting on 09/October/2018) “Virtual environments: influence of the optical aspects of image composition on the quality of user experience” coordinated by Dr. Marcia Campos. The objective was to collect data relative to the perception/reaction of the user versus levels of optical contrasts of the image composition. Therefore, to investigate this level of influence, an experiment was carried out in two stages: the first step consisted of producing 17 images with specific characteristics of visual representation covering various levels of contrasts of the optical aspects for each image. The second step was the application of questionnaires with users, with questions based on the images produced in the first stage. The results indicate evidences of the correlation of the optical aspects of image composition with the reactions choices from the users; such as, time spent for decision making and choices about the images regarding elements of the scene and level of visual and physical discomfort. The evidence point to the results that optical aspects of image composition, mainly the representation of its levels of contrasts, are relevant for the quality of the user experience concerning both physical and psychological perception issues.

M. Campos (✉)

Catholic University of Pernambuco, Recife, PE 50050-900, Brazil
e-mail: spot4m@gmail.com

F. Campos

Federal University of Pernambuco, Recife, PE 50670-901, Brazil

© Springer Nature Singapore Pte Ltd. 2021

S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation, Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_27

27.1 Introduction

The fast technological development of virtual reality devices and applications is expanding beyond the boundaries of video games, with demands from many areas such as healthcare, construction, auto industry, etc.

However, virtual reality devices face a number of issues regarding the user experience. For example, the uncanny valley phenomenon, break of sense of presence, and the undesirable symptoms caused by cybersickness. According to studies it is estimated that about 30–80% of users have suffered from symptoms of cybersickness. These annoyances are caused by interacting with virtual environments and can persist from a few minutes to even days after the use of the immersive devices [1]. Symptoms are eye fatigue, headache, disorientation, nausea and may even lead to vomiting. Although studies on cybersickness have advanced, the reasons that cause these symptoms in users of virtual reality devices are still mostly unknown [1, 2].

The visual aspects of the virtual environment impact the user experience, ranging from objective issues such as the perception of optical aspects of light/image, to subjective issues such as the interpretation and significance of this light/image [3]. With this in sight is plausible to verify what optical elements of conception and composition of the virtual environments are crucial factors to the effects provoked in the users.

Studies have found a relationship between the variation of the field of view angle and the level of cybersickness symptoms experienced by users. That is, the larger the field of view angle is, the effects of symptoms are proportionally stronger, and the smaller the field of view angle is, the symptoms become gradually softer [2, 4].

In this paper we verify the relationship between the data obtained on the influence of the field of view angle with the level of discomfort of cybersickness symptoms, and the studies on the optical aspects of image composition via their contrast levels. This method allows a deeper analysis of the preliminary information found on our previous researches [3, 5].

According to Campos [3] the representation of contrast levels of the optical aspect of depth of field or sharpness changes the contrast levels of the images as the field of view or framing angle changes. Then, as the field of view angle increases, the sharpness contrast becomes gradually lower or lighter, while as the field of view angle decreases, the sharpness contrast proportionally becomes higher or stronger [2, 6].

To make it clearer, the greater the angle of the field of view, the level of representation of the optical sharpness of the image becomes proportionally more subtle (the whole scene around the perfectly sharp object is just slightly blurred). And vice versa, the smaller the field of view angle, the level of representation of the optical sharpness of the image becomes gradually stronger (every scene around the object that is perfectly sharp is extremely blurred).

Despite technological advances producing increasingly realistic virtual images, those more realistic images are a major cause of the symptoms of cybersickness. The more realistic the virtual environment is, more it is prone to the cybersickness

[2]. Studies also confirm the direct relationship between the quality level of the representation of optical aspects of image composition via their contrasts, with the user's perception and interpretation of the level of image realism [5]. Thus, the relationship between the optical aspects of image composition via their contrasts, and the problems of virtual reality devices.

27.1.1 *Optical Aspects of Image Composition by Contrasts—Campos Method*

This method offers a new approach which allows more formal and objective observation and understanding about the optical effects that constitute the images (real or virtual images). It allows to do the processes of conception, analysis and understanding of images focusing on the objective aspects of light or image through their representation via contrasts.

The method mimics the way the human eyes perceive the images, that is, all the objects we see in our visual field, be them people, animals, things, colors, etc., absolutely everything is visual information or reflected light. Only after the capture of this light (the image) does the psychological and rational step of giving meaning to the captured light begins.

Within the logic of optics, the image is the result of the propagation and reflection of light in the environment and, since light is the image itself, all the optical aspects that form it are also elements of light.

Campos's [3] method is based on the analogy with the Contrast Theory developed by Itten [7, 8], this theory organizes colors via their contrast representations. Through the analogy with the Contrast Theory, Campos' method considered the representation of the optical aspects of the image as elements originating from light. That is, they are perceived via levels of contrasts or intensity, following the same gradient logic existing from the light source (white) to its absence (black) (see Fig. 27.1).

In this way, each optical aspect is represented and perceived through its variation of levels or intensity of contrasts.

The optical aspects of image composition are:

Depth-of-Field Optical Contrasts: that is the perception of optical sharpness of contrasts in the image. This optical aspect is represented by varying its contrast levels, which comprise a continuous gradient (in the same logic as the light propagation flow), ranging from the low depth-of-field contrast to its opposite extreme which is the High depth-of-field contrast (see Fig. 27.2).



Fig. 27.1 Representation of the levels of contrast of intensity of colors

Fig. 27.2 Representation of the levels of contrast of depth of field

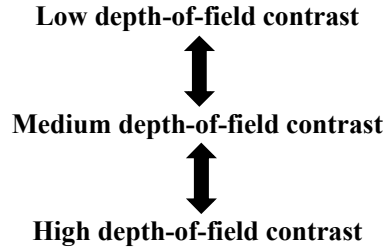
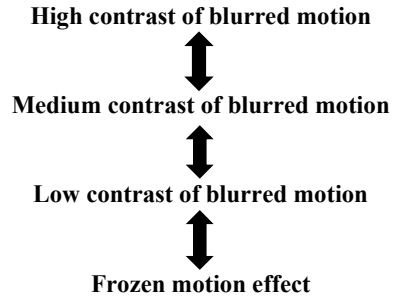


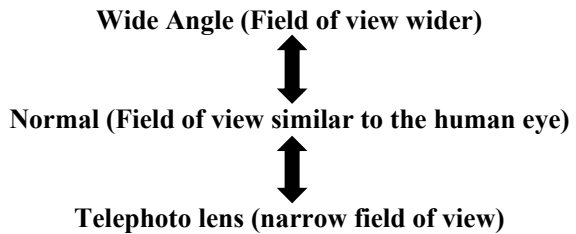
Fig. 27.3 Representation of the levels of contrast of motion effect



Optical Contrasts of Motion Effect: it is the perception of optical contrasts originated in the reflection of light coming from scenarios or objects that are in motion. This optical aspect is represented by varying its contrast levels that encompass a continuous gradient (in the same logic as the flow of light propagation), ranging from the high contrast of blurred motion effect to its opposite extreme, which is the Frozen motion effect (see Fig. 27.3). And always going through the medium contrast.

Optical Angle of View Contrasts (aka Lenses): Corresponds to the perception obtained from the size of the framed area by varying the degrees of the visual pyramid or the field of view. This optical aspect is represented by varying its contrast/degree levels, which encompass a continuous gradient (in the same logic as the light propagation flow), ranging from the widest field of view of the Wide Angle to its extreme opposite which is the narrow field of view of the telephoto lens (see Fig. 27.4).

Fig. 27.4 Representation of the levels of contrast of angle of view



Analyzing the image by its decomposition on the elements of contrast described so far, this method allows a better understanding of the image, from its objective information to its consequent subjective information. A wide range of visual information that contributes to the scientific enrichment of evaluation processes and image design solutions. Mainly in the perception and understanding of how the optical aspects of the image combined in image composition techniques impact the emotional and physical interpretation of the user. Bringing an innovative point of view that enables from the logic of contrasts, more accurate and formal analyzes of image quality, the measurement of its level of realism, greater efficiency in the process of image design and consequently its impact on the user experience. Positively reflecting on the process of creating the visual aesthetics of scenarios whether real or virtual. More efficiently and safely make visual artifact design decisions.

Therefore, this paper investigates the level of influence of optical aspects of image composition on issues and consequent problems inherent to the user experience. This work is part of a bigger research project “Virtual environments: influence of the optical aspects of photographic/image composition on the quality of the user experience”. The objective is to collect data regarding the perception and reaction of the users when observing the optical contrast levels of the image composition. The collected data focused on the user’s perception and reaction are: time taken for decision making, motivation for choosing objects in the scene and level of visual discomfort. The experiment was carried out in 2 steps.

27.2 First Step of the Experiment

The first stage consisted of the production of 17 photographic images within the visual representation characteristics of the contrast gradient levels of the optical aspects of photographic composition based on the Campos method [3–5, 9].

For this the 17 images were composed as follows:

- Three totally identical objects (plastic cups) were used in all pictures, without any kind of treatment or post production.
- The images were produced based on the composition method, technique and photographic equipment in order to produce images with the exact optical contrast levels desired [3].
- Three images were produced with the **high contrast of depth of field, or sharpness** (framed with telephoto—147 mm and Aperture f5.6). One photo with exact focus on the object in front (see Fig. 27.5), another photo with exact focus on the object in the middle (see Fig. 27.6), and another photo with exact focus on the back object (see Fig. 27.7).



Fig. 27.5 High contrast of sharpness—focus on the front



Fig. 27.6 High contrast of sharpness—focus on the middle

Three images were produced with the **low contrast of depth of field or sharpness** (framed at wide angle—17 mm and Diaphragm Aperture f5.6). One photo with exact focus on the object in front (see Fig. 27.8), another photo with exact focus on the object in the middle (see Fig. 27.9), and another photo with exact focus on the object on the back (see Fig. 27.10).

Three photos were produced with the **medium contrast of depth of field or sharpness** (framed with normal lens - 45 mm and Aperture f5.6). One photo with exact focus on the object in front (see Fig. 27.11), another photo with exact focus on



Fig. 27.7 High contrast of sharpness—focus on the back



Fig. 27.8 Low contrast of sharpness—focus on the front

the object in the middle (see Fig. 27.12), and another photo with exact focus on the object on the back (see Fig. 27.13).

Three photographs were produced **with total absence of sharpness** (framed with 120, 72 and 45 mm lenses). Being a photo with very strong blur effect (see Fig. 27.14), another photo with medium blur level (see Fig. 27.15) and another photo with slight blur (see Fig. 27.16).

Five photographs were produced with the optical aspect of motion effect contrast variation (exposure time—longer for blur effect and shorter for frozen effect x subject movement speed variation). Being a high contrast blurred motion effect photo (see



Fig. 27.9 Low contrast of sharpness—focus on the middle



Fig. 27.10 Low contrast of sharpness—focus on the back

Fig. 27.17), another medium blurred contrast effect photo (see Fig. 27.18), another low contrast blurred motion effect photo (see Fig. 27.19), another effect photo motion blur (see Fig. 27.20) and another high-contrast panning motion blur photo (See Fig. 27.21).



Fig. 27.11 Medium contrast of sharpness—focus on the front



Fig. 27.12 Medium contrast of sharpness—focus on the middle



Fig. 27.13 Medium contrast of sharpness—focus on the back



Fig. 27.14 Strong blur—fully without focus



Fig. 27.15 Medium blur—fully without focus



Fig. 27.16 Light blur—fully without focus



Fig. 27.17 High contrast blurred motion



Fig. 27.18 Medium contrast blurred motion



Fig. 27.19 Low contrast blurred motion



Fig. 27.20 Frozen motion contrast



Fig. 27.21 High contrast blurred panning motion

27.3 Second Step of the Experiment

This stage consisted on the application of questionnaires, with questions based on the photos produced in the first stage of the experiment. The purpose of the questionnaire was to collect data regarding the user's perception and reaction versus the variation of optical contrast levels in the photos. Regarding the time taken for decision making, motivation of the choice of objects in the scene and level of visual and physical discomfort felt.

The questionnaire had 17 questions with one photo in each. Where, in the first nine questions I asked participants to simply choose one of the cups at their wish, which one would be their personal choice. Questions 10, 11 and 12 were aimed at verifying the user's level of discomfort by observing the fully blurred photos for as long as they could and after choosing from five rating levels from very comfortable to very uncomfortable. And the last questions aimed to ask the level of speed, stillness or immobility perceived from the observed photos.

Here are three examples of questions used in the questionnaire:

1. **To answer the question below use your vision naturally to observe the photographs. Now choose one of the cups in the image below**



- Front Cup
 Middle Cup
 Back Cup
12. **To answer this question you should be looking at the photo without looking away from it for as long as you can stand and/or feel any visual discomfort. Indicate the level of comfort you felt**



- Very comfortable
 Uncomfortable
 Neutral (indifferent)
 Slightly uncomfortable
 Very uncomfortable

12. Now indicate what sensation or impression you noticed in the photo below



- Fully stationary/immobile
- Too slow/lagging
- Neutral
- Very fast/rapid
- Extremely fast/rapid.

27.4 Results and Analysis of the Experiment

The questionnaire was posted online and was answered by 73 people, mostly, university undergraduate students from diverse institutions and courses. The results were compiled in 5 tables, each one tabulating a different aspect, such as: Table 27.1 summarizing the preference of the level of focus in high contrast of depth of field or sharpness images, Table 27.2 the same focus preference but in images with low contrast of depth of field or sharpness, Table 27.3 regarding images with medium contrast of depth of field or sharpness, Table 27.4 related to totally unfocused photos,

Table 27.1 Results of questions (1, 2, and 3)—images with high contrast of depth of field or sharpness

Answer alternatives	Question 1. Focus on the front cup	Question 2. Focus on the middle cup	Question 3. Focus on the back cup
Front cup (%)	91.6	12.3	5.6
Middle cup (%)	4.2	86.2	5.6
Back cup (%)	4.2	1.5	88.8

The purpose of using boldface is to highlight the percentage result in the alternatives considered most appropriate

Table 27.2 Results of questions (4, 5, 6)—images with low contrast of depth of field or sharpness

Answer alternatives	Question 4. Focus on the middle cup	Question 5. Focus on the front cup	Question 6. Focus on the back cup
Front cup (%)	63.4	91.6	54.9
Middle cup (%)	33.8	5.6	31.0
Back cup (%)	2.8	2.8	14.1

Table 27.3 Results of questions (7, 8, and 9)—images with medium contrast of depth of field or sharpness

Answer alternatives	Question 7. Focus on the back cup	Question 8. Focus on the front cup	Question 9. Focus on the middle cup
Front cup (%)	35.2	88.7	47.9
Middle cup (%)	43.7	9.9	43.6
Back cup (%)	21.1	1.4	8.5

Table 27.4 Results of questions (10, 11, and 12) with totally unfocused images, through optical variations of blur contrast levels

Answer alternatives	Question 10. Medium contrast fully blurred	Question 11. Between medium to high contrast fully blurred	Question 12. High contrast fully blurred
Very comfortable (%)	4.2	4.2	1.4
Uncomfortable (%)	5.6	2.8	8.3
Neutral (indifferent) (%)	31.9	31.9	9.7
Slightly uncomfortable (%)	41.7	43.0	20.8
Very uncomfortable (%)	16.6	18.1	59.8

through optical variations of blur contrast levels, and Table 27.5 the same analysis but with images with variations of optical contrasts of motion effect.

The results summarized on each table will be discussed below.

27.4.1 Data Analysis of Table 27.1

Following are the discussion based on the results from Table 27.1:

- The majority of people (91.5, 86.1, 88.7%) chose the glasses that were exactly sharp on three cases.

Table 27.5 Results of questions with images produced with variations of optical contrasts of motion effect

Answer alternatives	Question 13. High contrast blurred	Question 14. Medium contrast blurred	Question 15. Frozen motion effect	Question 16. Low contrast blurred	Question 17. High contrast blurred/panning
Fully stationary/immobile (%)	6.9	9.7	59.7	27.8	37.5
Too slow/lagging (%)	2.8	40.3	26.4	44.5	2.8
Neutral (%)	2.8	9.7	13.9	20.8	2.8
Very fast/rapid (%)	52.8	37.5	0.0	6.9	19.4
Extremely fast/rapid (%)	34.7	2.8	0.0	0.0	37.5

- The results pointed to certainty in the choice and speed of decision making by most participants.
- It was evident in photos produced with high contrast depth of field the need for human vision for the visual comfort that only perfect focus provides.
- The presence of the high-contrast depth-of-field optical aspect in the image composition that produces an extremely unfocused context around the object with the exact focus made it possible to respond with faster and more accurate decision making.
- It was found that the search factor via perfect focus, has relevant influence on decreasing the level of indecision in the user’s choice and, consequently, in the faster decision-making of the user. In short, this indicates that users spend less time observing and interacting in scenarios with high contrast of depth of field.

27.4.2 Data Analysis of Table 27.2

Following are the discussion based on the results from Table 27.2:

- The majority chose the cup with exact sharpness only in question 5.
- In questions 4 and 6 the results show the tendency to choose the cups by the configuration in the image composition. That is, by the position in the field of view.
- The results of questions 4 and 6 indicate a higher level of indecision and, consequently, slower decision-making by participants.

- It can be deduced, then, that the choice of cups through their position in the field of view tends to a higher level of indecision in the choice and, consequently, more time spent in the user's decision making.
- The extremely relevant results in question 5 indicate certainty of choice and faster decision-making by most participants. Thus, it can be concluded that the sum of the factors: exact sharpness in the object positioned in the foreground of the field of view, allows more certainty of choice and, consequently, faster decision-making of the user.
- It was also noticed evidence that in images with low contrast of depth of field or sharpness, the user experience tends first to observe the positioning of all objects in the field of view. This means that the user spends more time observing and interacting with all the elements that make up the scenario. Even though it is slightly unfocused.
- The lower percentage results obtained by the perfectly focused cups in questions 4 and 6 indicate that a significant number of users spent more time observing and interacting with objects that are out of focus.

27.4.3 Data Analysis of Table 27.3

Following are the discussion based on the results from Table 27.3:

- The majority chose the cup with exact sharpness only in question 8.
- In questions 7 and 9 the results show the tendency to choose the cups by the configuration in the image composition. That is, by the position in the field of view.
- The results of questions 7 and 9 indicate a higher level of indecision and, consequently, less rapid decision-making by participants.
- The extremely relevant results in question 8 indicate certainty of choice and faster decision-making by most participants. Thus, it can be concluded that the sum of the factors: exact sharpness in the object positioned in the foreground of the field of view, allows more certainty of choice and, consequently, faster decision-making of the user.
- The image with medium contrast of depth of field or sharpness in its composition produces a context with a median level of blur around the object with the exact focus. Thus, it can be deduced that this optical aspect causes a greater level of indecision in the choices and, consequently more time spent on decision making by the user.
- In images with medium contrast of depth of field or sharpness, the user experience tends first to observe the position of objects in the field of view, and then to "search" for the exact focus. However, because the depth-of-field medium contrast aspect had intermediate sharpness gradient levels (between low and high contrasts), the percentages of choice of the precisely focused cups were slightly higher when compared to the results from Table 27.2.

- Due to the percentage increase of the answers by the cups with the exact focus, it can be deduced that both the level of indecision in the choice and the speed in the user's decision making improve when compared to the results of Table 27.2.

27.4.4 Data Analysis of Table 27.4

Following are the discussion based on the results from Table 27.4:

- All questions had percentage results relevant to the “slightly uncomfortable” and “very uncomfortable” response options.
- The results were extremely irrelevant for the “very comfortable” and “uncomfortable” options.
- The results show that the level of discomfort felt by users increases as the levels of optical contrasts in the out-of-focus range also increase.
- So, it can be considered that the observation time of fully blurred images that the user can handle, decreases as the levels of optical contrasts in the in the out-of-focus range increase.
- Therefore, it can be concluded that the level of discomfort is a relevant factor in the level of indecision of choice and the time spent in decision making in the user experience.
- Thus, the level of discomfort felt by the user interferes with the level of indecision of choice and, consequently, the time spent in decision making.
- Thus, the increased level of discomfort felt by the user, causes a decrease in the amount of time the user can observe a given image.

27.4.5 Data Analysis of Table 27.5

Following are the discussion based on the results from Table 27.5:

- The results were relevant and confirm the proportional relation of the visual sensation of movement speed variation felt by the users, with the variation of optical contrasts levels of movement effect perceived by them.
- In images with high contrast of blurred motion effect the vast majority opted for the “very fast” and “extremely fast” alternatives.
- Then we can confirm that the higher the optical contrast level of the blurred effect, the user's perception will be of higher movement speed. And so, successively, it follows the same logic with the other ranges of optical effect levels of motion contrast to the opposite end of the high blurred motion contrast, which is frozen motion.
- In question 17 we had results with equal percentages in the answers “fully stationary” and “extremely fast”, with 37.5% for each. This is a very intriguing result, as

the two response options are extremely opposite, since one refers to immobility and the other to rapid movement. That is, they are totally opposite perceptions, which characterize two types of emphasis on image observation: one was based on the object (glass) with the exact sharpness emphasizing the freezing perception. And the other was performed based on the perfect focus scenario that is with high contrast levels of the blurred motion effect, inducing the perception of high speed.

27.5 Conclusions

The results obtained in the experiment performed in the present work provides evidence about the existence of relevant influence of the optical aspects of image composition via contrasts in the questions inherent to the user experience.

It was possible to confirm by comparing the data in Tables 27.1, 27.2 and 27.3 that in the continuous gradient of the optical depth of field contrast (comprises from the low sharpness contrast to the opposite extreme which is the high sharpness contrast):

- The level of indecision at the user's choice increases as the continuous depth-of-field optical contrast gradient becomes lower. Consequently, the level of indecision in the user's choice decreases as the continuous depth-of-field optical contrast gradient becomes higher.
- -And with respect to the decision-making time spent by the user, it increases as the continuous depth-of-field optical contrast gradient gets lower. So, the decision time spent by the user decreases as the continuous depth-of-field optical contrast gradient gets higher.
- According to the data in Table 27.4, regarding the level of comfort felt by users, the condition of the presence of the element focused on the field of view to make the sensation of visual comfort possible is evident.
- The gradient level of the out-of-focus range profoundly impacts the level of visual discomfort felt by the user. The level of visual discomfort gradually increases as the level of blur contrast becomes high or strong. And consequently, the time spent observing decreases due to the increased level of discomfort felt by the user.
- Regarding the effects of the depth-of-field optical contrasts, the evidence from the experiment pointed to:
 - Images produced with high depth-of-field contrast make it possible to decrease the level of indecision in choosing and faster decision making. Because in images of this kind, the perfect-focus setting is extremely blurred, or extremely uncomfortable. Consequently, the user instantly seeks the focused element, which brings relief and total visual comfort.
 - Following the same logic in the opposite direction, images produced with low sharpness contrast greatly increase the level of indecision in choice and more time spent in decision making. Because, in this kind of image, the scenario around the perfect focus is subtly blurred, or slightly uncomfortable. Consequently, the user

can observe the entire context of the scenario for a longer time, and often does not perceive the element with exact sharpness.

Regarding the measurement of the decision making time, analyzed in this paper, it was made by comparing the statistical results collected in the questionnaire. Crossing the data from Table 27.4 with the results from Tables 27.1, 27.2 and 27.3.

- Where the level of comfort and discomfort felt by users (Table 27.4) reflects the ability to observe the image for longer or shorter time. So, decision making time is also affected by the level of discomfort felt.
- When we connect these assumptions Table 27.4 with the data in Table 27.1 it was found that images with high sharpness of contrast point to faster decision-making, due to the extremely unfocused environment around the object with exact focus (extremely uncomfortable environment around the comfortable area). Being reflected with great statistical significance.
- When we connect these assumptions in Table 27.4 with the data in Tables 27.2 and 27.3, it was found that images with low and medium sharpness point to less rapid decision making, due to the slightly and slightly blurred environments around the object with exact focus (slightly and moderately uncomfortable environments around the comfortable area). Being reflected with low statistical significance.
- That is, the measurement of decision making time is influenced by the level of visual discomfort felt by the user. And through the measurement made possible by the representations via contrast levels, coupled with the user experience, we can reach conclusions about time spent in decision making.

All the evidence point to the importance of optical aspects of image composition via contrasts in the context of perception from the users. For studies carried out so far [3–5, 7] point out its great importance for a deeper understanding of the causes of problems and, consequently, in the contribution of solutions. It also enables greater knowledge about imaging and how to enable measurement, such as the level of realism of a virtual environment via contrast levels of optical aspects. Avoiding the intuitive and abstract processes, which most often generate high uncertainty in decisions. On the other hand, the evidence from the experiment show that is possible to control or manipulate some perceptions by playing with the aspects of contrast of the images.

References

1. LaViola Jr., J.: A discussion of cybersickness in virtual environments. *SIGCHI Bull.* **32**(1), 47–56 (2000)
2. Tirol, A.: Effect of visual realism on cybersickness in virtual reality. IS Master's thesis, University of Oulu Faculty of Information Technology and Electrical Engineering (2018)
3. Campos, M.M.M.M.: A Fotografia na Concepção da Imagem dos Games. Universidade Federal de Pernambuco tese. Ph.D. <http://www.ufpe.br/sib/> (2014)
4. Kolasinski, E.: Simulator sickness in virtual environment. U.S. Army Research Institute for the Behavioral and Social Sciences (1995)

5. Campos, M., Campos, F., Van Gisbergen, M., Kovacs, M.: The relationship among the optical aspects of photographic composition and the quality, perception and interpretation of the realism in virtual images. In: Karwowski, W., Ahram, T. (eds.) IHSI 2019, AISC 903, pp. 1–6 (2019)
6. Davis, S., Nesbitt, K., Nalivaiko, E.: Comparing the onset of cybersickness using the oculus rift and two virtual roller coasters. In: 11th Australasian Conference on Interactive Entertainment (2015)
7. Birren, F.: The Elements of Color: A Treatise on the Color System of Johannes Itten Based on his Book the Art of Color. Van Nostrand Reinhold, New York (1970)
8. Freeman, M.: The Image: Collins Photography Workshop Series. William Collins, London (1988)
9. Campos, M.M.M.M., Teixeira, C., Carvalho, B.: Aspectos Fotográficos na Construção de Marcas: estratégias de aproximação com o consumidor. In: II Congresso Internacional de Marcas/Branding: conexões e experiências 2015 – Lajeado – RS. Anais do II Congresso de Marcas/Branding: conexões e experiências, 2015 – Lajeado: Ed da Univates, p. 277 (2016)

Chapter 28

Compositional Models: Iterative Structure Learning from Data



Václav Kratochvíl, Vladislav Bína, Radim Jiroušek, and Tzong-Ru Lee

Abstract Multidimensional probability distributions that are too large to be stored in computer memory can be represented by a compositional model—a sequence of low-dimensional probability distributions that when composed together try to faithfully estimate the original multidimensional distribution. The decomposition to the compositional model is not satisfactorily resolved. We offer an approach based on search traversal through the decomposable model class using likelihood-test statistics. The paper is a work sketch of the current research.

28.1 Introduction

Many real-life problems can be solved using a decomposing strategy, excellently summarized by George Pólya in his famous book [1]: *If you cannot solve a problem, then there is an easier problem you can solve: find it.* The basic idea is simple. A problem, or a complex system, can be decomposed into a sub-problems/subsystems that are easier to describe/understand. Unfortunately, the art of decomposing is not always straightforward.

Thanks to the massive use of computers, we have a huge amount of data in various areas of human activity. Using these data sets we can describe complex systems that

V. Kratochvíl (✉) · V. Bína · R. Jiroušek
Institute of Information Theory and Automation, Czech Academy of Sciences,
Prague, Czech Republic
e-mail: velorex@utia.cas.cz

V. Bína
e-mail: bina@fm.vse.cz

R. Jiroušek
e-mail: radim@utia.cas.cz

Faculty of Management, University of Economics, Prague, Czech Republic

T.-R. Lee
Department of Marketing, National Chung Hsing University, Taichung, Taiwan, ROC
e-mail: trlee@dragon.nchu.edu.tw

may appear as black-boxes to us. The key is to extract knowledge from the data set and use it as a support for future decision making or predictions.

A set of vital tools to work with large data-sets is accessible through a probability framework where records from the data set are considered to be realizations of random variables. In this paper, we assume problems/systems that can be described using a set of random variables. By an event, we understand a moment when we measure values of the random variables and we assume the existence of a data set with records of such measurements in history. As a typical example can serve a patient in a hospital with a database of various diseases, symptoms, and related laboratory test results. It is very difficult to cover all the dependencies between symptoms, test results, and diseases. Still, people are trying to do exactly that. A desire for a tool that, for example, automatically alerts you to a possible threat based on the results of a common medical test is obvious. Similarly, we can imagine the area of financial markets with records of stock market movements and related tool for automatic trading, etc.

In our case, we assume random variables with a discrete finite domain. Each random variable has a probability distribution, which specifies the probability of its values. Set of random variables has a joint probability distribution.

28.1.1 Knowledge Representation

Suppose that knowledge can be represented using a probability distribution defined over a set of corresponding random variables. Of course, the size of such a probability distribution would be enormous. Moreover, even if we were able to store it, we would need a similarly large amount of data to estimate its parameters well. This phenomenon is called *curse of dimensionality*. Here comes the concept of conditional independence. It is well known that in case of independence among variables we can express the corresponding probabilistic distribution as a product of smaller probability distributions (i.e. distributions defined over a smaller set of variables). To save even more space, some weak (conditional) dependencies can be modeled by independencies as well.

28.1.2 Compositional Models

The basic idea of compositional models is simple—to describe global knowledge from an application area using pieces of local knowledge. Local knowledge can be easily obtained, easily stored in a computer, and easily understood by a user/expert. On the other hand, in some cases, the global knowledge of the problem of interest is so complicated that it is beyond human capabilities to describe it. Note that the word *compositional* stands for the fact that probability distribution representing the knowledge about the system is *composed* from a set of low-dimensional distributions

and *model* because the composed probability distribution is of course just a simplification/estimate of the original multidimensional distribution. (To simplify the model, some weak conditional dependencies are modeled by independence relations.)

To handle the knowledge hidden in a compositional model (decomposed probability distribution) one can use the standard tools from probability framework like marginalization, conditioning, and inference. The methods are of course customized to handle the decomposed structure and efficiently implemented using *local computations*.

28.2 Notation and Essentials

Let us consider a finite system of random variables with indices from a non-empty set N . Each variable from this system $\{X_i\}_{i \in N}$ has a finite (and non-empty) set of values \mathbf{X}_i . All the probability distributions discussed in the paper will be denoted by Greek letters. For $K \subset N$, $\kappa(x_K)$ denotes a distribution of variables $X_K = \{X_i\}_{i \in K}$, which is defined on all subsets of a Cartesian product $\mathbf{X}_K = \prod_{i \in K} \mathbf{X}_i$. Thus x_K denotes a $|K|$ -dimensional vector of variable values $\{X_i\}_{i \in K}$ and \mathbf{X}_K represents the set of all such vectors. Having a probability distribution $\kappa(x_K)$ and $L \subset K$ we shall denote its *marginal distribution* by $\kappa(x_L)$. To emphasize the marginalization process, we can also use $\kappa \downarrow^L$.

The symbol $\Pi^{(K)}$ denotes the set of all probability distributions defined for variables X_K . For two distributions defined over the same set of variables $\kappa, \lambda \in \Pi^{(K)}$ we say that λ *dominates* κ ($\kappa \ll \lambda$) if $\forall x \in \mathbf{X}_N : (\lambda(x) = 0 \implies \kappa(x) = 0)$. The distributions $\kappa \in \Pi^{(K)}$ and $\lambda \in \Pi^{(L)}$ are said to be *consistent* if for all $x \in \mathbf{X}_{K \cap L}$ $\kappa(x) = \lambda(x)$.

Definition 1 (*Operator of Composition*) For arbitrary two distributions $\kappa \in \Pi^{(K)}$ and $\lambda \in \Pi^{(L)}$ for which $\kappa \downarrow^{K \cap L} \ll \lambda \downarrow^{K \cap L}$ their *composition* is defined by the following formula

$$(\kappa \triangleright \lambda)(x) = \frac{\kappa(x \downarrow^K) \lambda(x \downarrow^L)}{\lambda \downarrow^{K \cap L}(x \downarrow^{K \cap L})}. \tag{28.1}$$

Otherwise, it remains undefined.

The operator of composition is used to construct multidimensional compositional models. Composing two distributions, we can define a distribution of a dimensionality higher than any of the original ones. The resulting distribution is defined over the union of involved random variables.

By a compositional model of a multidimensional probability distribution we understand a sequence of low-dimensional distributions that assembled together using the operator of composition represent a multidimensional distribution that

$\kappa(\mathbf{M}) \ll \lambda(\mathbf{M})$ denoted that the distribution κ is absolutely continues with respect to distribution λ , which in our finite settings means that whenever κ is positive also λ must be positive.

would be difficult to handle otherwise. In another words, the multidimensional distribution which can be written in the following way

$$\kappa_1 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \dots \triangleright \kappa_n = (\dots ((\kappa_1 \triangleright \kappa_2) \triangleright \kappa_3) \triangleright \dots) \triangleright \kappa_n. \tag{28.2}$$

where we expect κ_i to be defined over variables with indices from K_i . The sequence $\kappa_1, \kappa_2, \dots, \kappa_n$ is called the *generating sequence* of the model.

In this paper, we will focus on the models composed from the marginal distributions of an input distribution obtained from data. Thus there are no inconsistent distributions and the operator of composition is always defined. The sequence of sets of variables (or precisely their indices) K_1, \dots, K_n is called the *structure* of the model. Note that the ordering of sets is important since operator \triangleright is neither commutative nor associative. Because of the nature of the paper, we can simplify the notation and highlight the structure we denote the model from (28.2) in the following manner:

$$(K_1 \cdot K_2 \cdot \dots \cdot K_n)_\kappa$$

Note that compositional models represent a generalization of Bayesian networks. In other words, every Bayesian network can be represented using an equivalent compositional model. Note that structure K_1, \dots, K_n has a similar meaning as graphs in case of Bayesian networks. It represents the system of conditional independencies valid for the model.

For the purpose of the following text, we will introduce a degenerated model, the so-called full model:

Definition 2 (*Full model*) Compositional model κ of the form $\kappa = \kappa(x_N)$ is called *full model*.

In the case of the full model, the sequence of sets of variable indices is formed only by one set N . It means that no composition is performed. Thus, the original data distribution (containing all variables) is a full model.

Definition 3 (*Running Intersection Property*) The sets L_1, L_2, \dots, L_n fulfill the *Running Intersection Property (RIP)* if

$$\forall i \in \{2, \dots, n\} \quad \exists k < i \quad L_i \cap \left(\bigcup_{j < i} L_j \right) \subseteq L_k.$$

Definition 4 (*Decomposability*) The compositional model $(K_1 \cdot K_2 \cdot \dots \cdot K_n)_\kappa$ is said to be *decomposable* if the ordering of sets in its structure K_1, K_2, \dots, K_n fulfills the RIP property.

Definition 5 (*Conditional independence (CI)*) For distribution $\kappa(x_K)$ and for mutually disjoint $A, B, C \subseteq K$ such that $A \neq \emptyset$ and $B \neq \emptyset$ we write $X_A \perp\!\!\!\perp X_B | X_C [\kappa]$ (groups of variables X_A and X_B are *conditionally independent* given X_C with respect to the distribution κ) if

$$\kappa(x_{AUBUC})\kappa(x_C) = \kappa(x_{AUC})\kappa(x_{BUC})$$

for all $x_{AUBUC} \in \mathbf{X}_{AUBUC}$. Note that in case of $C = \emptyset$ we speak about unconditional independence and we denote it as $X_A \perp\!\!\!\perp X_B[\kappa]$.

The use of the operator of composition embeds a conditional independence relation. This fact can be easily shown from both Definitions 1 and 5. (See also, e.g., Lemma 5.2 in [2] where also other basic properties of compositional models are formulated).

28.3 Decomposability

By a decomposition is usually understood the result of a process that, with the goal of simplification, divides an original object into its sub-objects. Thus, for example, a problem is decomposed into two (or more) simpler sub-problems, decomposition of a positive integer into prime numbers, etc. In the latter case, an elementary decomposition is a decomposition of an integer into two factors, the product of which gives the original integer. When repeating the process of decomposition long enough we end up with elementary sub-objects that cannot be further decomposed.

It can be easily deduced from the above-presented properties that the process of a repeatedly performed decomposition of an arbitrary (finite) object into elementary sub-objects (i.e., sub-objects that cannot be further decomposed) is always finite.

In case of a finite two-dimensional probability distribution $\kappa \in \Pi^{(k,l)}$ (k, l are singletons), simpler sub-objects are just one-dimensional distributions: a distribution of variable X_k and a distribution of variable X_l . The process of decomposition corresponds to marginalization—i.e. the sub-objects are $\kappa(x_k)$ and $\kappa(x_l)$. Note that the process of marginalization is well defined. Nevertheless, except for a degenerate case when $X_k \perp\!\!\!\perp X_l[\kappa]$, we cannot unambiguously reconstruct the original two-dimensional distribution from its one-dimensional marginals. In that case, a compositional model composed from one dimensional marginal would be just a very bad estimate of the original distribution.

Having a general probability distribution, one can be interested in the way how to decompose it into a set of its marginals in a way that if composed back together (using the operator of composition), it faithfully reflects the original distribution. Or in other words, if we convert a data set into a probability distribution using e.g. frequency analysis, we would like to learn its compositional model.

The following section deals with a special type of compositional models—decomposable models. The reason why we restricted ourselves to this subclass is clarified later.

28.4 Hierarchy in Decomposable Models Space

The notion of decomposability has been already established in a class of probabilistic models. Following Definition 4, one can notice that decomposability is a structural property in case of compositional models. I.e., it is related to the structure of a compositional model only, not to respective properties of probability distributions from its generating sequence.

Similarly, in the case of Bayesian networks (representant of another approach to probabilistic modeling), decomposability is also a structural property. Recall that in the case of Bayesian networks, a directed acyclic graph is used to represent its structure and we say that a Bayesian network is decomposable if the graph is decomposable. Graph decomposability is equivalent to many other strong graph properties: graph chordality, graph triangularity, the existence of a perfect elimination ordering of nodes, the existence of a junction tree of graph cliques, etc. Simply said, decomposability is a very strong structural property and, what makes it so special, it is closely related to efficient local computations.

By local computation, we understand a possibility to perform complex computations with a probability distribution represented by a compositional model (like marginalization, conditioning, and inference) without the necessity to apply the operator of composition between members of the model generating sequence. Every general compositional model is converted into an equivalent decomposable model before performing any computations with it. This is one of the reasons why we have decided to restrict the current research on structure learning algorithms on the class of decomposable models only.

Assume a compositional model $(K_1 \cdot \dots \cdot K_n)_\kappa$. We recognize the so-called *trivial sets* of the structure. We say that set K_i , ($i \in \{1, \dots, n\}$) is trivial in the structure if $K_i \subseteq \bigcup_{j < i} K_j$. Note that probability distribution corresponding to K_i has no impact on the compositional model. Indeed, considering the definition of the operator of composition (denote $\bigcup_{j < i} K_j$ as $K_{j < i}$ to simplify the formula) then, following (28.1),

$$\begin{aligned} ((\kappa_1 \triangleright \dots \triangleright \kappa_{i-1})) \triangleright \kappa_i(x) &= \frac{(\kappa_1 \triangleright \dots \triangleright \kappa_{i-1})(x \downarrow_{\mathbf{K}_{j < i}}) \kappa_i(x \downarrow_{\mathbf{K}_i})}{\kappa_i \downarrow_{\mathbf{K}_{j < i} \cap \mathbf{K}_i} (x \downarrow_{\mathbf{K}_{j < i} \cap \mathbf{K}_i})} \\ &= \frac{(\kappa_1 \triangleright \dots \triangleright \kappa_{i-1})(x \downarrow_{\mathbf{K}_{j < i}}) \kappa_i(x \downarrow_{\mathbf{K}_i})}{\kappa_i \downarrow_{\mathbf{K}_i} (x \downarrow_{\mathbf{K}_i})} \quad (28.3) \\ &= (\kappa_1 \triangleright \dots \triangleright \kappa_{i-1})(x) \end{aligned}$$

Nevertheless, following Definition 3 of RIP, by adding a trivial set into the structure of a decomposable model, its decomposability can be violated. Nevertheless, we can add a trivial set that is a subset of another set preceding it in the sequence. See the following auxiliary property:

Lemma 1 (Redundant marginal) *Having a set $K \subseteq L_\ell$ the model $(L_1 \cdot L_2 \cdot \dots \cdot L_n)_\kappa$ is decomposable if and only if $(L_1 \cdot \dots \cdot L_\ell \cdot \dots \cdot L_m \cdot K \cdot L_{m+1} \cdot \dots \cdot L_n)_\kappa$ is decomposable.*

Proof Following the same reasoning as in (28.3), we can end up with the simplified model where the *redundant marginal* $\kappa(x_K)$ was removed. Let us emphasize that none of the compositions on the right of the considered marginal is affected by its removal since the union of variables appearing in the model before remains the same. \square

Using the following theorem, one can create a decomposable model from a given decomposable model by introducing a new conditional independence relation into its structure. The proof is constructive. Note that the theorem has been already published in a slightly different form in [3].

Theorem 1 *Assume a decomposable compositional model $\hat{\kappa} = (K_1 \cdot K_2 \cdot \dots \cdot K_n)_\kappa$ where $\exists k \in \{1, \dots, n\}$ such that $|K_k| > 1$. Then there exist a pair of variables $\ell, m \in K_k$ such we can introduce another decomposable model $\hat{\kappa}'$ with one additional conditional independence relation $\{k\} \perp\!\!\!\perp \{\ell\} | (K_k \setminus \{\ell, m\})[\hat{\kappa}']$. We say that $\hat{\kappa}$ and $\hat{\kappa}'$ are in a neighborhood relation.*

Proof Without the loss of generality, we can assume that $k = n$. Indeed, because if it is not the case then we can take just the first k elements of the generating sequence and take it as the model of our interest. Such a generating sub-sequence represents always a marginal of the original model [2] and what holds for the marginal, it holds for the original model as well.

In case of a decomposable model, its structure K_1, \dots, K_n must fulfil RIP property. I.e. it holds

$$\exists i < k \quad K_k \cap \left(\bigcup_{j < k} K_j \right) \subseteq K_i. \quad (28.4)$$

Without loss of generality let us make two assumptions:

1. Let us assume that $K_k \not\subseteq K_i$. (If the opposite was true then K_k would be a trivial column and as such it could be omitted because it does not change the model. For more detail see Lemma 1). I.e. $\exists \ell \in K_k$ such that $\ell \notin K_i$.
2. Further, assume that $|K_k| \geq 2$ because if it is not the case then it has only one element $\ell \notin \bigcup_{j < k} K_j$ (with no intersection with any other set of indices) we can move K_k to any other place without affecting the model [4].

Under these assumptions (or rearrangements of the model) we can choose another element $m \in K_k, m \neq \ell$ and change the structure of the model by introducing new conditional independence relation

$$\{\ell\} \perp\!\!\!\perp \{m\} | K_k \setminus \{\ell, m\}$$

by replacing K_k with sets $K_k \setminus \{\ell\}$ and $K_k \setminus \{m\}$. How to read conditional independence relations from a model structure can be found in [4]. Thus, we obtain a new compositional model where the only change is the replacement of the last distribution in its generating sequence by a pair of its marginals

$$\hat{\kappa}' = (K_1 \cdot K_2 \cdot \dots \cdot K_{k-1} \cdot K_k \setminus \{\ell\} \cdot K_k \setminus \{m\})_{\kappa}.$$

The new structure fulfills RIP property as well, which makes $\hat{\kappa}'$ decomposable. Indeed, because the first part of the structure K_1, K_2, \dots, K_{k-1} remains unchanged, it is enough to check the newly added sets. Note that the intersect of $K_k \setminus \{\ell\}$ with the union of all preceding index sets is in K_i by (28.4). In case of the last set $K_k \setminus \{m\}$ the intersection with all prior sets lies in the set $K_k \setminus \{\ell\}$ and namely it is equal to $K_k \setminus \{\ell, m\}$.

Note that if a trivial set appears, it can be dropped without affecting decomposability of the model. □

For a more detailed view of decomposable models space see [3].

28.5 Mutual Information and Decomposibility

As it has been mentioned in the introduction, decomposable models are essential for efficient use of compositional models due to the possibility of local computations. As an example, we can take the following computations of likelihood-ratio test statistics.

Most of the machine learning methods for probabilistic models construction are, in a way, supported by notions and theoretical results from information theory. E.g. the value of mutual information helps to find pairs of variables that are tightly connected. The value of a multi-information may be used to select the best model from a considered group of models. Note that the basic notion is the famous Shannon entropy from which all the remaining ones are derived.

To help the reader to understand the notion of mutual information, it could be beneficial to highlight that it is the measure of similarity of two distributions. In probability theory, several measures of similarity for distributions have been introduced. One of them, having its origin in information theory, is a Kullback-Leibler divergence defined for $\kappa(K)$ and $\lambda(K)$ by the formula

$$\text{Div}(\kappa \parallel \lambda) = \begin{cases} \sum_{x \in \mathbf{X}_K} \kappa(x) \log \frac{\kappa(x)}{\lambda(x)}, & \text{if } \kappa \ll \lambda \\ +\infty, & \text{otherwise.} \end{cases} \tag{28.5}$$

It is a known fact that Kullback-Leibler divergence is always non-negative and equals 0 if and only if $\kappa = \lambda$ (see [5, 6]). Its only disadvantage is that it is not symmetric, i.e., generally $\text{Div}(\kappa \parallel \lambda) \neq \text{Div}(\lambda \parallel \kappa)$

Therefore, for testing whether the compositional model $\hat{\kappa}$ approximates faithfully original data distribution κ (both with variables from \mathbf{X}_K) one can use Kullback-Leibler divergence.

In our case, we take the full model for κ and we compare it with various decomposable models. Usually, the choice of the optimal model is accomplished either by the process of hypothesis testing or by using some information criterion.

In the following, we illustrate how to take the advantage of decomposability in case of compositional models to calculate Kullback-Leibler divergence using local computations while following the notion of a neighborhood of decomposable models introduced in Theorem 1.

Assume a decomposable compositional model $\hat{\kappa} = (K_1 \cdot K_2 \cdot \dots \cdot K_n)_\kappa$ such that $\exists i \in \{1, \dots, n\} : |K_i| \geq 2$. Following Theorem 1 one can introduce into this model one new conditional independence relation and get a new model $\hat{\kappa}'$ where the original set K_i was replaced by a pair of sets $K_i \setminus \ell$ and $K_i \setminus m$. Note that some of these sets may be trivial in the structure of the new model and appropriate probability distributions may be removed from the model generating sequence by Lemma 1 without affecting the decomposability.

Following Theorem 1, the new model $\hat{\kappa}'$ can be obtained by multiplication of the formula for model $\hat{\kappa}$ by a simple factor:

$$\hat{\kappa}' = \hat{\kappa} \cdot \frac{\kappa(x_{K_i \setminus \ell})\kappa(x_{K_i \setminus m})}{\kappa(x_{K_i \setminus \{\ell, m\}})\kappa(x_{K_i})}. \quad (28.6)$$

The Kullback-Leibler divergence for full model κ and the new model $\hat{\kappa}'$ is

$$\text{Div}(\kappa \parallel \hat{\kappa}') = \sum_{x \in \mathbf{X}_K} \kappa(x) \log \frac{\kappa(x)}{\hat{\kappa}'(x)}. \quad (28.7)$$

Note that the divergence is always defined because we work with marginals of κ . (28.7) can be rewritten using (28.6) into

$$\text{Div}(\kappa \parallel \hat{\kappa}') = \sum_{x \in \mathbf{X}_K} \left(\kappa(x) \cdot \log \frac{\kappa(x)\kappa(x_{K_i \setminus \{\ell, m\}})\kappa(x_{K_i})}{\hat{\kappa}'(x)\kappa(x_{K_i \setminus \ell})\kappa(x_{K_i \setminus m})} \right)$$

which can be further split into the sum of two logarithms

$$\text{Div}(\kappa \parallel \hat{\kappa}') = \sum_{x \in \mathbf{X}_K} \kappa(x) \log \frac{\kappa(x)}{\hat{\kappa}'(x)} + \sum_{x \in \mathbf{X}_K} \kappa(x) \log \frac{\kappa(x_{K_i \setminus \{\ell, m\}})\kappa(x_{K_i})}{\kappa(x_{K_i \setminus \ell})\kappa(x_{K_i \setminus m})}.$$

Notice that the left part is a Kullback-Leibler divergence of κ and the original model $\hat{\kappa}$. The right-hand sum can be further rewritten as

$$\text{Div}(\kappa \parallel \hat{\kappa}') = \text{Div}(\kappa \parallel \hat{\kappa}) + \sum_{x \in \mathbf{X}_{K_i}} \left(\left(\sum_{x \in \mathbf{X}_{K_i}} \kappa(x) \right) \cdot \log \frac{\kappa(x_{K_i \setminus \{\ell, m\}})\kappa(x_{K_i})}{\kappa(x_{K_i \setminus \ell})\kappa(x_{K_i \setminus m})} \right)$$

and the inner sum is equal to a marginal $\kappa(x_{K_i})$. I.e.

$$\text{Div}(\kappa \parallel \hat{\kappa}') = \text{Div}(\kappa \parallel \hat{\kappa}) + \sum_{x \in \mathbf{X}_{K_i}} \kappa(x_{K_i}) \log \frac{\kappa(x_{K_i \setminus \{\ell, m\}})\kappa(x_{K_i})}{\kappa(x_{K_i \setminus \ell})\kappa(x_{K_i \setminus m})}.$$

Following the last formula, we can easily and efficiently compute the divergence of the new model using the already computed divergence of $\hat{\kappa}$ and local computations concerning the replaced low-dimensional marginal defined by indices K_i only.

28.6 Model Complexity

By decomposing the original probability distribution into its marginals we reduce the number of its parameters. That is, by the way, the main reason to do the decomposition at all. The lower number of parameters, the faster the computations are, the easier one can store the model in computer memory. Realize that current models work with dozens or hundreds of variables.

In the case of our elementary approach, we will simply use the number of parameters needed to represent the compositional model in computer memory. Because every compositional model is represented using its generating sequence—a sequence of probability distributions—we will sum the size of respective probability distributions. In this paper, we restricted ourselves to discrete finitely valued random variables. Therefore, respective probability distributions can be represented using contingency tables, where the size of each table is connected with the number of distinct values of involved random variables.

Let r_k be the number of categories for variable k ($\forall k \in K : r_k = |\mathbf{X}_k|$). Then, in case of the full model $\kappa(x_K)$, we need a probability table with $\prod_{k \in K} r_k$ cells. Note that this number can be decreased by one—probabilities must sum up to one. The number of parameters needed to represent the full model is then given by formula

$$C_F = \prod_{k \in K} r_k - 1.$$

Assume a general compositional model

$$\hat{\kappa} = (K_1 \cdot K_2 \cdot \dots \cdot K_n)_\kappa.$$

Despite the fact that a compositional model can be expressed in the form of product of conditional distributions where the i th conditional distribution is a distribution of variables with indices from K_i not present in previous index sets and is conditioned by variables of K_i which already appeared in the previous parts of model, we use a standard representation using unconditional probability distributions. One of the reasons is that this representation makes local computations easier.

In this case, the number of parameters needed to represent compositional model $\hat{\kappa}$ is

$$df = \sum_{i \in \{1..n\}} \left(\prod_{j \in K_i} r_j - 1 \right)$$

28.7 Information Criteria

The goal of decomposition is to get a compositional model as simple as possible which is, of course, in direct contradiction to the faithfulness of the model. To get an optimal balance between these two measures one can become inspired by various approaches used in other probabilistic learning methods.

Another approach to estimating the optimal size of the compositional model can be found in [7]. In this paper, the authors suggest to use the famous Huffman code [8] to find in a way optimum code to encode the original data set. The procedure is rather simple and it belongs to the fundamental parts of information theory. The idea is rather simple. First, find the Huffman encoding of the data or full model. Nevertheless, we do not need the encoding, we just need the number of bits necessary to encode the data or the model. Note that in the case of Huffman code one has to keep also the coding table to reconstruct the original object. The total size of both objects gives us a hint about the space needed to encode the data and we can use it to restrict the size of the compositional model as well.

On the other hand, information criteria that began to appear in the '70s of the 20th century remind a famous Occam's razor. Based on this principle the simplest model is chosen from a class of models describing the data in the same quality.

Let us recall the famous Bayesian information criterion (BIC), Schwarz criterion [9], or Akaike (AIC) criterion [10]. The criteria are generally a difference between the distance of the model from data and the size of the space needed to store it. One variant of BIC in the notation of this paper can be

$$\text{BIC}_{\hat{\kappa}} = 2 \cdot \text{Div}(\kappa \perp \hat{\kappa}) - \log(n) \cdot df$$

where n is the number of observations in data.

Let us highlight, that we do not have any useful information criterion so far. We hope that we will receive one based on experiments performed in the next section.

28.8 Algorithms

Theorem 1 suggests to perform the *breadth-first search* through a tree of decomposable models where the root of the tree is the full model. Neighbors in the tree (also in the meaning of Theorem 1) differs from each other by additional conditional independence relation introduced to the structure. Using the theorem-proof, we can immediately construct the tree. The decomposability of all models is guaranteed. Moreover, we can use the advantage of local computations of test statistics needed for information criteria— df and Kullback-Leibler divergence.

The following questions arise: Does the tree contain all possible decomposable models? Or in other words, can we find a path from any decomposable model to a full model in such a tree? The answer is positive. Let us note that an inverse assertion to

Theorem 1 can be proven (see Theorem 12 in [11]). Both together they guarantee the existence of a path of neighbor models from a full model to an arbitrary decomposable model by repeated application of Theorem 1.

The exhaustive search among all decomposable models is computationally intractable. Indeed, the number of decomposable models is enormous—numerical results for the case of mathematically equivalent chordal graphs (the structure of an arbitrary decomposable model can be represented using a chordal graph and vice-versa) can be seen in [12] or [13]. Nevertheless, we expect that the tree traversal could be significantly speedup using various techniques like gradient descent method.

Because the optimal method does not exist, we hope that a sub-optimal method can be found. The algorithm would not go through all nodes of the tree, but it will go through a restricted sub-tree with e.g. the best values of test statistics (see [14]).

Let us start with the simplest possible algorithm—a *greedy search algorithm*. It is based on the idea to take the best optimal choice in each step to eventually reach the global optimum. The algorithm picks the best solution in each step regardless of the consequences. Using Theorem 1 we can design it as follows:

Algorithm 1 (Greedy search) *Start with the full model.*

1. **Generate all decomposable models in its neighborhood by adding a conditional independence relation.**
2. **Choose the best model according to an information criterion (the difference between df and Kullback-Leibler divergence from the full model).**
3. **Repeat steps 1 and 2 until the information criterion starts to increase.**

The other idea is to use a slightly modified greedy search enhanced with a history of previous search.

Algorithm 2 (Greedy search— k best) *Start with the full model.*

1. **Generate all neighboring decomposable models with an additional CI relation between a pair of variables.**
2. **Choose k best models according to a given information criterion.**
3. **Repeat steps 1 and 2 until the information criterion starts to increase.**

28.9 Experiments

We have performed several experiments to explore the possibility of usage of greedy search approach for structure learning of compositional models. We have used two data sets. The famous ASIA data set from [15]—an artificial data set generated from a probabilistic model based on a hypothetical medical situation. The data set has 8 variables (A,B,D,E,L,S,T,X) and 5000 records. Then, because of the computational complexity of an exhaustive search in the space of decomposable models over 8 variables, we have used also 6 variables data set REINIS from [14] with 1200 records.

Using a frequency analysis, we created full models. Then, the models were iteratively decomposed using Theorem 1, likelihood-test statistics, and the greedy search algorithm. In the case of REINIS data set, we have also performed the exhaustive scan through the whole space of decomposable models over 6 variables.

To generate all decomposable models, we have used the known fact that a sequence of sets satisfying RIP property corresponds to cliques in a chordal graph (ordered using maximum cardinality search algorithm). We used a catalog of all chordal graphs over six variables from [16]. Note that in case of six variables there are 18,395 decomposable models (without those with more than 4 singletons), based on 75 chordal graphs.

The run of the greedy search algorithm in case of REINIS data set is illustrated in Table 28.1. The first column corresponds to a structure of respective compositional model—respective probability distributions are marginals of the full model. The second column contains KL-divergence (also called relative entropy)—a measure of how one probability distribution (represented by a compositional model with a given structure) is different from a second, reference probability distribution corresponding to the full model. df is the above-defined number of parameters needed to represent the model. The last but one column contains the newly introduced independence relation to the structure of the model using Theorem 1 by splitting set K_i . The index i can be found in the last column. Note that the run was not stopped by any criterion and it was performed until the structure was split into a sequence of singletons.

To see the quality of greedy search approach, compare the results from Table 28.1 with Table 28.2 which contains the results of the exhaustive search in the class of

Table 28.1 Greedy search over the set of all decomposable models of REINIS data set

Structure	KL-diverg.	df	Suggest ind.	i
(A,B,C,D,E,F)	0.00000	63	$B \perp D K_i$	1
(A,B,C,E,F) (A,C,D,E,F)	0.00479	62	$C \perp D K_i$	2
(A,B,C,E,F) (A,D,E,F)	0.00759	46	$A \perp F K_i$	2
(A,B,C,E,F) (A,D,E)	0.01009	38	$E \perp F K_i$	1
(A,B,C,E) (A,B,C,F) (A,D,E)	0.01368	37	$A \perp B K_i$	1
(A,C,E) (B,C,E) (A,B,C,F) (A,D,E)	0.01616	36	$C \perp E K_i$	2
(A,C,E) (B,C) (A,B,C,F) (A,D,E)	0.01868	32	$C \perp F K_i$	3
(A,C,E) (B,C) (A,B,F) (A,D,E)	0.02143	24	$A \perp F K_i$	3
(A,C,E) (B,C) (B,F) (A,D,E)	0.02247	20	$B \perp F K_i$	3
(A,C,E) (B,C) (F) (A,D,E)	0.02432	18	$A \perp D K_i$	4
(A,C,E) (B,C) (F) (D,E)	0.03081	14	$D \perp E K_i$	4
(A,C,E) (B,C) (F) (D)	0.03582	12	$C \perp E K_i$	1
(A,C) (A,E) (B,C) (F) (D)	0.04432	11	$A \perp E K_i$	2
(A,C) (E) (B,C) (F) (D)	0.05113	9	$A \perp C K_i$	1
(A) (B,C) (E) (F) (D)	0.06190	7	$B \perp C K_i$	2

Table 28.2 Exhaustive search over the space of all decomposable models

Structure	KL-diverg.	<i>df</i>
(A,B,C,E,F) (A,C,D,E,F)	0.00479	62
(A,B,C,E,F) (A,D,E,F)	0.00759	46
(A,B,C,F) (A,C,E,F) (A,D,E,F)	0.01204	45
(A,B,C,E,F) (A,D,E)	0.01009	38
(A,B,C,F) (A,B,C,E) (A,D,E)	0.01368	37
(A,B,C,D,E) (B,F)	0.01457	34
(A,B,C,E) (A,B,D,E) (B,F)	0.01649	33
(A,B,C,D,E) (F)	0.01643	32
(A,B,C,E) (A,B,D,E) (F)	0.01834	31
(A,D,E,F) (A,B,C,E)	0.01495	30
(A,B,C,F) (A,C,E) (A,D,E)	0.01633	29
(A,C,E) (B,C,E) (A,D,E) (D,E,F)	0.01901	28
(A,B,C,E) (A,D,E) (B,F)	0.01747	25
(A,C,E) (A,D,E) (B,C,E) (B,F)	0.01995	24
(A,B,C,E) (A,D,E) (F)	0.01932	23
(A,C,D,E) (B,C,F)	0.02070	22
(A,D,E) (A,C,E) (B,C,E) (F)	0.02180	22
(A,D,E) (A,C,E) (B,C,F)	0.02157	21
(A,C,D,E) (B,C) (B,F)	0.02160	21
(A,D,E) (A,C,E) (B,C) (B,F)	0.02247	20
(A,C,D,E) (B,C) (F)	0.02346	19
(A,D,E) (A,C,E) (B,C) (F)	0.02432	18
(D,E,F) (A,C,E) (B,C)	0.02801	17
(A,B,C,E) (F) (D)	0.03083	17
(B,F) (B,C) (A,C,E) (D,E)	0.02895	16
(A,C,E) (B,C,E) (F) (D)	0.03330	16
(A,C) (B,C) (B,E) (D,E) (B,F)	0.03724	15
(A,D,E) (A,B,C) (F)	0.03047	15
(A,C,E) (B,C)(D,E) (F)	0.03081	14
(A,C) (B,C) (B,E) (D,E) (F)	0.03909	13
(A,C,E) (B,C) (D) (F)	0.03582	12
(A,C) (B,C) (B,E) (F) (D)	0.04411	11

all decomposable models of REINIS data set. More precisely, Table 28.2 contains a set of best models based on Kullback-Leibler divergence from a full model for each possible structure complexity df . One can see, that it is not true that for a smaller df the corresponding KL divergence has to be higher. Similarly, it seems to be difficult to decide which ratio of KL divergence and df is reasonable. There is no significant change in KL divergence considering decreasing df .

Table 28.3 illustrates the greedy search algorithm in the case of ASIA data set. Because the size of the set of all decomposable models for eight variables was for us computationally intractable, we have added two additional lines not corresponding to the greedy search algorithm. The last but one algorithm represents the original model used for data generation. Note that the model is not decomposable. Its decomposable version [4] is in the last row of the table. You can see that with this setting, the greedy algorithm is far from finding it. The biggest problem is located in a huge KL-divergence jump in the 9th step of the algorithm. This will require further investigation of the problem. Nevertheless, even if we compute the KL-divergence globally (not employing local computations) we end up with the same numbers.

28.10 Conclusion

This paper introduces a theoretic background for iterative compositional model learning. Although the introduced test statistics is feasible to efficiently compute using local computations, it seems that a simple greedy approach is not good enough. There are still several problems to be solved:

- to find a suitable criterion to stop the decomposition process,
- to check whether $k > 1$ will lead to better results and if not, to come with another algorithm, and
- to check the circumstances under which the greedy approach can provide solutions sufficiently close to the optimal solution.

To solve the problem we have to find a way how to efficiently generate the complete class of decomposable models for eight variables—probably using the catalog of chordal graphs by employing the fact that a chordal graph is just another mathematical representation of a decomposable structure. The problem is the number of permutations of eight variables, nevertheless, using [4], we should be able to determine structures from the same equivalence class and keep only one representant of each class.

Table 28.3 Greedy search algorithm run in case of Asia data set

Structure	KL-diverg.	df	i	Independence
(A,B,D,E,L,S,T,X)	0.00000	255	1	$A \perp E K_i$
(A,B,D,L,S,T,X) (B,D,E,L,S,T,X)	0.00000	254	2	$S \perp E K_i$
(A,B,D,L,S,T,X) (B,D,E,L,T,X)	0.00000	190	2	$L \perp X K_i$
(A,B,D,L,S,T,X) (B,D,E,L,T)	0.00000	158	2	$B \perp E K_i$
(A,B,D,L,S,T,X) (D,E,L,T)	0.00000	142	2	$E \perp D K_i$
(A,B,D,L,S,T,X) (E,L,T)	0.00000	134	1	$S \perp X K_i$
(A,B,D,L,S,T) (A,B,D,L,T,X) (E,L,T)	0.00036	133	2	$B \perp X K_i$
(A,B,D,L,S,T) (A,D,L,T,X) (E,L,T)	0.00063	101	2	$A \perp L K_i$
(A,B,D,L,S,T) (A,D,T,X) (E,L,T)	0.22641	85	2	$A \perp T K_i$
(A,B,D,L,S,T) (A,D,X) (E,L,T)	0.25171	77	1	$S \perp D K_i$
(A,B,L,S,T) (A,B,D,L,T) (A,D,X) (E,L,T)	0.25207	76	1	$A \perp S K_i$
(A,B,D,L,T) (B,L,S,T) (A,D,X) (E,L,T)	0.25260	60	3	$A \perp X K_i$
(A,B,D,L,T) (B,L,S,T) (D,X) (E,L,T)	0.25346	56	2	$T \perp L K_i$
(A,B,D,L,T) (B,S,T) (D,X) (E,L,T)	0.28748	48	2	$S \perp T K_i$
(A,B,D,L,T) (B,S) (D,X) (E,L,T)	0.28825	44	1	$A \perp T K_i$
(A,B,D,L) (B,D,L,T) (B,S) (D,X) (E,L,T)	0.28916	43	1	$A \perp D K_i$
(A,B,L) (B,D,L,T) (B,S) (D,X) (E,L,T)	0.28990	35	1	$A \perp L K_i$
(A,B) (B,D,L,T) (B,S) (D,X) (E,L,T)	0.29065	31	1	$A \perp B K_i$
(A) (B,D,L,T) (B,S) (D,X) (E,L,T)	0.29109	29	2	$T \perp L K_i$
(A) (B,D,T) (B,D,L) (B,S) (D,X) (E,L,T)	0.29274	28	2	$T \perp B K_i$
(A) (D,T) (B,D,L) (B,S) (D,X) (E,L,T)	0.29685	24	2	$T \perp D K_i$
(A) (T) (B,D,L) (B,S) (D,X) (E,L,T)	0.29971	22	3	$L \perp B K_i$
(A) (T) (D,L) (B,D) (B,S) (D,X) (E,L,T)	0.31149	21	6	$X \perp D K_i$
(A) (T) (D,L) (B,D) (B,S) (X) (E,L,T)	0.32927	19	3	$L \perp D K_i$
(A) (T) (L) (B,D) (B,S) (X) (E,L,T)	0.35411	17	7	$T \perp L K_i$
(A) (E,T) (L) (B,D) (B,S) (X)	0.70100	12	2	$T \perp E K_i$
(A) (T) (E) (L) (B,D) (B,S) (X)	0.73479	11	6	$S \perp B K_i$
(A) (T) (E) (L) (B,D) (S) (X)	0.86501	9	5	$B \perp D K_i$
(A) (T) (E) (L) (B) (D) (S) (X)	1.20740	8		
(A) (S) (B,S) (L,S) (T,A) (E,L,T) (D,B,E) (X,E)	0.00829	28		
(A,T) (E,L,T) (E,L,S) (B,E,S) (B,D,E) (E,X)	0.00789	34		

Acknowledgements The research was financially supported by grants GAČR no. 19-04579S (first author), GAČR no. 19-06569S (second author) and AV ČR no. MOST-04-18 (third and fourth author).

References

1. Pólya, G.: How to solve it, 2nd edn. Doubleday Anchor Books, Garden City (1957)
2. Jiroušek, R.: Foundations of compositional model theory. *Int. J. Gen. Syst.* **40**(6), 623–678 (2011)
3. Bína, V.: Multidimensional Probability Distributions: Structure and Learning. Ph.D. thesis, University of Economics, Prague, Faculty of Management (2011)
4. Jiroušek, R., Kratochvíl, V.: Foundations of compositional models: structural properties. *Int. J. Gen. Syst.* **44**(1), 2–25 (2015)
5. Kullback, S., Leibler, R.A.: On information and sufficiency. *Ann. Math. Stat.* **22**, 76–86 (1951)
6. Kullback, S.: An information-theoretic derivation of certain limit relations for a stationary Markov chain. *J. SIAM Control* **4**, 454–459 (1966)
7. Jiroušek, R., Krejčová, I.: Minimum description length principle for compositional model learning. In: International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, pp. 254–266. Springer (2015)
8. Huffman, D.A.: A method for the construction of minimum-redundancy codes. *Proc. IRE* **40**(9), 1098–1101 (1952)
9. Schwarz, G.: Estimating the dimension of a model. *Ann. Stat.* **6**(2), 461–464 (1978)
10. Akaike, H.: A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**(6), 716–723 (1974)
11. Bína, V.: Exhaustive search among compositional models of decomposable type. In 13th Czech-Japan Seminar on Data Analysis and Decision Making in Service Sciences, pp. 103–108, Otaru University of Commerce. University Hall (2010)
12. Kawahara, J., Saitoh, T., Suzuki, H., Yoshinaka, R.: Enumerating all subgraphs without forbidden induced subgraphs via multivalued decision diagrams. CoRR, abs/1804.03822 (2018)
13. Matsui, Y., Uehara, R., Uno, T.: Enumeration of perfect sequences of chordal graph. In Hong, S-H., Nagamochi, H., Fukunaga, T., (eds.) Algorithms and Computation, pp. 859–870, Berlin, Heidelberg (2008)
14. Havánek, T.: A procedure for model search in multidimensional contingency tables. *Biometrics* **40**(1), 95–100 (1984)
15. Lauritzen, S.L., Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc. Ser. B (Methodological)*, 157–224 (1988)
16. McKay, B.: Various simple graphs. <http://users.cecs.anu.edu.au/~bdm/data/graphs.html> (2019). Online accessed 27 Aug. 2019

Chapter 29

Consciousness Detection in Complete Locked-In State Patients Using Electroencephalogram Coherency and Artificial Neural Networks



V. S. Adama and Martin Bogdan

Abstract In this study, a method to uncover levels of consciousness using electroencephalogram (EEG) coherency and artificial neural network is presented. The subjects of interest are complete locked-in syndrome (CLIS) patients. These patients are characterized by complete paralysis and sufficiently intact cognition. Consequently, they are aware of themselves and their surroundings, but are unable to produce speech. A great challenge in the study of consciousness in patients with CLIS is that there are no certainty regarding their level of awareness at all time. In this paper, a method using EEG coherence matrices as input to a convolutional autoencoder to determine a patient's level of consciousness is presented. The ultimate goal of the research is to build a brain–computer interface-based communication device to allow interactions with CLIS patients.

29.1 Introduction

Much research in recent years has focused on trying to uncover states of consciousness in LIS and CLIS patients. Patients in locked-in state preserve their cognitive functions but are unable to move and to produce speech. Eye movements are however preserved, which could be used as a means of communication. Unfortunately, eye movements are lost as well when patients transitioned to complete locked-in syndrome (CLIS), which makes communication extremely challenging [1]. The physiological and behavioural changes that occur in such transition are described in [2]. Since most of their cognitive states are retained, brain–computer interfaces could be a solution to re-establish communication with the patients. Current approaches in consciousness detection in CLIS consist of analysing event-related potentials such as the P300 or to measure motor imagery [3, 4]. In spite of promising results, one major challenge is to uncover with certainty the consciousness states of these patients.

V. S. Adama (✉) · M. Bogdan
Department of Computer Engineering, Leipzig University, Augustusplatz 10, 04109 Leipzig,
Germany
e-mail: adama@informatik.uni-leipzig.de

In this paper, we present a proof of concept of a method based on artificial neural network and electroencephalogram (EEG) coherency to improve the probability of correctly predicting the patient’s level of consciousness. The general approach presented here was first roughly introduced in [5] used as part for a hybrid-based detection system. In this paper, we focus strictly on the method using two other patients to reassure the approach. The ultimate goal of the research is to design a reliable communication device to CLIS using brain–computer interface working in the case of reliably detected consciousness of the patient in question.

The paper is organized as follows. In Sect. 29.2, a description of the data used in this proof of concept is provided, followed by the signal preprocessing methods. Then, the feature extraction process is outlined before presenting the artificial neural network method used. The results are presented in Sect. 29.3. A brief discussion part is presented in Sect. 29.4 before concluding.

29.2 Tools and Methods

29.2.1 Description of the Approach

In this paper, an artificial neural network-based method using coherence values of EEG recordings is used in an attempt to uncover consciousness states in two patients suffering from complete locked-in syndrome.

Figure 29.1 shows the flow chart of the EEG signal analysis to do so. After preprocessing the signal, the imaginary part of the coherence is calculated and fed into a directional autoencoder. The imaginary part of coherence is used for several reasons. It reduces possible effects of volume conduction on the brain and can inform about the direction of information flow [6]. A visual inspection of videos made from consecutive electrocorticogram (ECoG) coherence matrices displayed changes in the direction of the flow depending on the state of consciousness of the patient in [5].

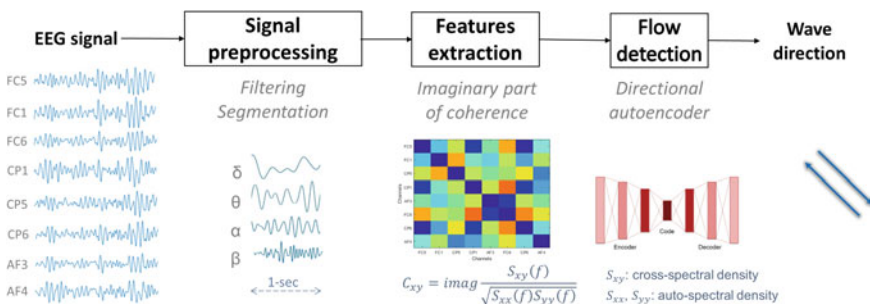


Fig. 29.1 Flow chart of the EEG signal analysis

A convolutional autoencoder was used to attempt to uncover these direction changes. In this paper, the same approach is used with EEG.

29.2.2 Data Description and Preprocessing

EEG data were acquired from two complete locked-in syndrome patients (*Patient 1* and *Patient 2*) during rest and while accomplishing the auditory paradigm described in [4]. The patients were asked questions that require a “yes” or “no” answer. First, questions with known (by the family) answers were asked and data recorded during this session are used to train a support vector machine (SVM) classifier. This latter will then be used in an open question session to predict the patient’s answer. Each session lasted about 10 min.

The challenge to acquire data from these types of patients is its difficulty to record regularly and with the same experimental settings for all different patients located all over Germany. The EEG was recorded at 200 Hz from electrodes located according to the 10/10 system at FC5, FC1, FC6, CP5, CP1 and CP6 for Patient 1, and at AF3 and AF4 in addition to those of Patient 1 for Patient 2 (cf. Fig. 29.2).

Data analysis is performed using MATLAB R2019a (Massachusetts, Texas, USA), the FieldTrip toolbox [7] and custom-written codes. Before extracting the features of interest, the data are re-referenced to the mean and band-pass filtered using a third-order Butterworth filter with cut-off frequencies between 0.5 and 50 Hz. Afterwards, the filtered signals are segmented into 1-s duration time slices.

29.2.3 Feature Extraction: Imaginary Coherence

The coherency is used to investigate the relation between two signals. For two signals x and y , it can be computed at a frequency f as [8]:

$$C_{xy}(f) = \frac{S_{xy}(f)}{\sqrt{S_{xx}(f) \cdot S_{yy}(f)}} \quad (29.1)$$

where S_{xx} and S_{yy} are the individual power spectral density of x and y , and S_{xy} is the cross-power spectral density of x and y at frequency f . The coherency C_{xy} is a complex number:

$$C_{xy}(f) = \Re(C_{xy}(f)) + i\Im(C_{xy}(f)) \quad (29.2)$$

$$C_{xy}(f) = |C_{xy}(f)|e^{i\phi} \quad (29.3)$$

In Eq. (29.3), ϕ represents the phase angle between x and y , and the magnitude is obtained using:

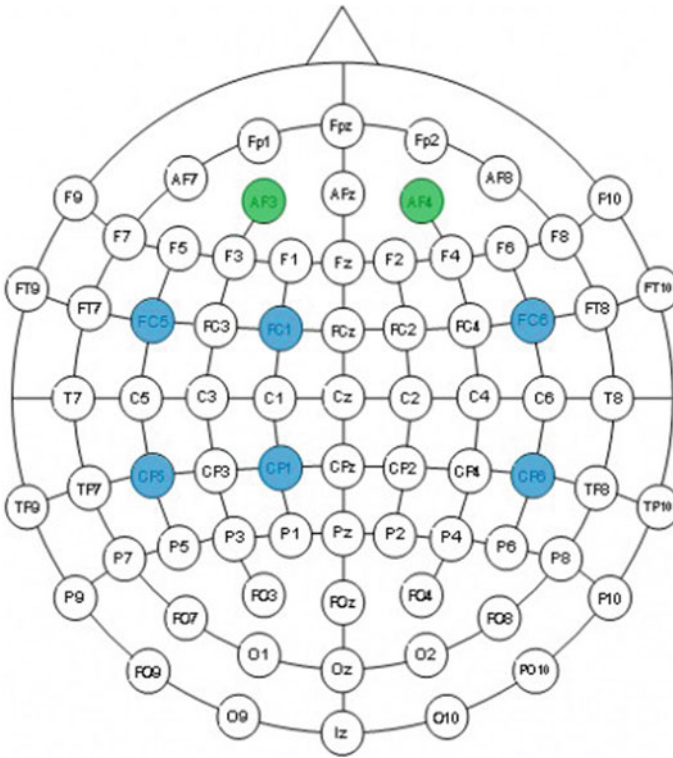


Fig. 29.2 Electrodes from which the EEG signal was recorded. In blue: common channels for both patients. In green: additional channels for Patient 2

$$|C_{xy}(f)| = \sqrt{\Re(C_{xy}(f))^2 + \Im(C_{xy}(f))^2} \tag{29.4}$$

In neuroscience, coherence can be used to measure functional relationship between two different brain regions. A higher value of coherence signifies that the functional relationship between these regions has increased. The most common coherence measure used in this case is the magnitude-squared coherence $|C_{xy}(f)|$. However, it is suggested to use the imaginary part of the coherence to reduce the effects of volume conduction in the recorded EEG signal and avoid false connectivity [6].

After preprocessing the EEG signals, for every 1-s segment of each channel pair x and y and for each frequency band δ (0.4–8 Hz), θ (4–8 Hz), α (8–12 Hz) and β (12–30 Hz), the imaginary part of the coherency $i\text{COH}$ is extracted from Formula (29.2):

$$i\text{COH}_{xy}(f) = \Im(C_{xy}(f)) \tag{29.5}$$

Its value ranges from -1 to $+1$. A positive value suggests that information is flowing from x to y [6].

This results in a feature array of size $n_{\text{timePoints}} * n_{\text{channels}} * n_{\text{channels}}$ for each session. This feature array is used as input to a directional autoencoder.

29.2.4 Convolutional Autoencoder

The idea behind using a directional autoencoder in this research is to uncover the wave direction flow from sliding windows of consecutive coherence matrices. It is assumed that a change of direction corresponds to a level of consciousness. An autoencoder is an artificial neural network that uses (un)supervised learning to efficiently code data by using its reduced representation [9]. In addition, here, we combine the autoencoder approach with the CNN approach.

Architecture The convolutional autoencoder consists of convolutional layers followed by some fully connected layers similar to [10]. Its architecture is illustrated in Fig. 29.3. The method consists of transforming a sliding window of the data that exhibits a specific motion direction into a low-dimensional representation. The sliding windows are stacked on top of each other to form a feature map for the convolutional layer. Under the assumption that the movements occur from the front to the back of the cortex or inversely, executing the sliding window backward would derive an opposite movement. In Fig. 29.3, x represents a specific time and x^{-1} its opposite. The directional neuron is given by $d = d_x$ which value ranges from -1 to 1 and should satisfy:

$$-d = d_{x^{-1}} \quad (29.6)$$

where $d_{x^{-1}}$ is directional neuron of the opposite movement.

Loss Function The loss function is used to evaluate the goodness of neural network predictions. Mean square error (MSE) is one of the most used loss functions [11]. It is calculated as:

$$\text{MSE} = \sum_{i=1}^M \frac{1}{2} (d_i - x_i) \quad (29.7)$$

for a neural network with M output nodes $y_i, i = 1, \dots, M, d_i$ being the predicted value and x_i the real value [12].

In this case, the total loss of the convolutional encoder is computed as:

$$L_{\text{total}} = L_{ae} + L_{ae}^{-1} + L_{\text{content}} + L_{\text{direction}} \quad (29.8)$$

where

$$L_{ae} = \text{MSE}(x, x') \quad (29.9)$$

$$L_{ae}^{-1} = \text{MSE}(x^{-1}, x^{-1'}) \quad (29.10)$$

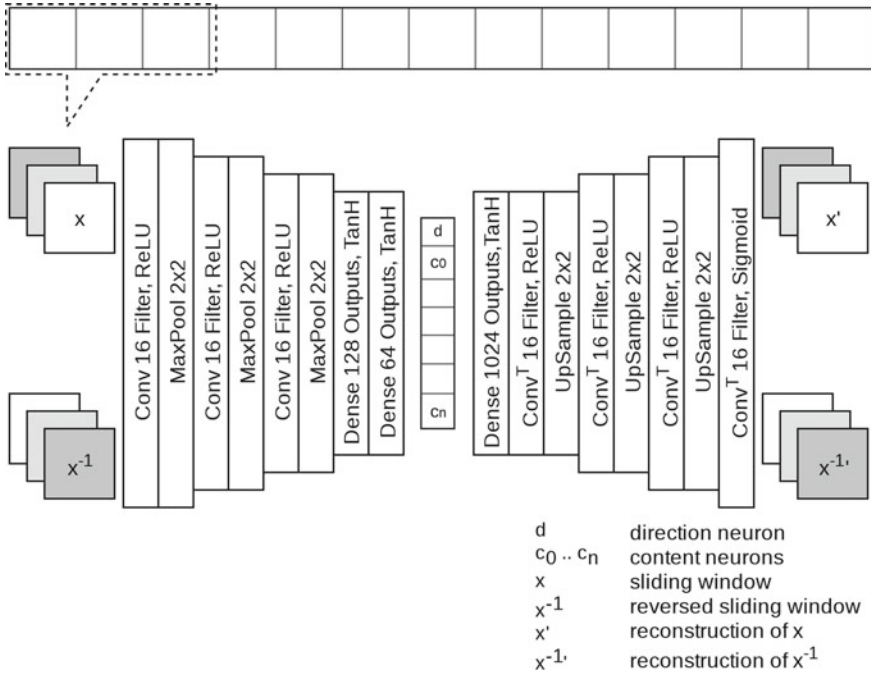


Fig. 29.3 Architecture of the convolutional encoder [5]

$$L_{\text{content}} = \text{MSE}((c_0 \dots c_n), (c_0^{-1} \dots c_n^{-1})) \tag{29.11}$$

$$L_{\text{direction}} = \text{MSE}(-d, d_{x^{-1}}) \tag{29.12}$$

Training The neural network is trained using the beta bands with a learning rate of 0.001 and 5000 epochs (or until no significant improvement of the loss was observed). The weights are minimized using the Adam optimizer [13] on an equally split training and test sets. The batch size is 512 randomly chosen sliding windows of size 3.

29.3 Results

The EEG brain signals of two locked-in syndrome patients were recorded for several days. The imaginary part of coherence of the EEG signals was computed for each pair of channels for all sessions and further analysed using a convolutional autoencoder. The goal of the study is to investigate the usability of the methods to evaluate consciousness states in these patients.

29.3.1 Imaginary Part of Coherence

The coherence matrix displays the connections between all pairs of electrodes. It consists of a 6×6 matrix for Patient 1 and a 8×8 matrix for Patient 2 (cf. Fig. 29.1). The value of coherence during rest was used as baseline to evaluate its variations.

The imaginary part of the coherence as a function of time for Patient 1 is shown in Fig. 29.4 for the θ band and in Fig. 29.5 for the α band. Lower values of α coherence are observed for Patient 1 during day 1 for the first 2 sessions. These values tend to progressively increase as time passes. No distinguishable differences are observed during day 2.

Figure 29.6 shows the imaginary coherence values across time for Patient 2. The values are fluctuating around the baseline value. No visible changes in coherence values are observed for Patient 2 for all days and for all frequency bands compared to the baseline (cf. Fig. 29.7).

29.3.2 Convolutional Autoencoder

The convolutional autoencoder is trained with the beta coherence feature maps. The level of consciousness is considered higher when the value is below the threshold. The outputs of the convolutional autoencoder are illustrated in Figs. 29.8, 29.9, 29.10 and 29.11.

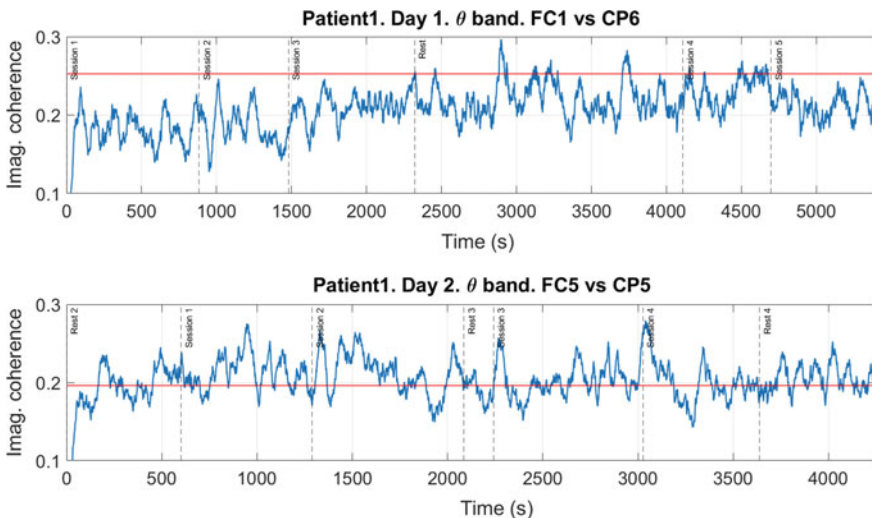


Fig. 29.4 θ coherence between FC1 and CP6 as a function of time during all sessions for Patient 1. The horizontal red line represents the mean coherence during rest. Vertical dotted lines: start/end of session

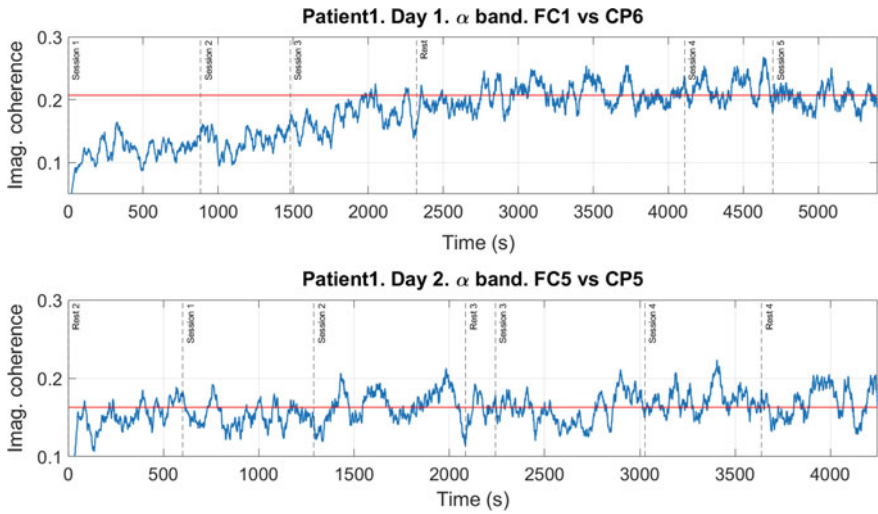


Fig. 29.5 α coherence between FC1 and CP6 as a function of time during all sessions for Patient 1. The horizontal red line represents the mean coherence during rest. Vertical dotted lines: start/end of session

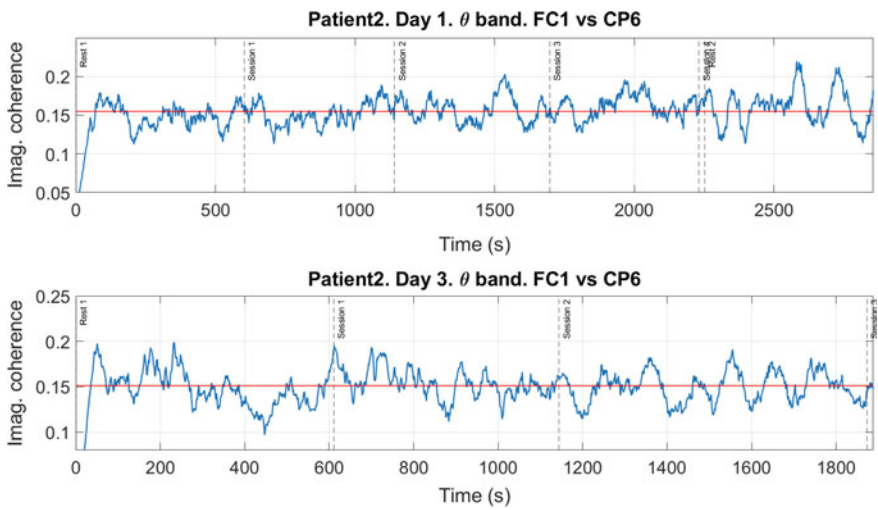


Fig. 29.6 θ coherence as a function of time during all sessions for Patient 2. The horizontal red line is the baseline computed as the mean coherence during rest. Vertical dotted lines: start and end session

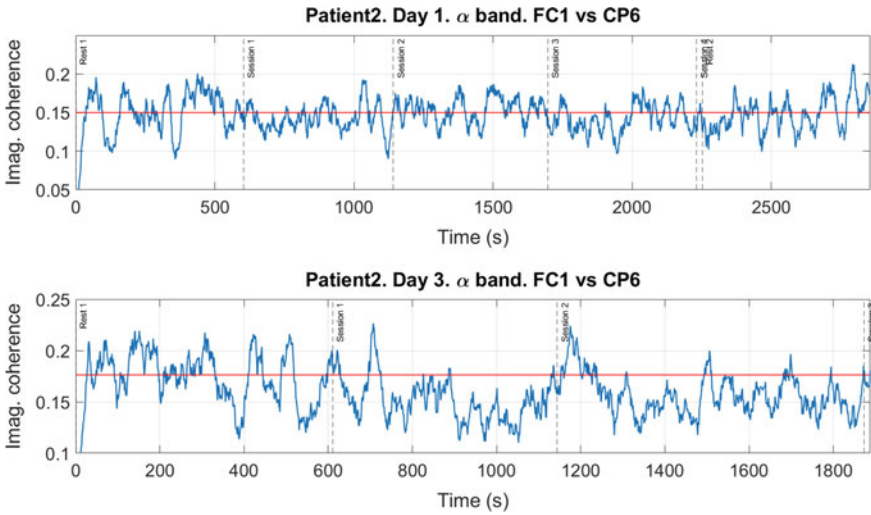


Fig. 29.7 α coherence as a function of time during all sessions for Patient 2. The horizontal red line is the baseline computed as the mean coherence during rest. Vertical dotted lines: start and end session

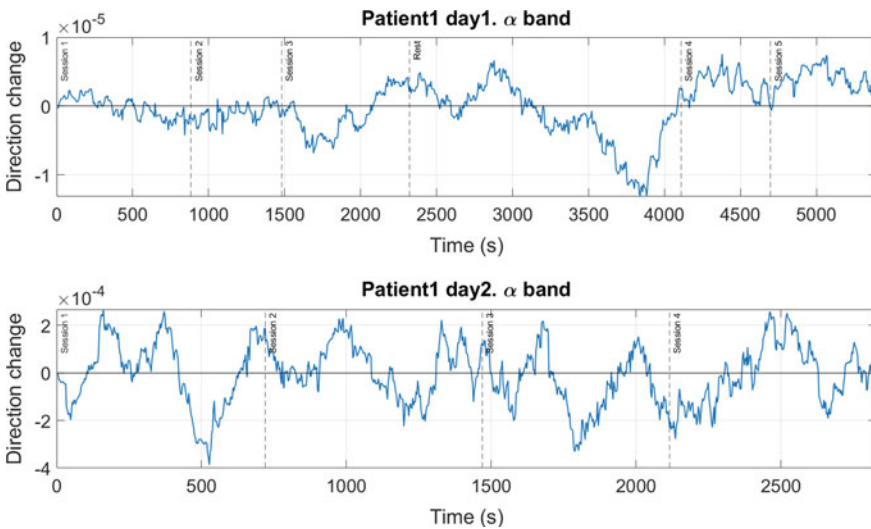


Fig. 29.8 Motion direction changes in the α bands for Patient 1. Time is represented in the x -axis, and y -axis represents the direction changes

During the experiments, the levels of consciousness of Patient 1 are relatively high some days, which would suggest that he/she was attending to the task these specific days. Other days reflect a lower level of consciousness.

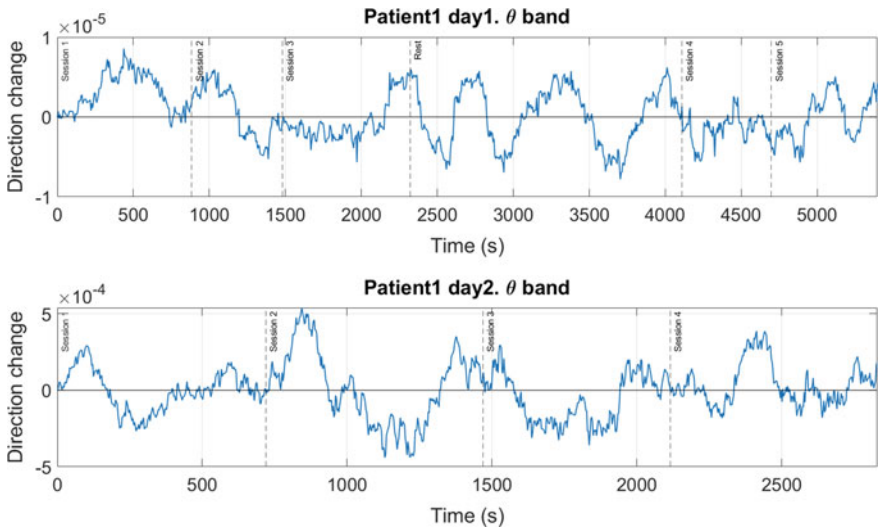


Fig. 29.9 Motion direction changes in the θ bands for Patient 1. Time is represented in the x -axis, and y -axis represents the direction changes

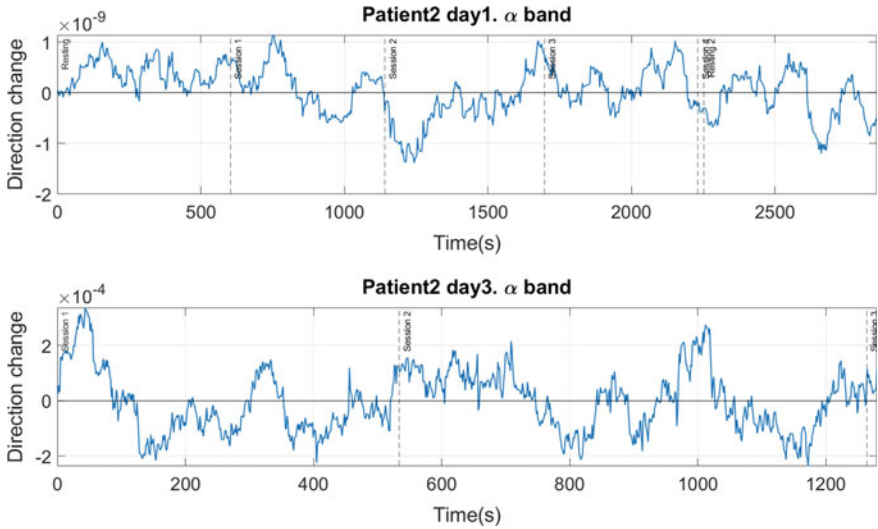


Fig. 29.10 Motion direction changes in the α bands for Patient 2. Time is represented in the x -axis, and y -axis represents the direction changes

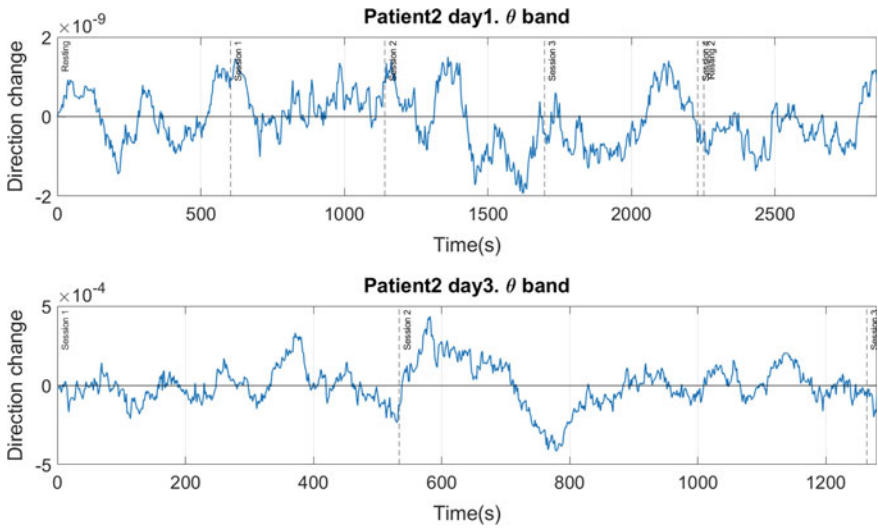


Fig. 29.11 Motion direction changes in the θ bands for Patient 2. Time is represented in the x -axis, and y -axis represents the direction changes

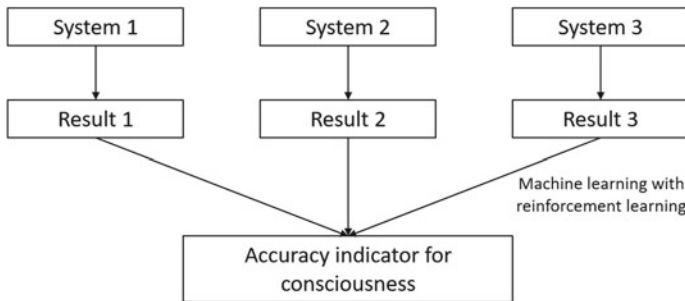


Fig. 29.12 Modus operandi of the proposed system [14]

29.4 Discussion

Lower values of coherence between frontal and parietal channels are believed to correspond to higher states of consciousness. The results show that for Patient 1, the level of consciousness decreases after some sessions (cf. Fig. 29.4). This may be due to decrease in attention or fatigue, which is one of the reasons why it is challenging to perform longer experiments with locked-in syndrome patients. On the other hand, the levels of consciousness of Patient 2 seem to be low for all sessions in Fig. 29.6. In addition, these results are relative to the baseline. It is unclear however if the patients were really “resting” during the recording.

The results from the wave direction sometimes contradict that of the imaginary coherence. For instance, in Fig. 29.5, the highest level of consciousness was estimated during the first 3 sessions, which seems logical. In Fig. 29.8 (top) however, the highest level of consciousness was estimated during rest. Nonetheless, throughout the last two sessions for the same example, both results concur. Low level of consciousness was observed. An indication of the correctness of these estimations could be done by comparing the patients' performance during this time. Unfortunately, this information was not available yet at the moment.

In [14], the results from the imaginary part of the coherence between a frontal and a parietal channel coincide. ECoG was used in that experiment as opposed to EEG in this research. In [15], intracranial and scalp brain signals were simultaneously recorded in rhesus macaque monkeys to compare coherence values obtained from both of them. Their results showed that scalp coherence did not reliably relate to the intracranial coherence in rhesus macaque monkeys even though they were recorded simultaneously. This may explain the slight differences in our results, but remains to be more deeply investigated.

29.5 Conclusion

This paper presents a method based on artificial neural networks and EEG signal coherency to identify consciousness states in two patients with complete locked-in syndrome. Their brain signals were recorded at rest and during task consisting of questions requiring a positive or negative answer that they should answer. Overall, the patients show some level of consciousness during the task according to the methods we used. Nonetheless, these results only show a probability of the patients being aware. To improve patients', relatives' and caregivers' lives, it is important to build a reliable communication system. Future work would consist of building a system like the one pictured in Fig. 29.12 from [14]. Consequently, an approach integrating several approaches would allow a higher probability of uncovering these consciousness states.

Acknowledgements Data were kindly provided by Prof. Dr. Dr. hc. mult. Niels Birbaumer and Dr. Ujwal Chaudhary from the Institute for Medical Psychology and Behavioural Neurobiology, University of Tübingen.

References

1. Laureys, S., Tononi, G.: *The Neurology of Consciousness. Cognitive Neuroscience and Neuropathology*, 2nd edn. Academic, Amsterdam, London (2009). <https://doi.org/10.1016/B978-0-12-374168-4.X0001-9>

2. Ramos Murguialday, A., Hill, J., Bensch, M., Martens, S., Halder, S., Nijboer, F., Schoelkopf, B., Birbaumer, N., Gharabaghi, A.: Transition from the locked in to the completely locked-in state: a physiological analysis. *Clin. Neurophysiol.* **122**, 925–933 (2011). <https://doi.org/10.1016/j.clinph.2010.08.019>
3. Guger, C., Spataro, R., Allison, B.Z., Heilinger, A., Ortner, R., Cho, W., La Bella, V.: Complete locked-in and locked-in patients: command following assessment and communication with vibro-tactile P300 and motor imagery brain-computer interface tools. *Front Neurol.* **11**(251) (2017). <https://doi.org/10.3389/fnins.2017.00251>
4. Chaudhary, U., Xia, B., Silvoni, S., Cohen, L.G., Birbaumer, N.: Brain-computer interface-based communication in the completely locked-in state. *PLoS Biol.* **15**(1) (2017). <https://doi.org/10.1371/journal.pbio.1002593>
5. Adama, V.S., Blankenburg, A., Ernst, C., Kummer, R., Murugaboopathy, S., Bogdan, M.: Motion detection in videos of coherence matrices in order to detect consciousness states in CLIS-patients—an Approach. In: 10th EUROSIM Congress 2019. Logroño, Spain (2019)
6. Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., Hallett, M.: Identifying true brain interaction from EEG data using the imaginary part of coherency. *Clin. Neurophysiol.* **115**(10), 2292–2307 (2004). <https://doi.org/10.1016/j.clinph.2004.04.029>
7. Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-M.: FieldTrip: open source software for advanced analysis of MEG, EEG and invasive electrophysiological data. *Comput. Intell. Neurosci.* (2001). <https://doi.org/10.1155/2011/156869>
8. Priestley, M.B.: Spectral Analysis and Time Series. Probability and mathematical statistics. Academic Press (1989)
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, London (2016)
10. Tensorflow Convolutional Autoencoder. <https://github.com/Seratna/TensorFlow-Convolutional-AutoEncoder>
11. Loy, J.: Neural Network Projects with Python. The Ultimate Guide to Using Python to Explore the True Power of Neural Networks Through Six Projects. Packt Publishing, Birmingham (2019)
12. Kim, P.: MATLAB Deep Learning. With Machine Learning, Neural Networks and Artificial Intelligence. Apress, New York (2017)
13. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. <https://arxiv.org/abs/1412.6980>
14. Adama, V.S., Wu, S.-J., Nicolaou, N., Bogdan, M.: Extendable hybrid approach to detect conscious states in a CLIS patient using machine learning. In: 10th EUROSIM Congress 2019. Logroño, Spain (2019)
15. Snyder, A.C., Issar, D., Smith, M.A.: What does scalp electroencephalogram coherence tell us about long-range cortical networks? *Eur. J. Neurosci.* **48**(7), 2466–2481 (2018). <https://doi.org/10.1111/ejn.13840>

Chapter 30

Computer Vision-Based Demersal Fish Length Measurement Technology



Sheng-Wen Jeng, Chih-Kai Chiu, and Kai-Siang Gan

Abstract This study employed underwater cameras and simple spatial mechanism design to capture real-time demersal fish (e.g., *Epinephelus lanceolatus* or the so-called giant grouper) image stream. Computer vision-based algorithm processes and analyzes the captured image to locate a fish inside a bounding box. Sophisticated shape analysis and machine learning technique were then applied to the bounding box region of image to identify and extract the shape contour of the fishes on the image. Next, this contour was used to obtain the location and length of the fishes in image coordinates. Then, by using the spatial information provided to the system earlier (calibration), the scaling information (mm/pixel) can be used to transform the image length to physical scale. Beside, the fish activity also can be calculated based on the difference between current and previous image frame. The application of the research results can provide information for understanding the growing trend of fishes for fish farmers. In addition, the conversion of fish length and weight and detection of fish activity performed during feeding can serve as critical references for precise feeding management to mitigate feed waste and impede water quality deterioration.

30.1 Introduction

In recent years, global ocean resources have endured heavy damage. Since 2014, the amount of fish obtained through fish farming for direct human consumption has surpassed that by capture fishery and is still growing. Taiwan once claimed the title of aquaculture kingdom because Taiwanese fish farming started early on and was benefited from farming technologies of Japan. However, as more and more countries entered this industry and the cost of aquaculture increased in Taiwan, the competitiveness of Taiwanese aquaculture industry in the global market rapidly

S.-W. Jeng (✉) · C.-K. Chiu · K.-S. Gan
Visual Interaction Department, Smart Microsystems Technology Center, Industrial Technology
Research Institute Southern Region Campus, Tainan, Taiwan
e-mail: itriA40264@itri.org.tw

became recession. Fish farming is highly vulnerable to the influence of environment and climate.

The conventional Taiwanese fish farming method greatly depends on the experience of individual workers, which limits the possibility of quantity- and quality-wise improvement in the industry. With the technological development of big data, Internet of things, and artificial intelligent (AI) images, Taiwan should introduce high-tech into its aquaculture industry and transform it into smart aquaculture industry to resolve current aquaculture-related obstacles. The focus of this study was the application of computer vision to identify and measure target fish underwater to provide benefits of computer vision technology for fish farmers in the future. This study established an equipment of power tool that digitally documents the growth information of fishes and can be integrated with other functions, such as water quality sensing and automatic feeding to make Taiwan an optimized and automatized smart aquaculture kingdom.

30.2 Literature Review

Taiwanese fishery industry is divided into four major categories: distant water fishery, offshore fishery, coastal fishery, and aquaculture. The aquaculture produces most percent of all domestic sales. To improve the aquaculture production capacity and quality and expand the overseas markets, the development of smart aquaculture technology is imperative.

In Taiwan, relevant research on detection technology for aquatic underwater is abundant. With respect to the application of counting, a study by National Sun Yat-sen University established a simple design of automatic counting system for fish larvae [1]. This system uses the gray-level value of the three primary colors as an indicator to identify the mean number of pixels present in a fish larva to estimate the total number of fish larvae from an image. This system uses the method adopted by conventional counting systems, in which counting is performed after identifying individual targets. The new system is beneficial for the counting of fish larvae cultivation. A study applied a fixed mechanically scanned imaging sonar system to determine size of targets and monitor the underwater small mobile objects (e.g., Thai shrimps) [2]. The slant range of this sonar system was set at 7.5 m to detect the environmental features of the field, such as the area, depth, and quality of water and the status of the bottom of the pond. These features can be used to analyze the number, location, movement, and behavioral modes of the target object. The monitoring system established in the aforementioned study effectively aided in describing and understanding the environment of the shrimp-fishing pond and the status of the shrimps.

Regarding the research on fish length, “Simulation Study on an Infrared-light Automatic System for Measuring Fish Size” [3] applied matrix-array infrared-light sensors to measure and analyze the size of a fish passing through nodes of sensors underwater. In Faisal Shafait paper [4], semi-automatic and manual stereo-video

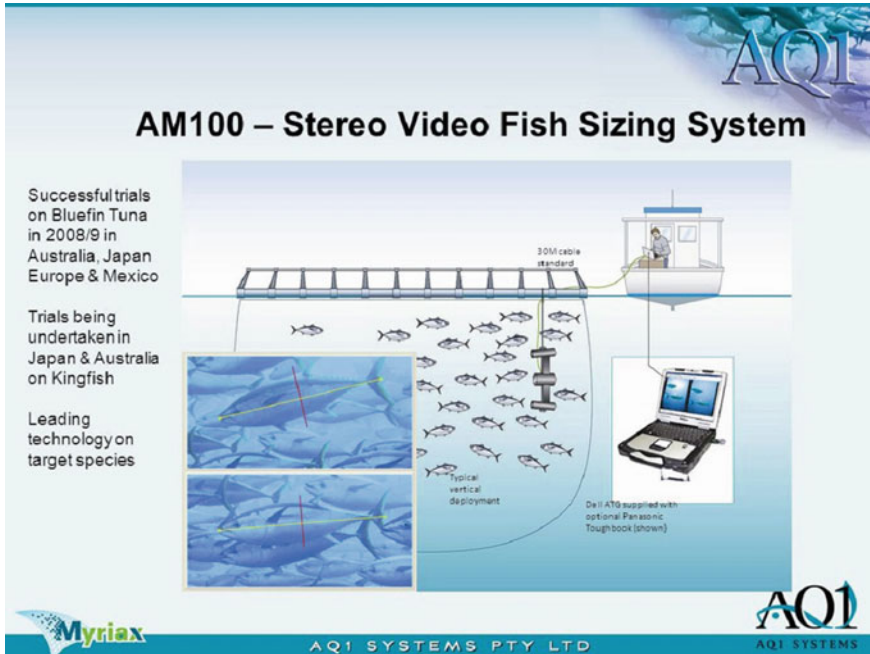


Fig. 30.1 AM100 system concept

measurements of fish length were compared for their efficiency. A company called AQ1 Systems [5] has product “AM100 Fish Sizing and Counting Technology.” Their analysis software allows users to size and count objects underwater. AM100 analysis software can count objects moving past a fixed point, which is great for fish transfer or movement situations or in ecological time-related density analysis. Sizing can be done by single or multiple point-to-point measurements that can be analyzed directly. But, the detail measurement method which is automatic or manual is not disclosure. Figure 30.1 shows AM100 system concept.

In this study, we used a two-dimensional photographic camera to capture images of demersal (benthic) fish in the fishponds. Compounded pixel fusion technology and machine learning were then adopted to effectively reduce the target errors caused by interference of light and noise of the fishpond environment. The correct fish target image than be processed to get accuracy of length measuring.

30.3 The Measuring System Framework

The measuring system framework established in the present study is presented in Fig. 30.2. An underwater camera was lowered into a fishpond, along with a background white plate, to emphasize the contrast between the target objects and the

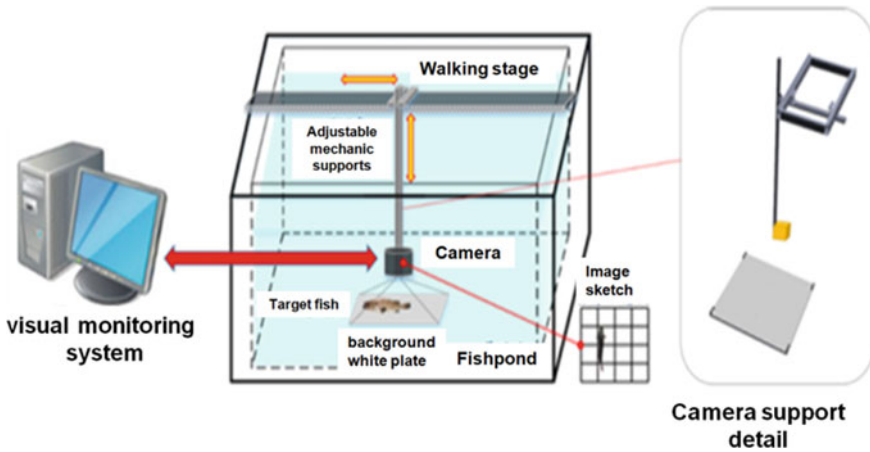


Fig. 30.2 A measuring system framework. The visual monitoring system located at a remote laboratory executes measuring task

background. The video signal of the camera was sent to a visual monitoring system placed in a computer laboratory.

The system functions by connecting the underwater camera to the DVR 4CH 1080P AHD digital recorder and then transferring video signals to fish measuring work station. Figure 30.2 shows a single measuring set in the fishpond. Our system can expand to allow maximum 8 measuring set in the future. The measuring results consequently send to the monitoring server and SQL Server database for save/analysis purpose. The overall fish monitoring/measuring system architecture is presented in Fig. 30.3.

This experiment was set up and conducted in a grouper pond of the Mariculture Research Center, Fisheries Research Institute of the Council of Agriculture, R.O.C. The fishpond is divided into 10 sections, with a 2.6-m-wide central pathway and 85-cm-wide side pathways. Moreover, the pond is approximately 5.5 m wide, 4.8 m length, and 2 m deep, with a wall of approximately 0.2 m and capacity of 50 tons.

30.4 Fish Length Measurement Algorithm

For the investigated technique of underwater fish length measurement, a fixed lens camera was used for image capturing. The condition of image capturing is when fishes are completely observable in the visible range, and the main target objects are the fishes swimming in the pool. When a fish appears under the camera field of view (FOV) with its full shape, the algorithm must be able to identify and capture the swimming fish. Next, the target object (i.e., the fish) should be segmented from the image by using some sophisticated image processing and analysis methods. The core

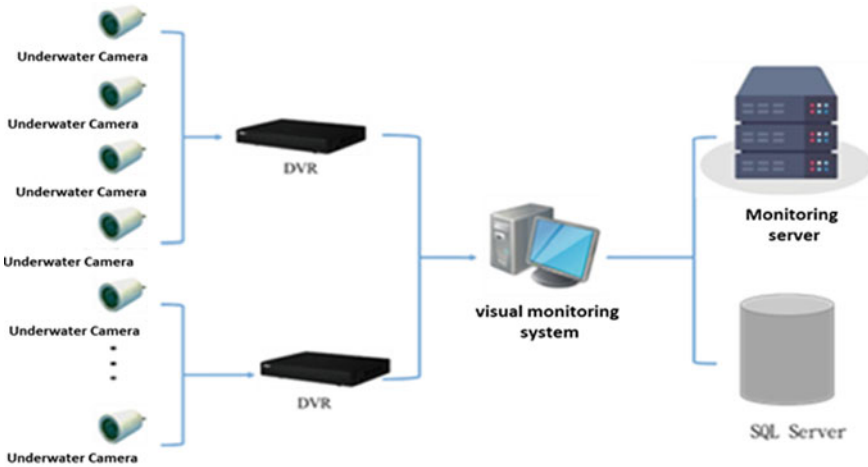
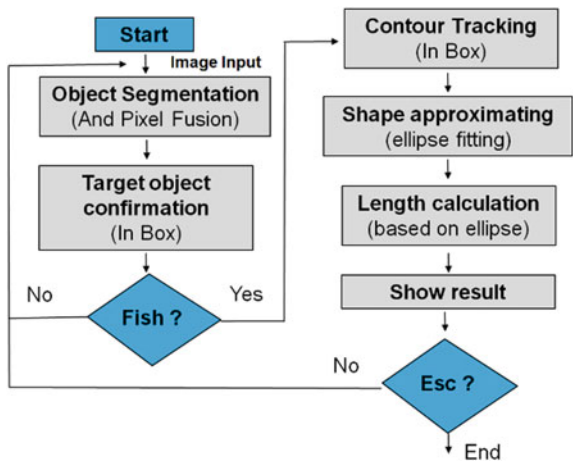


Fig. 30.3 Overall fish monitoring/measuring system architecture

fish length algorithm must then define the shape contour of the fish in this segment region and measure the fish’s body length (from the head to the tail). Figure 30.4 shows the flowchart of the measurement algorithm.

According to the aforementioned function descriptions, the image processing and analysis steps include image preprocessing, target object segmentation and identification, object contour tracking and shape estimation, etc. The identification of fishes can be performed using machine learning techniques such that the target initial locations can be found quickly and set as a bounding box. Furthermore, object tracking techniques may be used to prevent repeated sampling all area of an image but only

Fig. 30.4 Algorithm flowchart



near previous location area. Finally, the best intervals for fish body measurements were identified using the adaptive contour tracking method.

Image preprocessing to extract objects from image (object segmentation) can be performed using several conventional methods, such as the difference matting method. The difference matting method is suitable for static scene images without a solid color background. However, the perturbations and disturbances in the underwater environment can result in light and shadow changes, which may subsequently lead to a wide image noise area and inability to determine the fishes' positions. To conquer this problem, in Prabowo paper [6], objects were extracted by subtraction of a background mode. The motion detection method, another conventional image preprocessing method, can be used to extract moving objects from a series of consecutive images. However, the image extraction effect of the motion detection method can be easily compromised when the target object does not exhibit sufficient movements; this can result in broken extracted images of the target object.

A highly adaptive hybrid pixel fusion technique is proposed in the current study, in response to the deficiencies and misjudgments in image's target region marking (caused by the sporadic activity characteristic of the fishes' bodies). This method can overcome the identification limitations of the conventional single algorithms and effectively enhance the integrity and reliability of fish detection. As illustrated in Fig. 30.5, the brightness and color uniformity of the background images often cannot be maintained over time because of environmental changes (lighting or shadow, etc.). When the difference matting method is employed to resolve this problem, white target objects separated into several regions of the image and thus reduce the effectiveness of discriminating the fishes' boundary and position. By contrast, when the motion detection method is used, insufficient movements of the target object can result in the extraction of a broken shape contour. Thus, a more complete fish image can be obtained through the pixel fusion method—a combination of the two aforementioned methods: The motion detection method is first used to explicitly define the position of the target object, and then, the difference-matted image of the aforementioned position can be used to patch up the broken shape contour. As such, a more complete fish shape image can be obtained (Fig. 30.6).

Furthermore, after object segmentation (object with bounding box, bbox), the target object in the image may be broken into several separated regions of bboxes

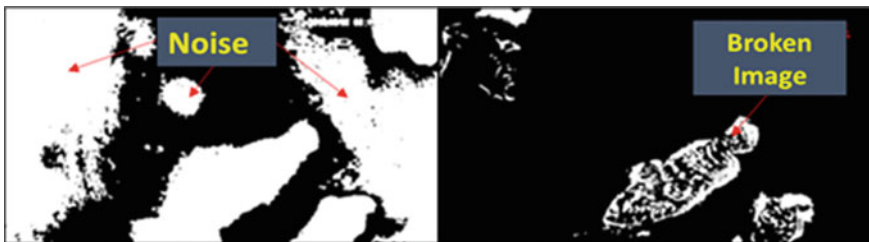


Fig. 30.5 Problem of difference matting (left) and motion detection (right)



Fig. 30.6 Proposed pixel fusion method

for some reasons (fish is not swimming fast enough or body color near background, etc.). Therefore, the complete target object (i.e., the fish) must be post-processed by using the target object identification algorithm; this algorithm comprises two steps.

1. **Target region designation:**

Merge the adjacent target regions that may belong to the same target object and re-label the target regions. Then, for the region which contains the objects that have highest possibilities of being the target objects (fishes), designate it as the target region to be subjected to subsequent processing. Regarding the details of the merging algorithm, this study referred to [7]; the basic concepts of the merging algorithm are illustrated in Fig. 30.7. If the adjacent target regions fit the merging requirements, merge them into a new target region. The simplest method for designating the most probable target region is to identify all target regions fitting the specified length-to-width ratio and select the region with the largest area among these identified target regions.

2. **Target object confirmation:**

The signals in the designated target region may be caused by real fishes or by noises (i.e., those caused by water wave perturbations). Therefore, various computer vision algorithms (machine learning) must be applied to confirm if the objects emitting signals in the target region are fishes, such that the system can decide whether to initiate the subsequent fish body length measurement algorithm. This step eliminates false signals and reduces the system misjudgment rate.

Fig. 30.7 Illustration of the blob merging



(a) Blob before merged

(b) Blob after merged

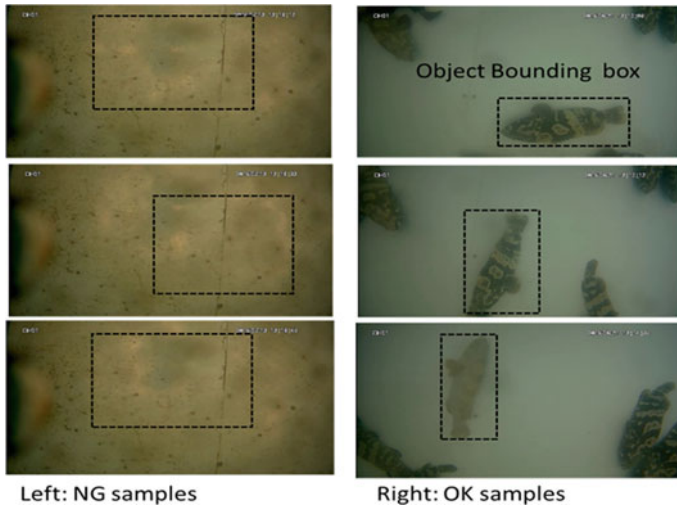


Fig. 30.8 Some sample images with object bounding box for HOG learning

The aforementioned machine learning algorithm adopted at this study can be described as follows: First, various sample fish images are fed manually to the system such that the system can learn about the image features of the fish models. We assume two classes model (fish and non-fish), and learning classes features by feeding about 30 images each class. Figure 30.8 shows some sample images for learning with object bounding box.

In this study, we implement classifier use histogram of oriented gradients (HOG) as features, the source code can be found at OpenCV, since it is became a well-known technique we don't want waste space to describe it. In [8], title "Histogram of Oriented Gradients" is a reference introduction document you can refer it.

The image features would be saved in the data base in the form of parameter files. In this study, file name "cvHOGClassifier.yaml" is used. When executing online identification, the system first segments out the possible target regions from the image (called the preprocessing step); this is followed by the extraction of image features—wherein the same image feature extraction algorithm employed for system learning is used. The identified image features are compared with those of the fish models in the database (by using the decision-making algorithm) such that the system can determine whether the object in the target regions is a fish.

After the target object was confirmed to be a fish, contour tracking was conducted to obtain the shape contour information. The acquired shape contour information was then used for ellipse fitting because an ellipse is similar to the fish body shape; the ellipse fitting was performed to obtain the major axis information, which was a rough estimated fish length in image. Then, two refine regions around two end points of the major axis were used to find exact end points position. The idea is presented in

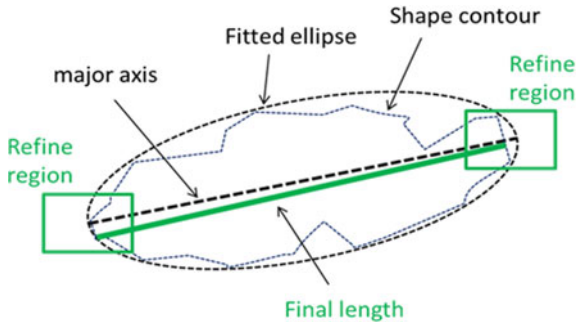


Fig. 30.9 Fish length fine tune

Fig. 30.9. The distance of two exact end points in pixels (final length) was converted to centimeters to be output as measured values.

The converting is as follows:

Assume fish length with image pixels: L_{im} (pixels),
 Camera resolution (mm/pixel): Reso, (pre-setting manually offline).

Then, the real fish Leng, $L_f = L_{im} \times Reso$ (unit: mm), by multiple $\times 10$ to get cm unit. Figure 30.10 shows the measure result with real physical length in cm.

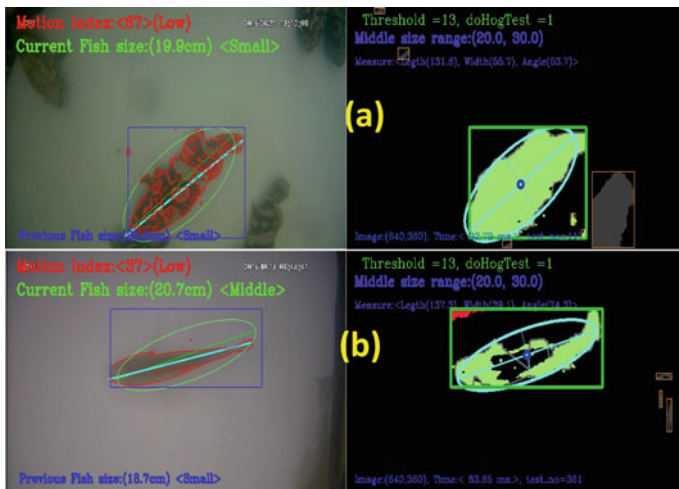


Fig. 30.10 Examples of measure result (Left side, white line segment). Note that a, b are different type of fishes

30.5 Fish Size Classification

Regarding the classification standards for fish body length, two parameter values, namely `middle_low` (i.e., 20.0 cm) and `middle_high` (i.e., 30.0 cm), were used for the classification of fish body length. A fish with a body length lower than the `middle_low` value and higher than `middle_high` is classified as a small and large fish, respectively; a fish with a body length that ranged between `middle_low` and `middle_high` values was classified as a medium-sized fish. The two parameters were specified in the `MeasureSetting.ini` system file to cater to the size of different fish species. In the current results, the approximate error of measurement was ± 2 cm, and the classification accuracy was as high as 97.5% for 200 test samples.

Examples of measure result of this measuring system are showed in Fig. 30.10. The fish length and classification (as small or middle) with processing middle steps are overlay on image at left side graphically. By the way, the fish activity is also showed at up of left side image with “motion index: <67> (Low)” string.

30.6 Conclusion and Future Work

The fish length measurement technology developed in this study is suitable applied to demersal fish (e.g., *Epinephelus lanceolatus* or the so-called giant grouper) in fishpond. The purpose of this system is long-term monitoring of the growing process of fishes with non-disturb way. So, passive measurement is adapted with simple hardware configuration including a camera, support mechanism, and white color background plate. The measurement activated when a complete fish swims into the camera field of view (FOV). The problems with current system include: water becoming dirty, lighting variation, fish type change, etc. So, system parameters and function routines may not cover general conditions for other cases.

Non-demersal fish is not suitable for this 2D camera system, because the distance between fish and camera cannot assure fixed when fish swims up and down. So, the measured length loses reference value, because a same fish near and far from camera will get different lengths. In the future, we will develop 3D stereo measuring system to solve this problem.

References

1. Huang, C.: A simple design of automatic counting system for fish larvae. Master's thesis. Institute of Marine Biology, National Sun Yat-sen University (2002)
2. Tian, W.-M., Tseng, C.-M.: Monitoring techniques for the detection of underwater small mobile objects (shrimp). In: Proceedings of the 39th Ocean Engineering Conference in Taiwan Hungkuang University (2017)
3. Young, C.W., Huang, C.S., Ay, C.: Simulation study on an infrared-light automatic system for measuring fish size. *J. Agric. Mach.* **14**(2) (2005)

4. Jubouri, Q.A.: Towards automated length-estimation of free-swimming fish using machine vision. In: 14th International Multi-conference on Systems, Signals and Devices (SSD) (2017)
5. AQ1 Systems, <http://www.aq1systems.com/farming/13510002>
6. Prabowo, M.R.: A moving objects detection in underwater video using subtraction of the background model. In: Proceedings of the EECSI (2017)
7. Zhou, Y., Li, Y.: A traversing and merging algorithm of blobs in moving object detection. Appl. Math. Inf. Sci. **8**(1L), 327–331 (2014)
8. Mallick, S.: Histogram of oriented gradients, <https://www.learnopencv.com/histogram-of-oriented-gradients/>

Chapter 31

A Haze Removal Method Based on Additional Depth Information and Image Fusion



Tian Tian and Bin Zhang

Abstract To address the problem of image degradation in foggy days, we propose a haze removal method based on additional depth information and image fusion. With recent advances in depth-sensing technology, it has been realized that sensing devices can produce depth images in which the depth value are quite accurate. We adopt the depth estimation dataset of Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) which contains images collected from different real-world environments. The additional information includes the LiDaR scanning points and original depth images which can be used to estimate the optical depth of each point in the scene. In this paper, we investigate how to use additional depth information to remove haze for a single image. Our method focuses on LiDaR depth imaging, image fusion, and the atmospheric scattering model. We use LiDaR scanning points as input and then deduce a rough depth image with prominent features. The rough depth image is then combined with original depth image to improve reliability of depth estimation by image fusion. Using the atmospheric scattering model, we can remove haze for a single image. Experimental results show that our proposed approach provides better performance of dehazing under different fog conditions and holding the details of remote sensing images than current research methods.

31.1 Introduction

Due to low visibility in foggy days, the reflecting and scattering of impurities in the air seriously reduce contrast and clarity of outdoor images causing the visual effect getting awful. As a result, haze removing has become pervasive in applications such as target detection, autonomous driving, and scene recognition. With recent advances in depth sensing technology, it has been realized that sensing devices can produce depth images very easily. Unfortunately, depth-sensing devices often miss data when capturing images which cause the depth value are accurate but incomplete.

T. Tian · B. Zhang (✉)
Xi'an Jiaotong University, Xi'an, China
e-mail: bzhang82@xjtu.edu.cn

The goal of our work is to use additional depth information captured by LiDaR and original depth image to estimate depth of scene and then remove haze for a single image. Though image dehazing has received lots of attention during the recent years, it has generally been solved by different types of methods that recover image by image enhancement or auxiliary information [1–6]. Newer methods have been proposed to remove haze from color images, for example, machine learning [7].

Depth image is widely used as a tool to represent 3D information of scene. Nowadays, with the continuous improvement of depth-sensing technology, additional depth information such as scene depth and multiple images are easy to obtain for practical applications.

According to the different depth sensing devices, collecting scene depth information can be divided into two categories: passive ranging sensing and active depth sensing. [8] The most commonly used method of passive ranging is Binocular Stereoscopic Vision [9]. In this method, two cameras at a certain distance are used to shoot at the same time, so as to obtain two images of the same scene from different perspectives. The corresponding pixel points in the two images are found by the stereo matching algorithm, and the parallax is calculated by the triangle similarity principle and then converted into the scene depth information. However, this method has some limitations on the range and accuracy of parallax map, so the reliability of the scene depth is low. In active depth sensing, the acquisition of depth image is independent of the acquisition of color image. Depth-sensing devices capture 3D information by emitting energy. For example, LiDaR ranging technology can calculate the distance by firing lasers into the space and recording the time interval between the starting point to the surface of objects in the scene and reflecting point back to the LiDaR. It has been widely used in outdoor 3D space sensing system because of its wide range and high accuracy.

31.2 Related Work

The current dehazing algorithms are mainly divided into three categories.

The first category adopts image enhancement to highlight details and improve contrast so as to improve the visual effect of images. There exist algorithms using generalizations of histogram equalization [1] and Retinex [2]. Retinex algorithms have advantages in improving image color constancy and enhancing image details, but it is extremely easy to produce halos when processing images under a condition with a strong light and dark contrast.

The second category adopts image restoration based on auxiliary information, such as using partial differential equation [3], prior information [4] and depth information [5]. Using depth information makes it feasible to deduce the medium transmission and the global atmospheric light which is essential to recover the haze-free images by the atmospheric scattering model [6]. Nonetheless, it has not been used widely because of limitation of depth-sensing devices.

The third category is based on machine learning which focuses on training models for atmospheric conditions [7]. It is worth noting that this method is only suitable for image degradation caused by fog and results in image distortion. In addition, this approach is costly and time-consuming.

Currently, the most simple but effective method is the dark channel priori proposed by He et al. [10]. Since this method cannot deal with the sky region and halo phenomenon very well, it may fail when the haze imaging model is invalid. Moreover, it has a large amount of computation. From the perspective of additional information of scene, Narasimhan et al. proposed a method for scene depth evaluation by discussing the influence law of atmospheric scattering on the contrast of different depths [11]. However, this method only considers the grayscale or color information which causes that the evaluation of depth is not reliable.

31.3 Method

In this paper, we use the additional information of KITTI dataset [12] which includes LiDaR scanning points and original depth images to estimate the scene depth. The process of our method includes LiDaR depth imaging and depth information fusing. First, use scanning points captured by LiDaR to obtain the rough depth image. We project the scanning point onto the original color image. The grayscale pixel value of the projected image can characterize the distance from the camera to surface of the object in the scene. Then, we can obtain a rough depth image with prominent features. Second, aiming at such phenomena as uneven LiDaR scanning points and fuzzy features of objects in original depth images (see Fig. 31.1), we propose to fuse the rough depth image with the original depth image so as to obtain an exact depth image. Finally, calculate the medium transmission from the exact depth image and recover haze-free image using the atmospheric scattering model.

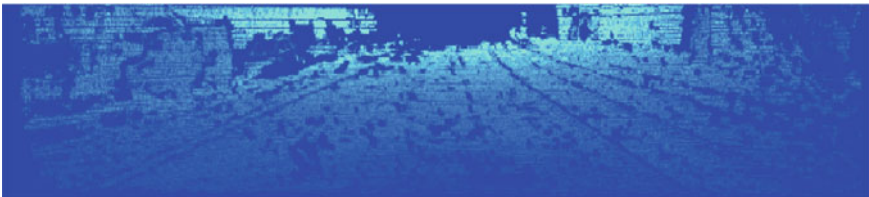


Fig. 31.1 An additional depth image in the scene of KITTI dataset

31.3.1 Haze Imaging Model

In computer vision, aiming at the problem of low visibility in foggy days, Narasimhan [6] explained the process and key elements of imaging by establishing a physical model and proposed that the reasons for image quality reduction include two aspects. On the one hand, it is the energy attenuation caused by the absorption and scattering of reflected light by atmospheric suspended particles. On the other hand, it is the blurring caused by the scattering of ambient light. This model can be described as:

$$E(d, \lambda) = E_0(\lambda)e^{-\beta(\lambda)d} + E_\infty(\lambda)(1 - e^{-\beta(\lambda)d}) \quad (31.1)$$

where E is the irradiance, d is the distance that the light travels, λ is the wavelength of light, $E_0(\lambda)$ is the illuminance of the light source ($d = 0$), $\beta(\lambda)$ is the total scattering coefficient, and $E_\infty(\lambda)$ is the maximum radiation of atmospheric light. The first term on the right side of (31.1) is the attenuation model of incident light, and the second term is the atmospheric light model.

The atmospheric scattering model is [10]:

$$I(x) = J(x)t(x) + A(1 - t(x)) \quad (31.2)$$

where x is the special coordinate of a pixel, I measures the observed intensity, J measures the scene radiance, A measures the global atmospheric light, and t measures the medium transmission describing the portion of the light that is not scattered and reaches the camera.

It can be inferred from (31.2) that the goal of haze removal is to recover J from I :

$$J = \frac{I - A(1 - t)}{t} \quad (31.3)$$

When the fog is uniform, the transmission t is [10]:

$$t(x) = e^{-\beta d(x)} \quad (31.4)$$

where β is the scattering coefficient of the atmosphere and d measures the scene depth. When the transmission $t(x)$ is close to zero, the scene radiance J is on the high side which will cause the image to transition to a white field. Therefore, we shrink the scene depth to a range of 0–1 when using Eq. (31.3) to recover J .

Therefore, the key of dehazing is to estimate scene depth and then substitute the medium transmission into the haze imaging model to recover the target image.

31.3.2 LiDaR Depth Imaging

LiDaR depth imaging is a process in which the scene scanning points projected onto the coordinate system of color image to obtain the rough depth image. The scanning points are stored in the form of point cloud in KITTI dataset.

The process of transforming from point cloud to depth image involves four coordinate systems: world coordinate system (X_w, Y_w, Z_w) , laser LiDaR coordinate system (X_v, Y_v, Z_v) , on-board camera coordinate system (X_c, Y_c, Z_c) , and color image coordinate system (u, v) . The same point has the same depth value in the camera coordinate system and the world coordinate system:

$$Z_w = Z_c \quad (31.5)$$

According to the sensor configuration status in KITTI acquisition platform, the installation distance between LiDaR and camera is 0.27 m. The transformation correlation between the coordinate system of laser LiDaR and the vehicle camera is:

$$X_v = Z_c + 0.27 \quad (31.6)$$

The depth value of points in world coordinate system can be deduced by Eqs. (31.5) and (31.6):

$$Z_w = X_v - 0.27 \quad (31.7)$$

The depth value (Z_w) of each point in scene can characterize the gray value of the depth image. The mapping correlation between point M in LiDaR coordinate system and point N in image coordinate system is [13]:

$$N = P \times R \times T \times M \quad (31.8)$$

where P is the corrected projection matrix, R is the corrected rotation matrix, and T is the translation matrix between LiDaR and camera.

Algorithm 1. LiDaR Depth Imaging

Input: LiDaR scanning points

Output: A rough depth image

(continued)

(continued)

Algorithm 1. LiDaR Depth Imaging

begin

1: Calculate the depth value of points in the world coordinate system:

$$Z_w = X_v - 0.27$$

2: Calculate the mapping correlation of points between LiDaR and image:

$$\mathbf{N} = \mathbf{P} \times \mathbf{R} \times \mathbf{T} \times \mathbf{M}$$

3: Use the pixel value limited in [0, 255] to characterize the depth value of points in image coordinate:

cols = gray;

4: Get a rough depth image.

end

31.3.3 Image Fusion

Even though the rough depth images can obtain relatively prominent object features, depth values are sparse. We notice that the original depth images have more scene depth values, while the object features are not prominent. Aiming at the complementary phenomenon, we propose to fuse the additional depth image and the rough depth image based on image fusion algorithms so as to obtain an exact depth image with prominent features.

Considered that pixel-level image fusion algorithms can retain as much detail information as possible which is conducive to further image analysis and understanding, we adopt Haar wavelet transform [14] in pixel-level image fusion. The steps of image fusion are as follows [15]:

- Wavelet decomposition
 - Use Haar wavelet transform for original depth image and rough depth image, respectively, to establish multi-scale two-dimensional wavelet decomposition.
- Build a wavelet-pyramid
 - Select different coefficients for fusion processing of each decomposition layer to build wavelet-pyramid.
 - For the high-frequency part, select the wavelet coefficients with large absolute value as the coefficient because the wavelet coefficients with large absolute value correspond to the detail characteristic of the objects with significant changes in the gray value of depth image.
 - For the low-frequency part, select the average value of wavelet coefficients as the coefficient.
- Wavelet reconstruction
 - Execute wavelet reconstruction for the wavelet pyramid. The reconstructed image is the fused image which includes more accurate depth value.

Algorithm 2. Image Fusion

Input: Two images (M1 and M2)

Output: A fused image (Y)

begin

1: Establish multi-scale two-dimensional wavelet decomposition using Haar wavelet transform:

[c0, s0] = wavedec2(M1, 3, 'Haar');

[c1, s1] = wavedec2(M2, 3, 'Haar');

2: Build a wavelet-pyramid based on different coefficients.

For the high-frequency part, compare the absolute values and then take the larger one as coefficient:

mm = (abs(MM1)) > (abs(MM2));

Y = (mm.*MM1) + ((~ mm).*MM2);

Coef_Fusion(s1(1,1) + 1:KK(2)) = Y;

For the low-frequency part, take the average value as coefficient:

Coef_Fusion(1:s1(1,1)) = (c0(1:s1(1,1)) + c1(1:s1(1,1)))/2;

3: The image fusion is achieved through wavelet reconstruction:

Y = (mm.*MM1) + ((~ mm).*MM2);

end

31.4 Experimental Results

Our experiments are implemented on a Windows 7 desktop computer with 3.4 GHz Intel Core i7-6700 CPU, 8 GB (space) RAM, and MATLAB R2013a (64 bit).

In this section, we conduct several experiments to verify the advantages of our method in dehazing performance on KITTI dataset. Each scene of KITTI consists of approximately 100 images as an image sequence. We added different concentrations of fog by photoshop to synthesize hazy image and tested on different images in respective sequence. In addition, we compared the results with algorithms of He et al. [10], Zhu et al. [16] (CAP), and Meng et al. [17] (BCCR).

PSNR and SSIM are used as image quality evaluation indexes. The calculation of PSNR is as follows [18]:

$$\text{PSNR} = 10 \times \log_{10} \left(\frac{(2^n - 1)^2}{\text{MSE}} \right) \quad (31.9)$$

where MSE is the mean square error between the original image and the processed image. The larger the PSNR value, the better the defogging effect.

SSIM is used to evaluate the retention degree of image structure information which can be calculated as follows [18]:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (31.10)$$

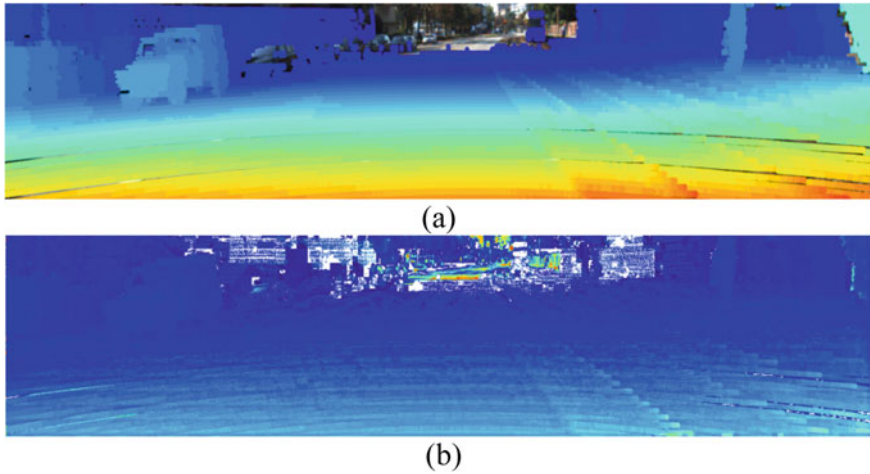


Fig. 31.2 Results of LiDaR imaging and image fusion

where μ_x and μ_y are mean values of image x and image y , σ_x^2 and σ_y^2 are variances of x and y , σ_{xy} is the covariance of x and y , c_1 and c_2 are both constants. The higher the SSIM value, the better the defogging effect.

In the process of LiDaR depth imaging, we use (31.7) and (31.8) to obtain the rough depth image, as shown in Fig. 31.2a. Due to the limited measuring range of LiDaR, the imaging process cannot be realized in the distant region (120 m away). We use image fusion to obtain the accurate depth image, as shown in Fig. 31.2b. In order to appreciate the accurate depth image, we convert the rough depth image and the accurate image to jet colormap.

We use the accurate depth image to deduce the medium transmission and recover a haze-free image based on the atmospheric scattering model, as shown in Figs. 31.3, 31.4, and 31.5. Figure 31.3 shows results of different dehazing methods at fog condition of 80%, and Fig. 31.4 shows results at fog condition of 50%. Figure 31.5 shows partial results in the same sequence which concludes approximately 100 images at fog condition of 50%. In Fig. 31.3, 31.4, and 31.5, (a) shows the original RGB-image with no haze. (b) shows the responding accurate depth image. (c) shows the hazy image with 80% fog concentration. (d) shows results of our method. (e) shows results using method of He et al. [10]. (f) shows results using method of Zhu et al. [15] (CAP). (g) shows results using method of Meng et al. [16] (BCCR). Table 31.1 shows the image quality evaluation index of different algorithms.

According to the dehazing results in Fig. 31.3, 31.4, and 31.5 and the image quality evaluation index value in Table 31.1, we can see that the average PSNR value (average of 100 images in the same sequence) of our method is significantly higher than others. It demonstrated that our method can effectively improve image clarity and avoid color distortion.

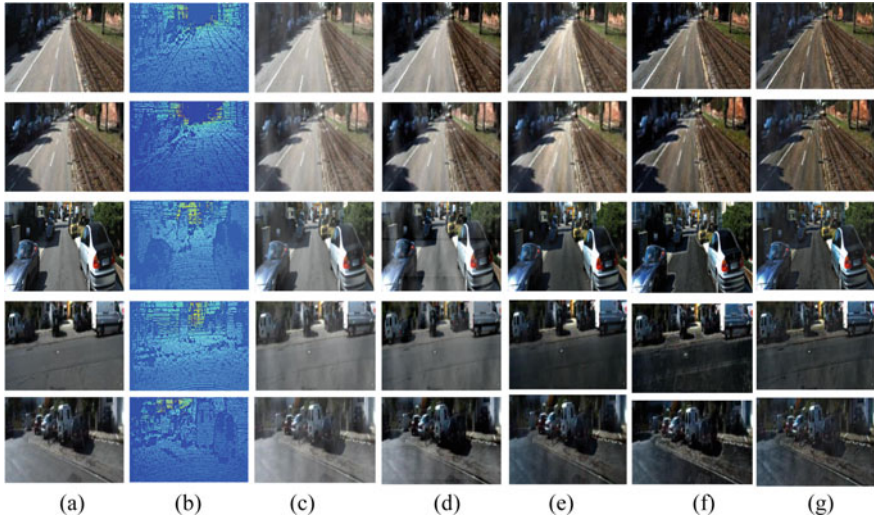


Fig. 31.3 Comparison of dehazing results for images in different scenes with 80% fog concentration

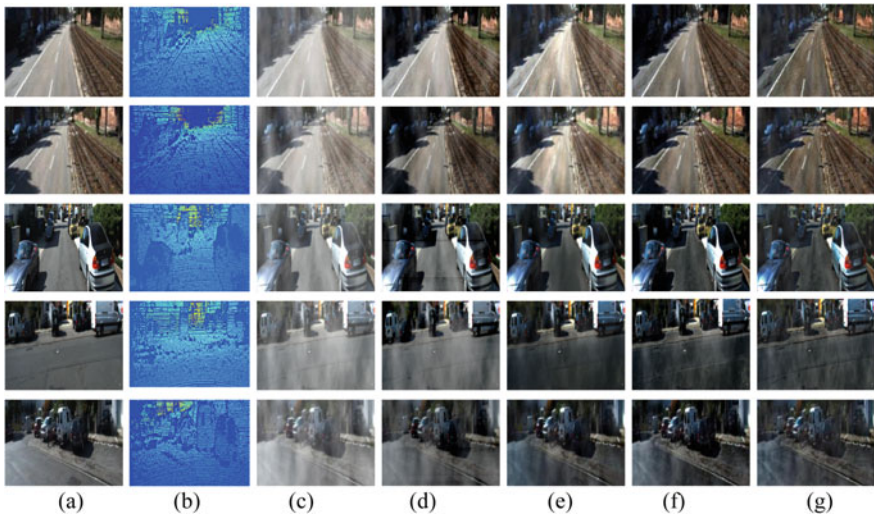


Fig. 31.4 Comparison of dehazing results for images in different scenes with 50% fog concentration

All of the algorithms we mentioned can properly remove the fog, but one main concern is the color distortion (such as surface of road) caused by He et al. [10], Zhu et al. [15] (CAP), and Meng et al. [16] (BCCR). By comparison, the method proposed in this paper can not only effectively remove fog but also avoid image distortion to some extent.

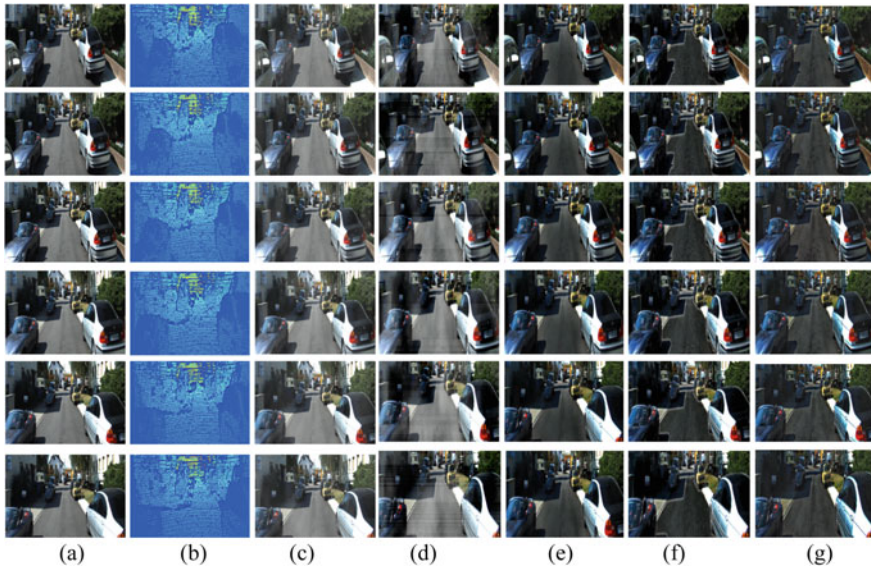


Fig. 31.5 Comparison of de-hazing results of an image sequence with 50% fog concentration

Table 31.1 Quantitative evaluation index value of different methods

Metric	Images	HE	CAP	BCCR	Ours
PSNR	No. 006	24.8198	25.2075	20.4942	27.6063
	No. 027	24.6488	24.7741	21.7419	24.3777
	No. 069	21.6596	22.0607	24.1417	25.4896
	No. 112	22.5323	22.2851	21.9657	25.3335
	Average	23.4151	23.5751	22.0859	26.2021
SSIM	No. 006	0.6827	0.9464	0.9013	0.9504
	No. 027	0.7547	0.9348	0.9258	0.9286
	No. 069	0.8371	0.8783	0.9256	0.9347
	No. 112	0.9259	0.8947	0.9086	0.9268

31.5 Conclusion

In this paper, we propose a haze removal method using depth information provided by the LiDaR and original depth image to obtain a more accurate depth image so as to recover haze-free images based on hazing image model. Experimental results show that our method can not only deal with the close-range and foggy scene but also have good performance to improve the contrast and color saturation of the image. Compared with other algorithms, our method has better performance of enhancing the veins feature and holding the details of remote sensing image. Due to the limitation

of the LiDaR scanning mechanism, our method needs to be improved for operating speed and scene images which are long-range or under non-uniform fog conditions.

References

1. Stark, J.A.: Adaptive image contrast enhancement using generalizations of histogram equalization. *IEEE Trans. Image Process.* **9**(5), 889–896 (2000)
2. McCann, J.: Lessons learned from mondrians applied to real images and color gamuts. In: 7th Color and Imaging Conference, pp. 1–8. (1999)
3. Schechner, Y.Y., Narasimhan, S.G., Nayar, S.K.: Instant dehazing of images using polarization. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 325–332. IEEE, Los Alamitos (2001)
4. Thiebaut, E., Conart, J.M.: Strict a priori constrains for maximum—likelihood blind deconvolution. *J. Opt. Soc. Am. A-Opt. Image Sci. Vision* **12**(3), 485–492 (1995)
5. Kopf, J., Neubert, B., Chen, B., Cohen, M., Cohen-Or, D., Deussen, O., Uyttendaele, M., Lischinski, D.: Deep photo: model-based photograph enhancement and viewing. *ACM Trans. Graph.* **27**(5), 1–10 (2008)
6. Narasimhan, S.G., Nayar, S.K.: Vision and the atmosphere. *Int. J. Comput. Vision* **48**(3), 233–254 (2002)
7. Chen, J., Chau, L.: Heavy haze removal in a learning framework. In: *IEEE International Symposium on Circuits and Systems*, 1590–1593 (2015)
8. CSDN Homepage. https://blog.csdn.net/zuochao_2013/article/details/69904758. Last accessed 11 July 2019
9. Huang, P.C., Jiang, J.Y., Yang, B.: Research status and progress of binocular stereo vision. *J. Opt. Instrum.* **40**(4), 81–86 (2018)
10. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(12), 2341–2353 (2011)
11. Narasimhan, S.G., Nayar, S.K.: Contrast restoration of weather degraded images. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(6), 713–724 (2003)
12. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant CNNs. In: *International Conference on 3D Vision*. IEEE Conference. Qingdao (2017)
13. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. *Int. J. Rob. Res.* **32**(11), 1231–1237 (2013)
14. Li, H., Manjunath, B.S., Mitra, S.K.: Multi-sensor image fusion using the wavelet transform. *Graphical Models Image Process.* **13**(16), 51–55 (1995)
15. CSDN Homepage. <https://blog.csdn.net/Chaolei3/article/details/80961941>. Last accessed 21 July 2019
16. Zhu, Q., Mai, J., Shao, L.: A fast-single image haze removal algorithm using color attenuation prior. *IEEE Trans. Image Process.* **24**(11), 3522–3533 (2015)
17. Meng, G., Wang, Y., Duan, J., Xiang, S., Pan, C.: Efficient image dehazing with boundary constraint and contextual regularization. In: *IEEE International Conference on Computer Vision*, pp. 617–624. ICCV, Sydney (2013)
18. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)

Chapter 32

Vehicle Detection Based on Area and Proportion Prior with Faster-RCNN



Hao Yuan, Bin Zhang, and Ming Liu

Abstract With the development of neural networks, detection accuracy and speed constantly improved. However, the detection effect is still insufficient in some special scenarios such as traffic environment. Therefore, we combine neural network with prior knowledge to improve its performance in vehicle detection. In this paper, we propose two effective prior: proportion and area prior to enhance the vehicle detection ability of neural network in traffic environment. The proportion and area prior is the statistical data of the vehicle at different angles and distances from the camera. In the traffic monitoring video, the proportion of most vehicles is mainly divided into several values. The area of all vehicles is also included between the thresholds. Experimental results demonstrate the effect of prior. Detection effect for vehicle in traffic environment of sample network in this paper increase by 6.56%.

32.1 Introduction

Vehicle detection is the technique that uses computer to simulate human eyes to acquire vehicles in images captured in different scenarios. This technology is highly desired in intelligent transport system and intelligent video surveillance. First, detecting vehicle can significantly increase work efficiency of traffic department and save the manpower. If traffic condition is represented by accurate statistics, more useful information can be read from traffic monitor video. Second, unmanned vehicles [1] also need vehicle detection technology. When unmanned vehicle is running, how to find the barrier ahead is the essential factor. During the last year, unmanned vehicle have developed much, vehicle equipped with a variety of sensors can travel well without driver. But when comes to the environment of road monitoring, only a few

H. Yuan · B. Zhang (✉)
School of Software, Xi'an Jiaotong University, 710049 Xi'an, China
e-mail: bzhang82@mail.xjtu.edu.cn

M. Liu
Command and Dispatch Section of Science and Technology Branch of Traffic Police, Branch of Xi'an Public Security Bureau, 710049 Xi'an, China

methods try to detect vehicle based on aerial image. In recent research, most vehicle detection methods are based on object detection which are mainly divided into traditional methods and deep learning methods.

In this paper, we propose two prior knowledge: proportion and area prior aiming at improving detection effect on traffic environment. Two priors are based on statistics of vehicles in traffic monitor video. We find that, different to images from unmanned vehicle or aerial photography, vehicles in traffic monitor video have multi-proportion and size based on height and angle of cameras on road. Different angles and different distances generate different values of prior. Therefore, vehicles in every part on image have a different prior. If we conclude these values and combine them with network training, we can improve the detect effect of vehicle in on-road monitor video.

This paper has applied method which introduced on the above. We use faster-RCNN object detection framework as example, as experimental data shows, our approach is valid and able to detect distant vehicle in image, which shows the enhancement of detect ability.

The rest of this paper is organized as follows: In Sect. 32.2, it introduces the background of the object detection in past years from two aspects: one is traditional method and the other is object detection by neural network. Recent research on vehicle detection is also introduced. In Sect. 32.3, definition and extraction of prior knowledge are discussed in this part. In Sect. 32.4, the training and prediction method is evaluated and analyzed by the experiment. Section 32.5 draws the conclusion of this paper.

32.2 Related Work

32.2.1 Background on Object Detection

Recent years, the way to detect target in image mainly is divided into traditional method and deep learning method. Two methods are different in many aspects.

In implementation of object detection, traditional algorithm mainly detects the target through hand-designed feature extractor and classifier. When image is inputted, pre-processing such as denoising, image enhancement is performed, and then the possible region of target is selected by sliding window. For each region, its feature is extracted by extractor based on factors like target shape, illumination and background. Specific process is shown in Fig. 32.1a. Feature extractor commonly used are SIFT, HOG, etc. Finally, features extracted were used to trained the classifier (like SVM, AdaBoost) to obtain the category and location information of detected target. In addition, background modeling is also a common method used in video object detection. As shown in Fig. 32.1b, this method firstly obtains the background model through image sequence training, then obtains the target by subtracting new image from model. However, in practical applications, traditional methods still have the following drawbacks:

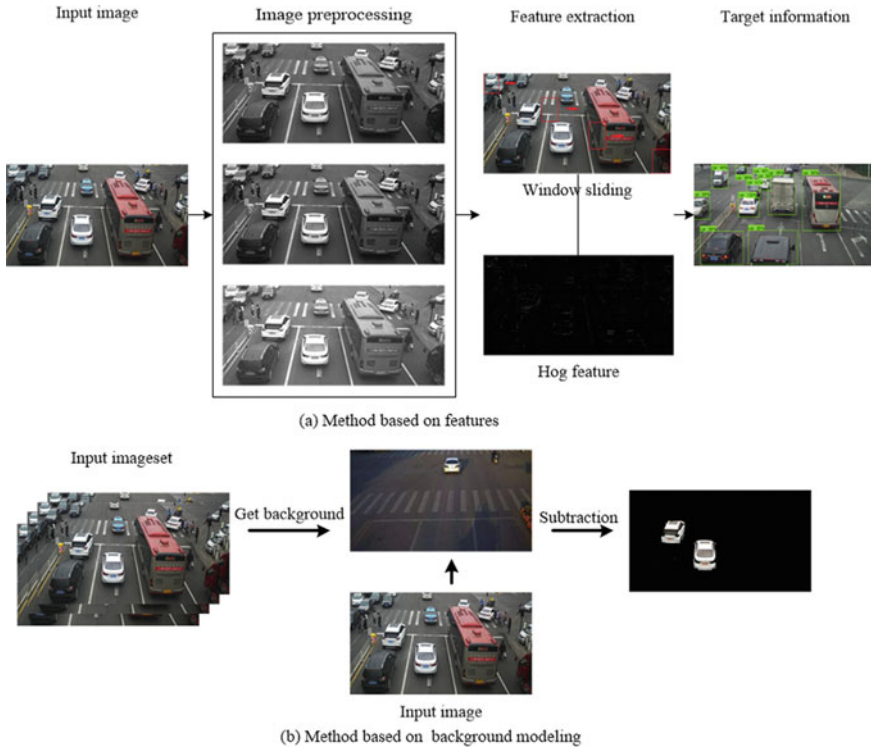


Fig. 32.1 Traditional object detection method

1. Area selection strategy based on sliding window is not targeted for identifying objects which leads to high time complexity and window redundancy.
2. Traditional methods use artificially designed feature extractors, which need to switch between different targets. Similarly, different designs have different effects. Therefore, traditional algorithms are often not robust enough. For example, in 2003, Paul Viola and Michael-Jones use Viola-Jones (VJ) [2] detector in paper published on CVPR. This method uses sliding window to select target region, then extracts Haar feature and uses AdaBoost to classify targets. In pedestrians detection [3, 4], Dalal and Triggs proposed HOG feature [5] and classified it by SVM classifier. In addition, for targets that are generally detachable into multiple different components, the deformable parts model (DPM) algorithm [6] is used. This algorithm divides the object into several components, for example, vehicles are divided into the body, window, wheel and so on.

In addition to traditional algorithms, deep learning methods [7] in real-world object detection [8] are also an important research direction in field of computer vision. In 2006, Hinton proposed deep learning [9], which uses deep neural networks to automatically get high-level features from large volumes of data. Compared with

traditional method, neural network obtains more comprehensive and accurate features, the detection effect improve greatly. Subsequently, Professor LeCun proposed convolutional neural network [10], which further developed deep learning. This network is a hierarchical model consisting of input layer, convolution layer, pooling layer, fully connected layer and output layer. It is specifically designed for image processing. The features received by convolutional layer in network are localized by previous layers. The feature is obtained by convolving shared weights. In convolutional neural network, image is extracted by initial convolution and pooling layers. As the maturity deepens, feature map generated by each layer becomes smaller and the extracted feature level becomes higher. The level of abstraction increases continuously. These features are then classified by fully connected layers and output layers to produce a one-dimensional vector which represents the current picture category. According to function of each layer, convolutional neural network can be divided into two parts: feature extraction part and classifier part.

Later, Girshick et al. proposed the RCNN model [11], which uses selective search to obtain candidate regions from input image, then the candidate regions are converted into uniform sizes and extracted feature by convolutional neural networks. Finally, multiple SVM classifiers are used to achieve multi-object detection.

In order to further improve the detection speed, faster-RCNN [12] adds a network for finding candidate boxes after convolutional layer, named region proposal network (RPN) [13]. By training this network, faster-RCNN can directly get the candidate area. We use faster-RCNN as an example to introduce how prior knowledge [14] improve the detection effect of vehicles [15, 16].

32.2.2 Background of Vehicle Detection

In recent years, due to the success of neural network, vehicle detection has developed rapidly, a variety of detect methods based on neural network sprung out. Some methods aim at detecting vehicle in on-board camera in order to build the driver assistance system [17] or intelligent vehicle [18]. This kind of methods are mainly used in driving environment. In addition, some methods focus on vehicle detection in aerial images like [19, 20], these methods mainly enhance the ability of neural network in detecting small vehicle. Some researchers also improve the detection speed by optimizing the structure of network.

32.3 Prior Knowledge

32.3.1 Definition of Prior Knowledge

In this paper, we use faster-RCNN to detect vehicles in images taken from road surveillance cameras to prove that using prior knowledge in neural network can enhance its effect on vehicle. Following types of features can be used as prior knowledge when detecting vehicles:

Prior knowledge 1: If size of image has not been clipped or scaled by network during training, the aspect ratio of rectangular frame which can contain the entire vehicle is considered as the prior knowledge. If it has been cropped, ratio is calculated based on cropped image.

Prior knowledge 2: When range of target size in training set is large, area of rectangular frame which can contain the target is also the prior knowledge. Size of area of rectangular frame corresponding to minimum target is taken as lower limit, and size of area of largest target corresponds to the rectangular frame is the upper limit.

Above are the proportion and area prior. All prior knowledge can be described by digital quantification. Prior knowledge one can optimize the interest region which generated by RPN in faster-RCNN and then improve the detection accuracy of vehicle. Prior knowledge two improved the size range of vehicle in detection. Through this prior, neural network can detect extremely small or large target better. By introducing features of detection target as prior knowledge, we can enhance the network's tendency to vehicle in different states.

Different values of area and proportion prior are mainly caused by the following factors: (a) Distance between shooting device and vehicle. When distance is large, vehicle area in image is small, vice versa. (b) Angle and height, when shooting device is in different directions of vehicle or at different shooting heights, proportion of vehicle changes greatly. For example, image taken from rear of font has quite different ratios.

32.3.2 Extracting Prior Knowledge from Object

When extracting prior experience as mentioned above, we divide vehicles in image set into five parts: vehicles far away from camera which result in small vehicle area in image, this kind of vehicles represents the minimum area threshold in detection; vehicle close to camera having the largest area in image represents the maximum area threshold; vehicles have same or opposite direction to camera; vehicle which is turning at 45 or 90° to camera. These five types of vehicles basically represent all kind of vehicles in image set.

Therefore, when neural network is training, we first get proportion of vehicle in different situations. Then obtain maximum and minimum values of vehicle area

in image. If data set has more priori value, we can set multiple scale values. As for neural network, automatically acquiring the prior knowledge by algorithm in training process still needs further research.

32.4 Prediction Method

32.4.1 Neural Network Used in Experiment

This paper combines prior knowledge to the training process of faster-RCNN-inception-v2 as example. Specific training process is shown in upper part of Fig. 32.2. During training, prior knowledge is added as a prior condition to the entire learning process, adding a different prior knowledge may generate different models.

In faster-RCNN, network firstly obtains original features from input image through feature extraction network (such as VGG [21] and resnet [22]). Then feeds the feature maps into RPN to get candidate boxes. Rest of network may extract the feature of candidate box and classify them into two outputs: recommended borders and categories (foreground and background). The processing flow is shown in lower half of Fig. 32.2.

This paper mainly modifies the regional proposal network (RPN) with prior knowledge in training process. In faster-RCNN, RPN network generates N candidate boxes in different shapes and sizes for each pixel in feature map through sliding window and then performs subsequent processing. Where N is determined by number of proportion prior and multiple. Calculation method is shown as follows:

$$N_{\text{boxes}} = N_{\text{ratios}} \times N_{\text{scales}} \tag{32.1}$$

Changing the size and shape of candidate boxes generated by regional proposal network by combining two prior knowledge can enhance the detection effect of neural network on vehicle and improve its ability on vehicles whose size is too large/small.

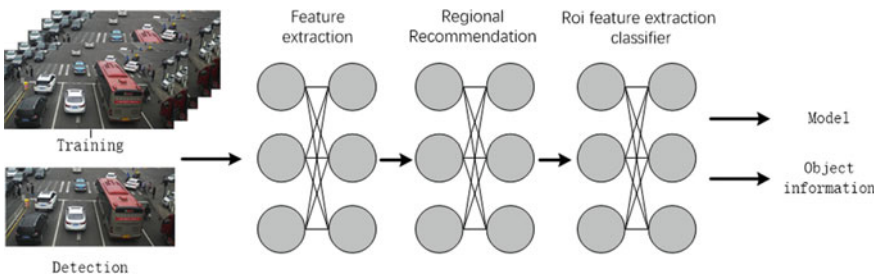


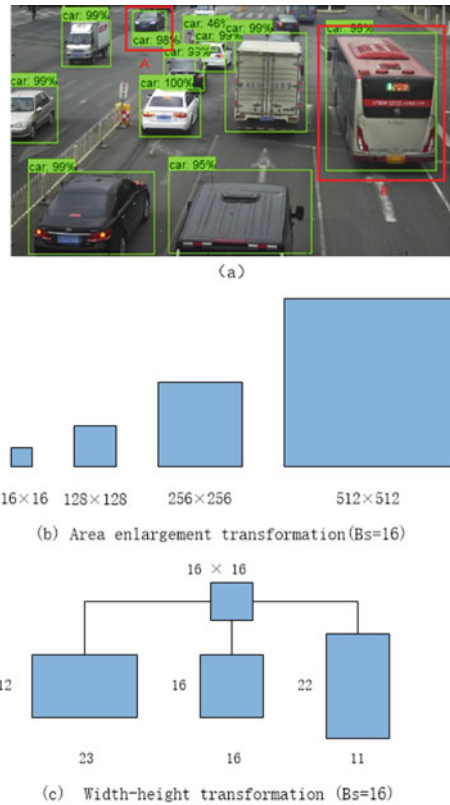
Fig. 32.2 Faster-RCNN training and detection process

32.4.2 Anchors Generation Affected by Area and Size Prior

When starting training, regional proposal network obtains feature map of input image from feature extraction network. Due to multiple convolution and pooling operations, each pixel of feature map represents a considerable area of original image. For each pixel, RPN network generates N candidate boxes corresponding to the corresponding positions of original image according to the set correlation parameters. Area prior is combined in this step to change the area size of candidate boxes. B_s parameter in following formula is the side length of generated minimum square candidate boxes, S is the area of generated candidate boxes, S_a is the magnification of other size of candidates based on B_s . When candidate box is enlarged in area, default candidate shape of boxes is square, and the area of different candidate boxes is the square of B_s , multiplied by the multiple. Amplification process is shown in Fig. 32.3b.

$$S = B_s^2 \times S_a \tag{32.2}$$

Fig. 32.3 Size and area transformation



Taking Fig. 32.3a as an example, vehicle contained by red box in upper left corner has the minimum area, size of this area is set to be the minimum value of area prior. Vehicle in red box on the right side of picture has the maximum area, so we set it to be the maximum value in area prior. Assuming that, minimum area of image is 250 and the maximum is 250,000, we can make B_s , 16 and S_a , 8, 16, and 32, respectively. Thus, after the area is enlarged, the largest candidate box can contain an area having the value of 262,144, which is larger than maximum value in area prior. By combining area priors, network can better detect vehicles that are too small/large.

The proportion prior mainly affects width and height of candidate boxes when the area has been determined. Transformation manner is shown in the following formula. P_r is proportion prior, which determines the aspect ratio of square box after the value of area has been determined by B_s , L_w represents the width of candidate boxes, and L_h is the height. Figure 32.3c is an example of width-height transformation.

$$L_b = \frac{S_a \times B_s}{\sqrt{P_r}} \quad (32.3)$$

$$L_h = L_b \times P_r = \sqrt{P_r} \times S_a \times B_s \quad (32.4)$$

Formula (32.3) represents width of candidate boxes and Formula (32.4) represents the height. S is value of area.

Then, the candidate boxes were sent to subsequent network for classification and scoring.

32.4.3 Neural Network Training

In this paper, sample network is trained with manually annotated vehicle data from road monitoring video. The algorithm is implemented on tensorflow which is an advanced deep learning framework. In network config, we set momentum of 0.9, a weight decay of 0.0005 and learning rate of 0.0001. Hardware used in experiment is a 1080ti GPU and an i7-7700 3.6 GHz CPU. In the RPN part of the network structure, we modified its P_r , B_s , and S_a parameters based on a different prior knowledge.

32.4.4 Experimental Result

In the experiment, there are 1200 images in total to be trained and tested. Among the images, there are 1000 images used for training, and 200 images are used for test. The number of steps for training is 200,000 in total. Through GPU acceleration, the network training has been finished in 20 h. The test result is shown in Tables 32.1 and 32.2. Except parameters combined with prior knowledge, the other parameters in neural network which participated in comparison are consistent. Different network's

Table 32.1 Detection effect of network under different ratios

Network	Ratio	Identification quantity
	0.5, 1.0, 2.0	1083
	1.8, 1.0, 1.1	1150
Faster-RCNN-inception-V2	0.7, 1.2, 1.5	1086
	0.5, 1.5	1113
	0.5, 1.0, 1.5, 2.0	1103
	0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0	1099

Table 32.2 Detection quantity comparison

Network	Identification quantity	Increased percentage
Resnet-101	1089	5.60
Resnet-50	1115	3.14
Faster-RCNN-inception-V2	1150	–

detection effects under different S_a and P_r are evaluated. We run each experiment four times on same data with each parameter and take the average value as final data.

The result of experiment to tests the detection effect with different is in Table 32.1. First row in Table 32.1 is the result of unmodified network. The other lines are results with different ratios. In the test results, average detection quantity of unmodified network is less than other comparative network. In addition to comparison with different values, we also test on other networks, the result is shown in Table 32.2. In Table 32.2, we can see that on same test set, our network is 3.14 and 5.60% more detect than the other two networks. Combing the results of Tables 32.1 and 32.2, we can draw the following conclusions: By adding prior knowledge, the detect effect of the neural network can be enhanced, and the detect ability of distant vehicle can be improved.

As for P_r , we can also find that when value of P_r is closer to actual proportion, network has better performance. When the number of P_r is increased to fit all possible proportions, performance of network is also promoted. In addition, according to prior knowledge acquisition rules, appropriate value of B_s can improve the ability of network on detecting too small/large targets. Figure 32.4 is the detection result



Fig. 32.4 P_r are 0.5, 1.0, 2.0



Fig. 32.5 P_r are 0.8, 1.0, 1.1

of network whose P_r value has not been modified. Figure 32.5 shows the result of network whose P_r value has been modified with prior knowledge. From two results, we can find that network combined with prior knowledge detect more cars in small size. In other words, the network have better performance on extremely small or large vehicle.

In addition, another conclusion can be obtained through the experiment. Compared to using actual proportion values of vehicle as prior knowledge directly, increasing the number of P_r only enhances the detection effect in a small range. Data represented by histogram in Fig. 32.6 shows the influence of different S_a on detection result under the same test set. When P_r is approximately or included in 0.7–1.5, value of detection result is higher. When P_r deviates from actual ratio, network detection ability drops significantly.

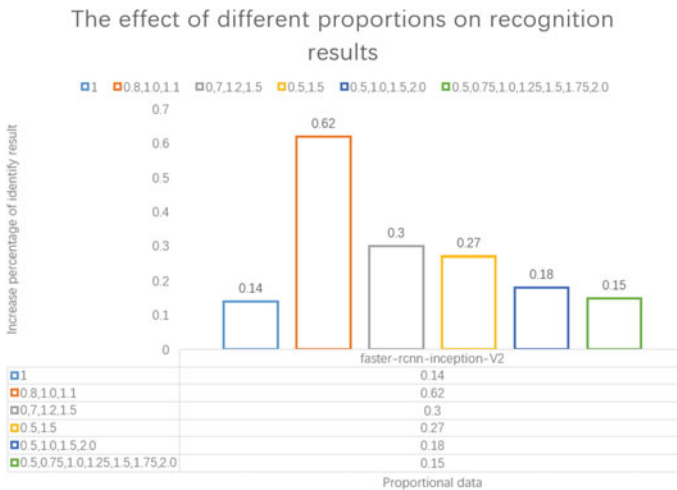


Fig. 32.6 Effect of different proportions on detection results

32.5 Conclusion

In this paper, we propose the method which adds prior knowledge into training process of neural network to enhance its detection ability for on-road vehicle. Proportion and area information of vehicle at different angles are manually extracted from training set and feed to neural network as prior knowledge. When tested on same test set, number of targets detected by modified network increased by 6.56% compared to initial network. From this result, we can reach the conclusion that the ability of network on detecting vehicle can be improved by adding prior knowledge.

In latter work, we will try to make the neural network automatically acquire prior knowledge and optimize it. Attempt will be made to train the network with prior knowledge as training data alone. Other prior knowledge should be introduced more than proportion and area prior. In addition, we will combine prior knowledge with other neural networks to test whether the proposed method is equally valid for most networks. Similarly, pedestrian and multi-object detection will also serve as our future research direction.

Compared with other algorithms, neural networks have very considerable potential advantages in task of object detection. We will try to reduce the network structure when detecting single target later to shorten the training time and improve the detection speed. At the same time, for specific goals, definition and extraction method of prior knowledge are also the future research content. Although the neural network improves the ability to detect vehicles by adding prior knowledge, we are still not sure whether the method is effective for non-vehicle targets. Therefore, this method still has great room for development.

References

1. Sun, Z., Bebis, G., Miller, R.: On-road vehicle detection: a review. *IEEE Trans. Pattern Anal. Mach. Intell.* **5**, 694–711 (2006)
2. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. *CVPR* **1**(511–518), 3 (2001)
3. Zhao, X., Li, W., Zhang, Y., et al.: A faster RCNN-based pedestrian detection system. In: 2016 IEEE 84th Vehicular Technology Conference (VTC-Fall), pp. 1–5. IEEE (2016)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE (2005)
5. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: European Conference on Computer Vision, pp. 428–441. Springer, Berlin, Heidelberg (2006)
6. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Cascade object detection with deformable part models. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2241–2248. IEEE (2010)
7. Schmidhuber, J.: Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015)
8. Ren, S., He, K., Girshick, R., et al.: Object detection networks on convolutional feature maps. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(7), 1476–1481 (2016)

9. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436 (2015)
10. Sermanet, P., Chintala, S., LeCun, Y.: Convolutional neural networks applied to house numbers digit classification. arXiv preprint [arXiv:1204.3968](https://arxiv.org/abs/1204.3968) (2012)
11. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587 (2014)
12. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* 91–99 (2015)
13. Ma, J., Shao, W., Ye, H., et al.: Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimed.* **20**(11), 3111–3122 (2018)
14. Kohara, K., Ishikawa, T., Fukuhara, Y., et al.: Stock price prediction using prior knowledge and neural networks. *Intell. Syst. Account. Financ. Manage.* **6**(1), 11–22 (1997)
15. Huo, Z., Xia, Y., Zhang, B.: Vehicle type classification and attribute prediction using multi-task RCNN. In: *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 564–569. IEEE (2016)
16. Jun, S., Pei, G., Zhili, X., et al.: Vehicle detection based on faster-RCNN. *J. Chongqing Univ.* **40**(7), 32–36 (2017)
17. Huang, D.Y., Chen, C.H., Chen, T.Y., et al.: Vehicle detection and inter-vehicle distance estimation using single-lens video camera on urban/suburb roads. *J. Vis. Commun. Image Represent.* **46**, 250–259 (2017)
18. Zhang, R.H., You, F., Chen, F., et al.: Vehicle detection method for intelligent vehicle at night time based on video and laser information. *Int. J. Pattern Recognit. Artif. Intell.* **32**(04), 1850009 (2018)
19. Tang, T., Zhou, S., Deng, Z., et al.: Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining. *Sensors* **17**(2), 336 (2017)
20. Chen, X., Xiang, S., Liu, C.L., et al.: Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **11**(10), 1797–1801 (2014)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
22. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)

Chapter 33

Mining High-Utility Itemsets of Generalized Quantity with Pattern-Growth Structures



Ming-Yen Lin, Tzer-Fu Tu, and Sue-Chen Hsueh

Abstract High-utility mining is an important issue in pattern discovery. Quantitative high-utility mining further discovers quantitative itemsets that contribute high utility to the database. In this paper, we address the issue of generalized quantity in mining quantitative high-utility itemsets. The relaxation on quantity counting enables novel and generalized findings on quantity combinations of itemsets. We propose an algorithm called QHIM to find out all of the high-utility itemsets of generalized quantity. QHIM maintains a pattern-growth-based structure to store generalized quantity information for the final discovery. The experimental results show that QHIM may efficiently discover these high-utility itemsets. A level-wise method is compared in the experiments. In average, QHIM outperforms the compared algorithm by four times faster for small data and is two orders of magnitude faster than the compared algorithm when the minimum threshold is very small.

33.1 Introduction

Frequent itemset mining [1–4] discovers itemsets that appear with high frequency in transactional databases. However, the utilities of itemsets are not considered in this mining because both the quantity and the utility of items are ignored. A typical customer transaction may have multiple items in various quantities, and each item is associated with a certain level of profit (i.e., utility), such as the profit of “bread and butter” is 5 and that of “birthday cake” is 30. Suppose “bread and butter” occurred in 6 transactions and “birthday cake” occurred in 2 transactions, “bread and butter” has a higher frequency in frequent itemset mining. Nevertheless, the total profit of “birthday cake” is 60, and that of “bread and butter” is 30. The profit contributed by an infrequent itemset like “birthday cake” can be more than that contributed by

M.-Y. Lin · T.-F. Tu

Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan

S.-C. Hsueh (✉)

Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan
e-mail: schsueh@cyut.edu.tw

© Springer Nature Singapore Pte Ltd. 2021

S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation, Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_33

447

a frequent itemset like “bread and butter.” To discover all of the itemsets with high profits is called *high-utility mining* [5–15].

Traditional high-utility mining consists of finding the itemsets with utility values no less than a minimum threshold from the transactional database [5–7]. Table 33.1 shows an example of the transactional database, and Table 33.2 lists the utility of each item. Each transaction contains the associated quantities of the items. For example, the total quantity of item C is 29 and the utility of one C is 3, and this gives the total utility of item C as $29 * 3 = 87$. Other than this calculation, the quantity is useless in traditional high-utility mining. *Quantitative association rule* (abbreviated as QAR) mining [16, 17] has been presented to uncover relationships of itemset quantities in frequent pattern mining, without the consideration of utility. The mining is to extract quantity combinations of itemsets for those itemsets having high frequency. For example, in Table 33.1, the frequency of (A:6) is 2, that of (D:1) is 3, and that of (B:1, D:1) is 2, where (A:6) means the quantity of item A is 6. Those item-with-quantity combinations are called *quantitative itemsets* (or called *qitemsets*). The mining of quantitative association rules is to discover popular qitemsets (i.e., highly frequent), so that promotions on the specific item–quantity combinations are feasible.

One problem of mining QAR is that the quantity values are too diverse. For example, the quantity of item A could be 1, 2, 4, 7, 8, and 10. The support (i.e., frequency) of each combination, from (A:1), (A:2), to (A:10), is calculated independently, such that the number of qitemsets passing the minimum support threshold can be very few. Some studies, therefore, combine a continuous range of quantities to constitute a qitemset of enough support. The discovered qitemsets thus are presented like (A:1–4), (A:8–9, C:6–10), etc. Qitemset (A:1–4) indicates that transactions having

Table 33.1 A transaction database DB with quantities as numbers

	A	B	C	D	E
T01	0	2	4	1	5
T02	2	0	8	0	4
T03	7	0	9	0	0
T04	6	1	0	1	0
T05	0	0	6	0	4
T06	0	1	2	1	0
T07	6	0	0	0	3
T08	2	1	0	0	1

Table 33.2 Utility table

Item	A	B	C	D	E
Utility	5	11	3	20	4

A of quantity ranging from 1 to 4 occurred frequently. Note that the computation disallows the combination of (A:1–4) if either (A:2) or (A:3) is missing.

Despite that the qitemsets can be found, the utility of some qitemsets might be low because only frequency is considered in the mining. Finding qitemsets of high utility is far more interesting for profitable strategies. This problem is called *quantitative high-utility mining* [18, 19], which aims at discovering qitemsets of high utility. The mining, however, suffers from two issues. First, like QAR mining, the quantity values are too diverse so that most studies combine items of consecutive quantities to form qitemsets in the computation of high-utility qitemsets. Consequently, the mining results might not easy to be applied in accurate quantity combination promotions. For example, if (A:8–9, C:6–10) is a high-utility qitemset, should we promote the bundle of (A:8, C:6), (A:9, C:6), or (A:9, C:10)? Combining different quantities solves the issue of quantity-diversity but introduces the difficulty of promotion combinations.

Second, most studies in quantitative high-utility mining may bias toward including itemsets having high quantity with less frequency consideration. That is, an item with large quantity has higher chance to become a high-utility qitemset, with respect to all quantity-values of that item. For example, both qitemsets (C:2) and (C:9) occurred once in Table 33.1. (C:9) may pass the utility threshold of 25 and becomes a high-utility qitemset while (C:2) would fail to be in the answer set. (A:2) occurred twice but it would also fail since its total utility is 20. Intuitively, it seems no problem since total utility is the product of utility, quantity, and frequency. Yet, promoting an item of high quantity is more difficult. The number of customers affordable for (C:2) normally is larger than that for (C:9). If the number of transactions supporting (C:2) is high, promoting this qitemset is more attractive and profitable as a consequence. This leads to the motivation of the study, mining high-utility itemsets of generalized quantity.

By looking into the transactions, a transaction (A:6, B:1) means the customer buys 6 of item A and 1 of item B together. It supports the counting of (A:6), (B:1), and (A:6, B:1). Given a transaction (A:6), it can be interpreted as (A:2, A:4). That is, a customer buys 6 of item A should be generalized as supports the counting of (A:2) and (A:4) in this case. A customer buys a larger quantity L can be interpreted as the customer is able to buy any quantity no more than L . When we are counting the support of an item having quantity q , any transaction containing the item at least quantity q may “contribute” to the support counting. With such a generalized view, (A:2) would occur 5 times in total in Table 33.1, with T02 and T08 having (A:2), T03 having (A:2, A:5), and T04 and T07 having (A:2, A:4). Note that (A:1), (A:3), (A:4), and (A:5) are missing in all the transactions so that the support counting is not performed against these combinations. In other words, only the supports of (A:2), (A:6), and (A:7) have to be considered. In this way, the total utility of (A:2) is $5 * 2 * 5 = 50$, that of (A:6) is $3 * 6 * 5 = 90$, and that of (A:7) is $1 * 7 * 5 = 35$. When the minimum utility threshold is 70, only (A:6) is a high-utility qitemset. The (A:6) promotion can be exercised since both the quantity and the frequency are considered. Overall, customers purchasing 7 of item A would be happy to receive a promotion of (A:6) and purchasing additional one A. Hence, we adopt the generalized quantity for support computation in the mining of high-utility qitemsets so that the above two

issues can be handled. Specific item–quantity combinations rather than quantity–ranges can be decided, and the mining is not biased to high quantity. Only if the itemset has a certain amount of quantity and frequency, the itemset may become a high-utility qitemset.

In this paper, we propose an algorithm called quantitative high-utility itemset mining (QHIM) to discover all high-utility itemsets of generalized quantity. QHIM constructs a QU-tree to store the database information. We extend the pattern-growth methodology to discover all the candidate high-utility qitemsets with the generalized definition of quantity. Only two database scans are required to discover all of the candidate itemsets. Most of the high-utility mining algorithms need one more database scan to identify the actually high-utility itemset after discovering candidate itemsets. QHIM is able to identify the actual high-utility qitemsets without additional scan because it may calculate the actual utility value of the qitemsets directly. In the experimental evaluation, a level-wise algorithm to mine the qitemsets of generalized quantity is designed to compare with the QHIM algorithm. The experimental results show that QHIM is, on average, four times faster than the naïve algorithm in the small dataset. QHIM is even hundreds of times faster than the naïve algorithm when the minimum utility threshold is small for a large dataset.

The remainder of this paper is organized as follows. Section 33.2 presents the related studies. We define the problem of high-utility qitemset mining of generalized quantity in Sect. 33.3. Section 33.4 describes the QHIM algorithm in detail. The experimental evaluations are discussed in Sect. 33.5. Section 33.6 summarizes this study.

33.2 Problem Definition

Let $\Psi = \{i_1, i_2, \dots, i_n\}$ be a set of items. The *utility* of item i_p , denoted by $e(i_p)$, represents a number measure of the usefulness of i_p . The utilities of all of the items are collected in a table called the *utility table*. A transaction database $DB = \{T_1, T_2, \dots, T_m\}$ contains m transactions, where each transaction T_q is a subset of Ψ , and the *quantity* of item i_p in transaction T_q is denoted by $o(i_p, T_q)$. A *quantitative itemset* X is $\{i_1:q_1, i_2:q_2, \dots, i_g:q_g\}$, where $i_x \in \Psi$, q is one of the quantities of the item i_g in DB , and g is quantitative itemset length. Given a quantitative itemset $Y = \{i_1':q_1', i_2':q_2', \dots, i_h':q_h'\}$ where $h \geq g$. We say that Y contained X if Y exists as one subset equal to X . The itemset utility of quantitative itemsets

$$iu(X) = \sum_{i_g \in T_q} (e(i_g) * Q_g) \quad (33.1)$$

The total utility of quantitative itemset X

$$tu(X) = iu(X) * F(X) \quad (33.2)$$

$F(X)$ is the number of the quantitative itemset X in DB where the quantity in the transaction is no less than the quantity in the X .

For example, we are given $\Psi = \{A, B, C, D, E\}$, a transactional database DB with the quantities displayed in Table 33.1, and the utility table in Table 33.2. The *quantity* of item A in transaction T02, $o(A, T02)$, is 2. Thus, $\{A:2\}$ is a quantitative itemset, but $\{A:4\}$ is not because $\{A:4\}$ is not contained in the DB, as the quantities of A in the DB are just 2, 6, and 7. The *utility* of item A, denoted by $e(A)$, is 5, as shown in Table 33.2. The itemset utility of $\{A:2\}$ $iu(\{A:2\}) = 5 * 2 = 10$. The total utility of $\{A:2\}$ $tu(\{A:2\}) = iu(\{A:2\}) * F(\{A:2\}) = 10 * 5 = 50$ because there are five transactions containing item A, and the quantity is no less than 2. Therefore, $F(\{A:2\}) = 5$. In the other example, the itemset utility of the quantitative itemset $\{C:2, D:1\}$ is $iu(\{C:2, D:1\}) = 3 * 2 + 20 * 1 = 26$. The $F(\{C:2, D:1\}) = 2$ because T01 and T06 contain the C, D and the quantity is no less than the $\{C:2, D:1\}$. Therefore, $tu(\{C:2, D:1\}) = 26 * 2 = 52$.

The transaction *utility* of T_q is defined as

$$u(T_q) = \sum_{i_p \in T_q} (o(i_p, T_q) * e(i_p)). \tag{33.3}$$

The *total database utility* is the sum of all of the transaction utilities in DB, i.e.,

$$u(DB) = \sum_{T_q \in DB} u(T_q). \tag{33.4}$$

A quantitative itemset X is a *quantitative high-utility itemset* if the total utility of X is greater than or equal to the *minimum database utility* $\rho = (\delta * \text{total database utility})$, where δ is a user-specified *minimum percentage threshold*. Otherwise, X is a quantitative low-utility itemset for $tu(X) < \rho$. The goal of mining is to find the set of all of the *quantitative high-utility itemsets* in the DB.

From the previous example, the *transaction utility* of T02 is $u(T02) = 10 + 24 + 16 = 50$. Table 33.3 shows the transaction utility for each transaction in Table 33.1. Thus, the *total database utility* is 385. Given $\delta = 10\%$, the minimum database utility is $\rho = 385 * 10\% = 38.5$. Quantitative itemsets $\{A:2\}$ and $\{C:2, D:1\}$ are quantitative high-utility itemsets because the total utility of $\{A:2\}$ is 50 and $\{C:2, D:1\}$ is 52. Both total utility values for quantitative itemsets are greater than the minimum database utility (38.5).

Table 33.3 Transaction utilities in the example database

TID	Transaction utility	TID	Transaction utility
T01	74	T05	34
T02	50	T06	37
T03	62	T07	42
T04	61	T08	25

In frequent itemset mining, the search space can be minimized by the downward closure property: any superset of an infrequent itemset cannot be a frequent itemset. In high-utility mining, however, the superset of a low-utility itemset cannot be pruned, for instance, if the minimum database utility is 38.5. Itemset $\{C:2, D:1\}$ is a quantitative high-utility itemset because its total utility is 52, whereas itemset $\{C:2\}$ is not a quantitative high-utility itemset because $F(\{C:2\}) = 5$. Therefore, $tu(\{C:2\}) = 6 * 5 = 30$. The total utility of itemset $\{C:2\}$ is less than the minimum database utility. In high-utility mining, most of the research solves the downward closure property problem by the transaction-weighted utility framework [6, 7]. The property also holds in the quantitative high-utility mining [18, 19].

In the transaction-weighted utility framework, the *transaction-weighted utility* of a quantitative itemset X , which is denoted by $TWU(X)$, is defined as the sum of the transaction utilities of all of the transactions containing quantitative itemset X . That is,

$$TWU(X) = \sum_{T_q \in DB \wedge X \subseteq T_q} u(T_q). \quad (33.5)$$

For example, $TWU(\{C:2\}) = u(T01) + u(T02) + u(T03) + u(T05) + u(T06) = 74 + 50 + 62 + 34 + 37 = 257$ and $TWU(\{C:2, D:1\}) = u(T01) + u(T06) = 74 + 37 = 111$. When the *transaction-weighted utility* of a quantitative itemset X is greater than or equal to the minimum database utility, i.e., $TWU(X) \geq \rho$, X is called a *quantitative high TWU itemset* (or *QHTWU itemset*).

Lemma 1 *Quantitative high-utility mining holds the two downward closure property.*

Lemma 2 *Quantitative high-utility itemsets must be a quantitative TWU itemset.*

Because the transaction-weighted utility of X is no less than its total utility, i.e., $TWU(X) \geq tu(X)$, if $TWU(X) < \rho$, then X cannot be a quantitative high-utility itemset and can be pruned. Moreover, the *transaction-weighted utility framework* has a downward closure property: any subset of a *QHTWU* itemset must be a *QHTWU* itemset. Therefore, the transaction-weighted utility can be employed as a pruning strategy in quantitative high-utility mining. In addition, if X is a quantitative high-utility itemset, X must be a *QHTWU* itemset because when $tu(X) \geq \rho$, $TWU(X) \geq \rho$. The set of all of the quantitative high-utility itemsets is a subset of the set of all of the *QHTWU* itemsets.

33.3 Related Works

The mining of the quantitative association rule was proposed in 1996 [16]. Quantitative association rule mining was tried to discover if most of the customers were purchasing some specific quantity of an item at the same time. Each item with a different quantity will be regarded as a different item we call a quantitative item. If the

frequency of the quantitative item or itemset is larger than the user-specific minimum frequency threshold, the item or the itemset is frequent. Quantitative association rule mining can discover which quantity combinations of itemsets are popular. Because the frequency of each of the quantitative itemsets is smaller than the total frequency of the itemsets, we must reduce the minimum frequency threshold to discover the quantitative itemsets. However, if we reduce the minimum threshold, we will find many unnecessary results. To avoid discovering too many unnecessary itemsets, [17] combined the frequency of the quantitative itemsets to become one itemset. The quantity of the quantitative itemsets will form in a quantity range. The notation $\{A:3-6\}$ means quantitative itemset A with a quantity range of 3–6. However, the combination may generate a large range quantitative itemset. In real-world applications, use of a large range quantitative itemset to establish business strategies is difficult.

The high-utility mining was proposed for discovering which itemsets have a high-utility value. High-utility mining can find the itemsets that have a low frequency but contribute high utility. The first efficient algorithm UMining was proposed by [7]. UMining calculates the upper bound of each itemset to identify whether the itemset is a candidate high-utility itemset. However, UMining does not solve the problem of downward closure property. The two-phase [6] method presents a transaction-weighted utility (abbreviated as TWU) to solve the downward closure property. The downward closure property can be used to prune candidate patterns with the transaction-weighted utility. In the first phase, two-phase uses the TWU downward closure property to find all of the high TWU itemsets from the database. The second phase scans the database once to identify the truly high-utility itemsets from the high TWU itemsets. High-utility mining considers the item quantity, but the relationship of itemset quantity is not to be considered.

Quantitative high-utility itemset mining algorithm HUQA [19] follows the definition of the quantitative association rules mining. HQUA uses two minimum utility thresholds to find quantitative high-utility itemsets and weak high-utility itemsets. If the utility value is larger than the minimum utility threshold by combining the utility of a continuous quantity of the quantitative items, the continuous quantitative items are combined into one quantitative item. Although they combined the quantitative itemset to avoid reducing the minimum utility threshold, the large range or fragment problems still possibly occur in the algorithm. Moreover, long length quantitative itemsets have a high chance of becoming quantitative high-utility itemsets to form the unbalanced chance of the itemset becoming high utility because the long length quantitative itemset contains a bigger utility value.

33.4 Proposed Algorithm

In this section, we will describe the algorithm QHIM quantitative high-utility itemsets mining (QHIM) algorithm. QHIM is a pattern-growth-based algorithm that consists of two steps. The first step is to construct the global header table and global quantitative utility tree (QU-tree) to store database information. The second step is to discover

all of the QHTWU itemsets by generating a local QU-tree and local header table. The quantitative high-utility itemsets can be identified directly when the QHTWU itemsets have been discovered in the second step. Therefore, we do not need another database scan to identify the actual quantitative high-utility itemsets.

33.4.1 Global Header Table and QU-Tree Construction

To construct the QU-tree, we need to scan the database twice. The first scan is to calculate the TWU value of each quantitative item and build a global header table. The global header table contains five parts: {item ID, quantity, TWU value, TID list, QU-tree pointer}. The TID list not only contains the actually TID of the quantitative item but also contains the TID of the quantity larger than the quantitative items. For example, the header table of quantitative item {C:2} consists of {C, 2, 257, (T01, T02, T03, T05, T06), QU-tree pointer} in the global header table. The QU-tree pointer is pointing to the {C:2} node in the QU-tree. The header table can be ordered by alphabet, TWU value ascending, TWU value descending, or other methods. In this paper, we use the alphabet and the quantity descending order to construct the header table. When the first database scan is completed, we have constructed the global header table shown in Table 33.4. Suppose the minimum database utility is 60.

After we finished the first database scan, we had constructed the global header table. Next, we must build the global QU-tree to discover all of the QHTWU itemsets.

Table 33.4 Global header table

Item ID	Quantity	TWU	TID
A	7	62	T03
A	6	165	T03, T04, T07
A	2	240	T02, T03, T04, T07, T08
B	2	74	T01
B	1	197	T01, T04, T06, T08
C	9	62	T03
C	8	112	T02, T03
C	6	146	T02, T03, T05
C	4	220	T01, T02, T03, T05
C	2	257	T01, T02, T03, T05, T06
D	1	172	T01, T04, T06
E	5	74	T01
E	4	158	T01, T02, T05
E	3	200	T01, T02, T05, T07
E	1	225	T01, T02, T05, T07, T08

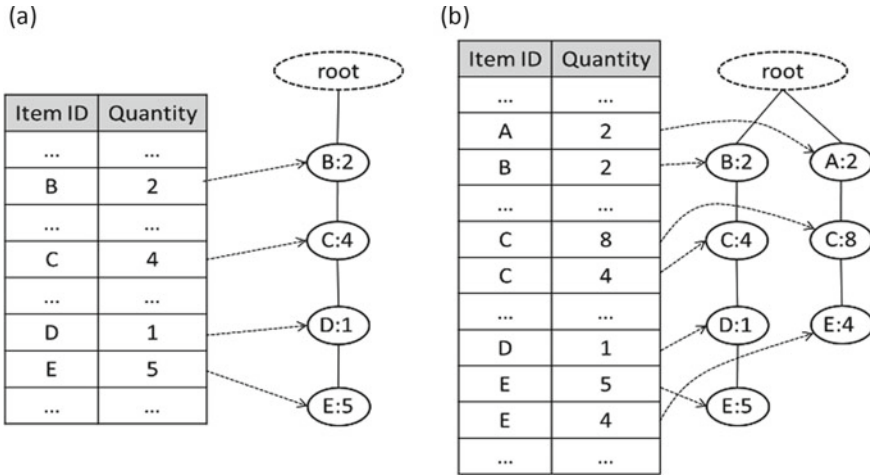


Fig. 33.1 a Global QU-tree (T01), b global QU-tree (T02)

Four elements are contained in the QU-tree: header table, tree node, path link, and item link. Tree nodes are the quantitative items. Path link links to the tree node that occurs in the same transaction. The item link is linked from the global header table and connects the tree nodes that stored the same quantitative item.

The QU-tree is built transaction by transaction when we scan the database the second time. For example, when we scanned the first transaction T01 from Table 33.1, T01 contained four quantitative items: {B:2}, {C:4}, {D:1}, and {E:5}. We built a corresponding path in the global QU-tree. The global QU-tree after scanning T01 is shown in Fig. 33.1a. There is one path in the global QU-tree. Next, we scan T02. Because T02 and T01 do not have the same prefix nodes, we construct a new path to store the T02. Figure 33.1b is the global QU-tree, where we scanned the T01 and T02. After we scan the database, we will construct a complete global QU-tree. Figure 33.2 is the complete Global QU-tree.

33.4.2 Discover QHTWU and Quantitative High-Utility Itemset

After two database scans in the first step, we have completed the global header table and the global QU-tree. The second step discovers all of the QHTWU itemsets. We generate the conditional local header table and local QU-tree based on each quantitative item bottom-up in the global header table. The structure of the local header table is the same as the global header table. However, all of the information is based on the condition of quantitative items.

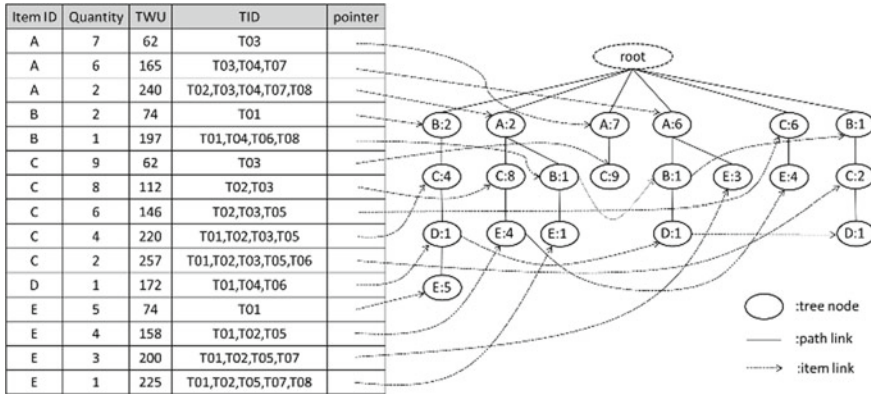


Fig. 33.2 Global QU-tree. (After database scan)

We follow the example of Fig. 33.2. First we make a projection based on {E:1}. We trace the {E:1} item link from the global table. There is one path containing {E:1}. The path contains three quantitative items: {A:2}, {B:1}, and {E:1}. We show the {E:1} projection in Fig. 33.3a. Next, we build the local header table based on {E:1} because each quantitative item in the local header table must occur with {E:1}. Therefore, we AND each TID list in the local header table with {E:1}. For example, the TID list of {A:2} is (T02, T07, T08) that was AND by the {A:2} and {E:1} TID lists. Base on the TID list, the TWU value of {A:2} in the local header table is 117.

After we constructed the local header table and the QU-tree based on the quantitative item {E:1}, we can discover the quantitative itemsets {A:2, E:1} containing TWU value 117 and TID list (T02, T07, and T08); {B:1, E:1} containing TWU value 99 and the TID list (T01 and T08) because both TWU values of two quantitative itemsets are larger than minimum database utility (60). In the local header table and QU-tree, if we cannot directly discover all of the QHTWU itemsets, we may recursively generate the header table and the QU-tree by the local header table. From the {E:1} local header table, we can generate the projected and the conditional

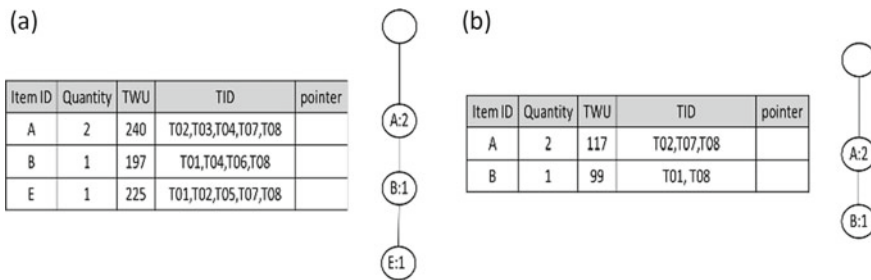


Fig. 33.3 a {E:1} Projection QU-tree, b local header table, and QU-tree ({E:1})

header tables by $\{B:1\}$. The local header table is based on $\{B:1, E:1\}$. Therefore, we can discover the quantitative itemset $\{A:2, B:1, E:1\}$. Because the TID list of $\{A:2, B:1, E:1\}$ is (T08) and the TWU value is 25, $\{A:2, B:1, E:1\}$ is not a QHTWU itemset.

In the mining step, once the QHTWU itemset has been discovered, we can directly calculate the actual utility of the quantitative itemset. For example, the itemset utility $iu(\{A:2, E:1\})$ is $2 * 5 + 1 * 4 = 14$. The TID list of $\{A:2, E:1\}$ is (T02, T07 and T03). From TID list, the count of $\{A:2, E:1\}$ is 3. Therefore, the total utility $tu(\{A:2, E:1\})$ is $14 * 3 = 42$. $\{A:2, E:1\}$ is not a quantitative high-utility itemset. In other examples, the itemset utility of $\{B:1, D:1\}$ is $iu(\{B:1, D:1\}) = 31$. The TID list of $\{B:1, D:1\}$ is (T01, T04, and T06). Therefore, $tu(\{B:1, D:1\}) = 31 * 3 = 93$. Because the total utility of $\{B:1, D:1\}$ is larger than the minimum utility of 60, $\{B:1, D:1\}$ is a quantitative high-utility itemset.

33.4.3 Algorithms of QHIM

Table 33.5 outlines the QHIM algorithm and the global header table, and the QU-tree construction is shown in steps 1–7. We discover that QHTWU and the quantitative high-utility itemset are shown from steps 8 to 18.

The first database scan is steps 1–3. We construct the global header table transaction by transaction. We may calculate the transaction utility of each transaction (step 2), then store the transaction utility value and quantitative item to the global header table to build the global header and calculate the TWU value of each quantitative item (step 3). After the first database scan, we may calculate the total database utility by transaction utility and the minimum database utility (step 4). In step 5, we remove the quantitative items from the global header table. The second database is steps 6 and 7. For each transaction, we may build the global QU-tree according to the order of the global header table. In this paper, we use the alphabet order to construct the header table.

After constructing the global header table and the QU-tree, we may discover the QHTWU itemsets and output quantitative high-utility itemsets (steps 8–17). If the global QU-tree is a single path, we can easily discover QHTWU itemsets and quantitative high-utility itemsets. Therefore, if the global QU-tree is more than one path, we may project and construct the local header table and the QU-tree (steps 10–13) from the global header table bottom to up (step 9). When we construct the local header table and the QU-tree, we can directly calculate the actual utility value of the quantitative itemsets (step 15). If the utility value of the quantitative itemset is no less than the minimum database utility, we output the quantitative high-utility itemsets (step 6). After we output all the quantitative high-utility itemsets from the local header table, we recursively construct the local header table and the QU-tree until the local header table size is 0 or the local QU-tree is a single path (18). When we finish the QHIM algorithm, we will discover all of the quantitative high-utility itemsets.

Table 33.5 Algorithm QHIM

QHIM algorithm
Input: Database DB; minimum percentage threshold δ
Output: All quantitative high-utility itemsets
//Phase 1: Global header table and QU-tree construction
//First database scan
//Input: DB, δ
//Output: Global header table
1: For each transaction T_q in DB
2: Calculate transaction utility $u(T_q)$
3: Store each quantitative item i in global header table
4: Calculate minimum database utility $\rho = (\delta * \text{total database utility})$
5: Remove quantitative item that $\text{tu}(i) < \rho$
//Second database scan
6: For each transaction T_q in DB
7: build global QU-tree
//Phase 2: Discover QHTWU and quantitative high-utility itemsets
//Input: Global or local header table, global or local QU-tree, ρ
//Output: Quantitative high-utility itemsets
8: If there is more than one path in QU-tree
9: For each quantitative item i in header table
10: generate projection header table and QU-tree
11: AND TID list in projection header table by TID list of i
12: Remove quantitative item j that $\text{TWU} < \rho$ from projection header table
13: Construct local QU-tree
14: For each quantitative item j in local header table
15: Calculate utility value of quantitative itemset X
16: if $\text{tu}(X) \geq \rho$
17: output X
18: Recursive process from step line 8 until the size of header table is 0 or the QU-tree is a single path

For steps 1–7 (phase 1) in Table 33.5, we may ensure that we minimize the search space of the QU-tree (Lemma 3) and keep the complete database information in the header table. By Lemma 4, phase 2 (steps 8–18 in Table 33.5) of the QHIM algorithm ensures that we may discover the complete set of the quantitative high-utility itemsets.

33.5 Performance Evaluation

We have evaluated the performance of the proposed algorithm. We maintain the naïve method to compare with the QHIM algorithm. The naïve method was designed by the a priori-like method. We find quantitative high-utility itemsets level by level. In the first phase, we discover all of the 1-length QHTWU and quantitative high-utility itemsets. Then, we generate the 2-length candidate itemsets and scan the database to

determine the 2-length QHTWU and quantitative high-utility itemsets in the second phase. The naïve algorithm will stop until the number of candidate itemsets is 0.

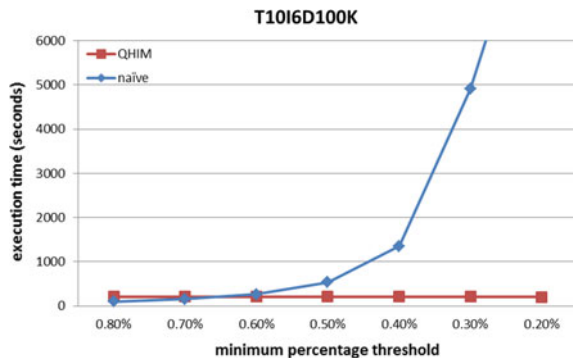
All of the experiments were executed on a PC with an Intel Core i5-2400 CPU and 2 GB of main memory running Windows 7. The QHIM algorithm and naïve method were implemented in C++. Both synthetic and real datasets were used in the experiments. The synthetic datasets were generated using the IBM Quest Data generator [20]. The quantities of the items in a transaction were randomly assigned a number between 1 and 5, and the utility values of each item in the utility table were randomly generated from 0.01 to 10, equal to those generated in the previous high-utility mining algorithms.

In this paper, we evaluate the QHIM algorithm on the following datasets: T10I6D100K with 100 items, and T10I6D1000K and T20I6D1000K with 1000 items, where T10 indicates that the average length in the transactions is 10, I6 represents that the average length of the maximum potential large itemsets is 6, and D100K (D1000K) indicates that the total number of transactions is 100K (1000K). The real datasets, which were collected from a grocery chain store, were received from U-MineBench version 2.0 [21]. The dataset includes 1,112,949 transactions with quantities, 46,086 items, and a utility table. The average length of the transactions is 7.2.

33.5.1 Execution Time Evaluation

We show the QHIM and naïve method execution times for T10I6D100K dataset in Fig. 33.4. We varied the minimum percentage threshold from 0.8 to 0.2%. The execution time of QHIM is stable close to 200 s for any minimum percentage thresholds. The naïve method is faster than QHIM when the minimum percentage threshold is larger than 0.7%. However, the number of quantitative high-utility itemsets is too small. When the minimum percentage threshold is 0.7%, we just found 685 quantitative high-utility itemsets. The naïve method will increase a significant amount

Fig. 33.4 Execution time of T10I6D100K



of time to find the quantitative high-utility itemsets when the minimum percentage threshold decreases. Therefore, when the minimum percentage threshold is lower than 0.6%, the QHIM is faster than the naïve method. The number of the quantitative high-utility itemsets is 1053 when the minimum percentage threshold is 0.6%. When the minimum percentage threshold is 0.3%, we discovered 6869 quantitative high-utility itemsets, and the naïve method cost 4909 s. Even when the minimum percentage threshold is 0.2%, the execution time is more than one day. Therefore, we cannot show the naïve execution time when the minimum percentage threshold is 0.2%.

Next, we evaluate the execution time for the T10I6D1000K dataset. Figure 33.5 shows the result of the execution time. We varied the minimum percentage threshold from 0.9 to 0.3%. The naïve method was faster than QHIM when the minimum percentage threshold was 0.9–0.6%. However, just less than 100 quantitative high-utility itemsets have been found. If the minimum percentage threshold is less than 0.6%, the naïve method must cost more than one day to mine the result. Therefore, we do not show the execution results in Fig. 33.5. Because the QHIM has a stable execution time, the QHIM will be faster than the naïve method when the minimum percentage threshold is less than 0.6%. Figure 33.6 is the execution time of the QHIM algorithm for a real chain store dataset. Because the naïve method cannot finish in

Fig. 33.5 Execution time of T10I6D1000K

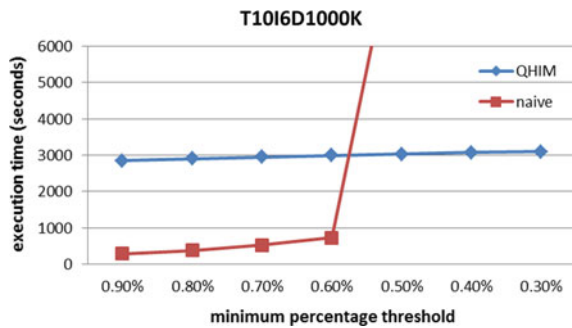
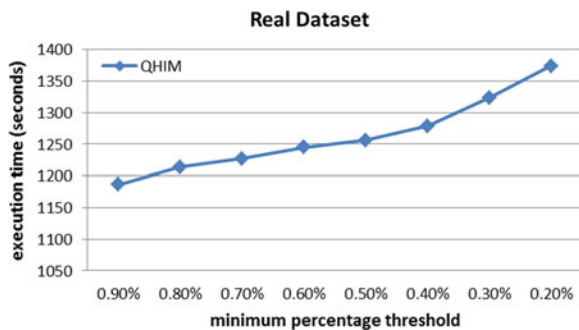


Fig. 33.6 Execution time of real dataset



any minimum percentage thresholds, we just show the QHIM results. The QHIM can finish the process in a rational time.

33.5.2 Memory Evaluation

In quantitative mining, every item may have several different quantities. Each different quantity item may seem to be an independent item. If a dataset contains 1000 items, in quantitative mining, that dataset will become several times larger. Therefore, the memory usage is a very important challenge. Figure 33.7 shows the memory usage in the T10I6D100K dataset. We evaluate the memory usage from the minimum percentage threshold of 0.9 to 0.3%. The naïve method needs a large space to store the candidate itemset information, especially when the minimum percentage threshold is small. The naïve method needs several times the space to store candidate itemsets. When the minimum percentage threshold is 0.9, the naïve method generated approximately 0.6 million candidate itemsets. More than 10 million candidate itemsets were generated when the minimum percentage threshold is 0.3. Because we do not need a candidate generation step, we just store the QU-tree in the memory. Therefore, the QHIM algorithm just needs stable memory space to build the QU-tree. Figure 33.8 is the memory usage of the QHIM algorithm for a real chain store

Fig. 33.7 Memory usage of T10I6D100K

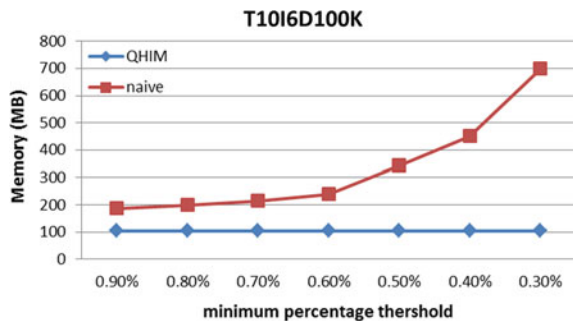
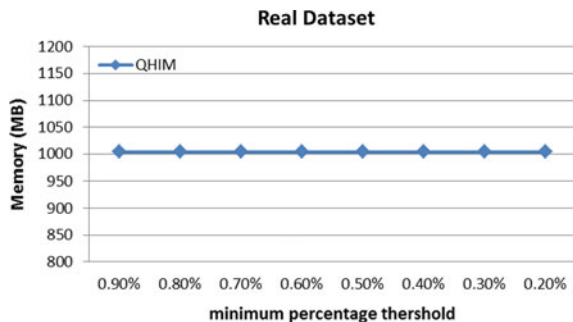


Fig. 33.8 Memory usage of real dataset



dataset. Because the naïve method cannot finish, Fig. 33.8 just shows the results of the QHIM algorithm. The QHIM algorithm still uses stable memory space in any minimum percentage threshold.

33.5.3 Performance Discussion

The naïve algorithm is a level-wise-based method to find quantitative high-utility itemsets. The mining time of the level-wise-based algorithms depends on the number of database scans and the number of the candidate itemsets. Therefore, in a large minimum percentage threshold, the naïve method may have better performance than the QHIM algorithm. However, the naïve method may find few quantitative high-utility itemsets, or even not find any itemsets. The result cannot be the reference for the business strategies. In most situations, the performance of the QHIM algorithm is better than the performance of the naïve algorithm.

Not only does the mining time of the QHIM algorithm outperform the naïve algorithm but the search space usage also outperforms the naïve algorithm. The naïve method must excuse the candidate generating step, and the candidate generating step will generate a huge number of the candidate itemsets. Especially in the quantitative mining environment, there are large numbers of quantitative items. Therefore, the naïve algorithm can just process with a small database, or the naïve algorithm may not finish the process. The QHIM algorithm just needs the search space to store the QU-tree and the header table. Therefore, the QHIM algorithm can finish the mining.

33.6 Conclusion

In this paper, we proposed the QHIM algorithm to discover all of the high-utility itemsets of generalized quantity. The novel definition of generalized quantity balances the consideration of both quantity and frequency in the mining of high-utility itemsets. The QHIM algorithm uses the QU-tree, based on the pattern-growth method to discover high-utility itemsets. Because no previous methods are designed to find itemsets of generalized quantity, the QHIM algorithm is compared with a level-wise method against both synthetic and real datasets in the experimental results. The experimental results show that QHIM outperforms the naïve method. Especially when the minimum percentage threshold is small, the naïve method must cost huge amounts of time to find the results. The QHIM algorithm can finish the process in an allowable time. Moreover, the naïve method needs a huge memory space to store information for all of the candidate itemsets. Therefore, the proposed QHIM algorithm is an efficient algorithm for discovering high-utility itemsets of generalized quantity.

Acknowledgements The authors appreciate the valuable comments from the reviewers. This research is supported partly by the Ministry of Science and Technology, R.O.C. under grant MOST-107-2221-E-035-072.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile, pp. 487–499 (1994)
2. Grahne, G., Zhu, J.: Efficiently using prefix-trees in mining frequent itemsets. In: Proceedings of the 2003 IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03) (2003)
3. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generating: a frequent-pattern tree approach. *Data Min. Knowl. Discov.* **8**(1), 53–87 (2004)
4. Ayubi, S., Muyebe, M.K., Baraani, A., Keane, J.: An algorithm to mine general association rules from tabular data. *Inf. Sci.* **179**(20), 3520–3529 (2009)
5. Liu, Y., Liao, W., Choudhary, A.: A fast high utility itemsets mining algorithm. In: Proceedings of the First International Workshop on Utility-Based Data Mining, Chicago, Illinois, Aug 2005, pp. 90–99
6. Liu, Y., Liao, W., Choudhary, A.: A two-phase algorithm for fast discovery of high utility of itemsets. In: Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Hanoi, Vietnam, pp. 689–695 (2005)
7. Yao, H., Hamilton, H.J.: Mining itemset utilities from transaction databases. *Data Knowl. Eng.* **59**(3), 603–626 (2006)
8. Erwin, R., Gopalan, P., Achuthan, N.R.: A bottom-up projection based algorithm for mining high utility itemsets. In: Proceedings of the 2nd International Workshop on Integration Artificial Intelligence and Data Mining, pp. 3–11 (2007)
9. Erwin, R., Gopalan, P., Achuthan, N.R.: Efficient mining of high utility itemsets from large datasets. In: Advances in Knowledge Discovery and Data Mining, pp. 554–561 (2008)
10. Li, Y.-C., Yeh, J.-S., Chang, C.-C.: Isolated items discarding strategy for discovering high utility itemsets. *Data Knowl. Eng.* **64**(1), 198–217 (2008)
11. Bac, L., Huy, N., Tung, A.-C., Bay, V.: A novel algorithm for mining high utility itemsets. In: Proceedings of the First Asian Conference on Intelligent Information and Database Systems, Quang Binh, Vietnam, Apr 2009, pp. 13–17
12. Li, H.-F., Huang, H.-Y., Lee, S.-Y.: Fast and memory efficient mining of high-utility itemsets from data streams: with and without negative item profits. *Knowl. Inf. Syst.* **28**(3), 495–522 (2011)
13. Lin, M.Y., Tu, T.F., Hsueh, S.C.: High utility pattern mining using the maximal itemset property and lexicographic tree structures. *Inf. Sci.* **215**, 1–14 (2012)
14. Tseng, V.S., Shie, B.-E., Wu, C.-W., Yu, P.S.: Efficient algorithms for mining high utility itemsets from transactional databases. *IEEE Trans. Knowl. Data Eng.* **25**(8), 1772–1786 (2013)
15. Tseng, V.S., Wu, C.-W., Fournier-Viger, P., Yu, P.S.: Efficient algorithms for mining top-K high utility itemsets. *IEEE Trans. Knowl. Data Eng.* **28**(1), 54–67 (2016)
16. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, pp. 1–12 (1996)
17. Tsai, P.S.M., Chen, C.-M.: Mining quantitative association rules in a large database of sales transactions. *J. Inf. Sci. Eng.* **17**(4), 667–681 (2001)
18. Li, H., Wu, C.-W., Tseng, V.S.: Efficient vertical mining of high utility quantitative itemsets. In: Proceedings of the 2014 IEEE International Conference on Granular Computing, pp. 155–160 (2014)

19. Yen, S.-J., Lee, Y.-S.: Mining high utility quantitative association rules. In: Proceedings of the 9th International Conference on Data Warehousing and Knowledge Discovery, pp. 283–292 (2007)
20. IBM Synthetic Data Generation Code: <http://www.almaden.ibm.com/software/quest/Resources/index.shtml>
21. Pisharath, J., Liu, Y., Parhi, J., Liao, W.-K., Choudhary, A., Memik, G.: U-MineBench Version 2.0 Source Code and Datasets. <http://cucis.ece.northwestern.edu/projects/DMS/MineBench.html>. Accessed Jan 2016

Chapter 34

Large-Scale Instance Selection Using a Heterogeneous Value Difference Matrix



Chatchai Kasemtaweekchok, Nitiporn Sukkerd, and Chatchavin Hathorn

Abstract Data classification of a large-scale dataset is a common problem nowadays because the classifier model takes an overwhelming amount of time to completely learn all the data. The instance selection algorithm is a well-known technique that addresses this issue by reducing the size of the training set. Instance selection methods decrease the difficulty of data classification and improve the quality of the training data. This paper proposed a novel instance selection method using a heterogeneous value difference matrix (HVDM) distance function. The proposed method selected a set of median HVDM values in each partition as a reduced training set. We compared the proposed method with the condensed nearest neighbor (CNN) and instance-based learning (IB3) methods. Five large-scale datasets from the UCI data repository were tested with three classifier models (decision tree, neural net, and support vector machine). The accuracy and kappa of the proposed method were better than those of the other two methods, and the proposed method had a moderate reduction rate. However, the accuracy and kappa of the proposed method were nearly equal to those of the original training set.

34.1 Introduction

The extent of Internet commerce provides many extremely large datasets from which information can be extracted using data mining techniques. In addition, many operations in everyday life, such as mobile transactions by a large number of clients, can lead to the generation of large amounts of data in a database system. Even if these data are useful for data classification, the large amount of data makes the classifier model inappropriate for these problems. Because of the complexity of the learning process, it can take excessive time to create a classifier model. Moreover, the large size of the dataset results in the learning process requiring a large amount of memory space.

C. Kasemtaweekchok (✉) · N. Sukkerd · C. Hathorn
Faculty of Science at Sriracha, Kasetsart University, Sriracha Campus, Sri Racha, Thailand
e-mail: Chatchai.kase@ku.th

Data reduction is recognized as an effective method to improve the learning process for large training datasets. The instance selection algorithm is one of the most common data reduction techniques to address this difficulty. Instance selection is the main technique that decreases the complexity of data classification and improves the quality of the training data because it removes missing, redundant, and noise data from a training set. Accordingly, the classifier model focuses on collecting only the instances that have affected the classification score [1]. The selected instances enable the classifier model to predict unseen with nearly equal accuracy to the original training set.

In this paper, a new approach is presented to select a subset of the training data to keep as representative of the training set using a heterogeneous value difference matrix (HVDM) called HVDM-IS. The training set is separated into disjoint partitions based on the number of instances in the dataset. In each partition, a class representative is selected which minimizes the sum of HVDM values to other instances. After the selection process is complete, the subset of training data presents the whole quality of the data in the original training set. The proposed method was evaluated on five large-scale datasets, and the performance was compared with two other algorithms. The performance of the full training set was used as the baseline value. The reduction rate shows the reduction capacity of all concerned algorithms. Accuracy and kappa were measured from three classified models: decision tree, neural net, and support vector machine.

The rest of this paper is as follows. Section 34.2 reviews briefly previous instance selection methods. The proposed method is described in Sect. 34.3. The experimental materials and methods are explained in Sect. 34.4. The results of this experiment are shown in Sect. 34.5. Lastly, Sect. 34.6 presents the conclusion.

34.2 Instance Selection

Instance selection methods can help classifier models by reducing the size of a training set. The reduction scheme of the previous instance selection methods can be categorized into three schemes: condensation, edition, and hybrid [2].

Condensation selection focuses on collecting instances near the decision boundaries. Condensed nearest neighbor (CNN) focuses on removing the instances that are correctly classified by their nearest neighbors [3]. The group of remaining instances is collected as a consistent subset. The condensation methods collect a small amount of training data, but it is sensitive to noise.

Edition selection removes the border instances and outlier instances. Edited nearest neighbor (ENN) removes instances that are misclassified in the original training set [4]. ENN removes instances that are noisy or disagree with neighbors with high accuracy. However, the reduction rate of the edition scheme is low.

Hybrid selection includes the strengths of the above two methods with internal and noise removal processes. Hybrid selection removes noise better than condensation selection and selects a subset of training data that is smaller than for edition selection.

Instance-based learning (IB3) [5] collects the subset of instances which provide a good classification record. IB3 removes the instances which have a poor classification record later. The iterative case filtering (ICF) algorithm removes the instances which have the number of nearest neighbors in the same class greater than the number of nearest enemies (the nearest neighbors of a different class) [6]. The distances to the nearest neighbors and enemies are also used in the concept of a local set. The local set of an instance x is the set of nearest instances in the same class where the distance to x is shorter than the distance between x and its nearest enemy. The idea of a local set is used as the selecting criterion in some methods [7, 8]. Moreover, the hit miss networks (HMN) method selects instances using the graph that has a directed edge from each instance to the nearest neighbors of each different class [9]. The hit degree of a node is the number of edges directed to the same class node. The miss degree is the number of edges directed to a different class node. The HMN method removes instances if the miss degree value is greater or equal to the hit degree value. The classification accuracy of the hybrid selection is comparatively higher than those of the condensation and edition selections.

In recent years, a novel instance selection algorithm has been used that selects representative instances in each partition based on the nearest enemy information near the decision boundary [10]. Furthermore, a density-based algorithm for instance selection analyzes the density of instances in each class and keeps only the densest instances of a given neighborhood within each class [11]. Some researchers presented a new method to select a representative instance of each of the densest spatial partitions [12]. In addition, a novel instance selection method uses metric learning for transforming the input space which addresses the decision boundaries between classes. The inter-class and intra-class separation criteria are used to select the instances near to the decision boundaries [13]. The above-mentioned methods achieved high classification accuracy and reduction rates when they were tested with datasets having less than 20,000 instances. Because of the small size of the tested datasets, these instance selections did not allow any scalability analysis.

34.3 Proposed Method

This section presents the idea of the heterogeneous value difference matrix instance selection (HVDM-IS) algorithm. Figure 34.1 shows the process flow of HVDM-IS, which has two main processes: the suitable partition size calculation (SPSC) process and the median of HVDM value selection (HVDMS) process. The first process calculates a suitable partition size and forwards it onto the next process. The second process uses the number of partitions for splitting the training data. Lastly, it selects the median HVDM value as the class representative in each partition.

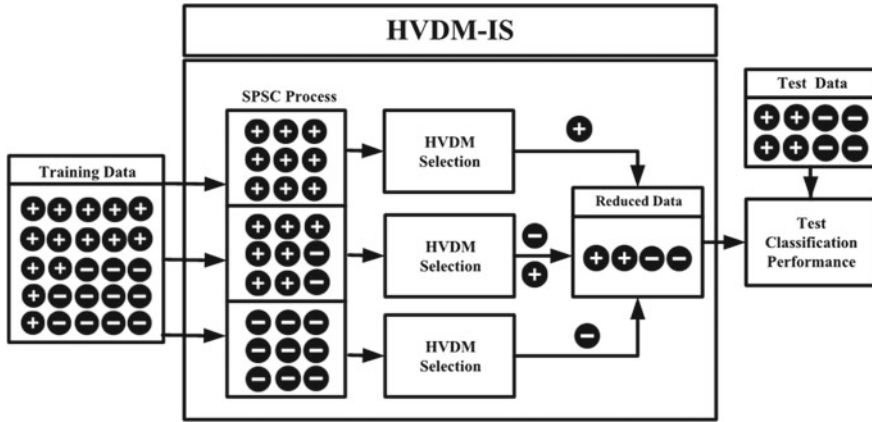


Fig. 34.1 Process flow of the HVDM-IS algorithm

34.3.1 Suitable Partition Size Calculation

The first process uses the Taro Yamane formula to calculate the number of selected instances n . The formula of Taro Yamane is shown in (34.1). Consequently, the HVDM-IS calculates the number of instances in each partition and divides the training data into n disjoint partitions. Because of its simplicity, the SPSC process is able to calculate the suitable partition size rapidly [14].

$$n = \frac{N}{1 + N(e^2)} \tag{34.1}$$

where n is the number of selected instances, N is the population (number of the overall training data) size, and e is the level of precision.

34.3.2 HVDM Selection

This section describes two related subjects: the heterogeneous value difference matrix (HVDM) and the HVDM selection process. First, the definition and formula of the HVDM distance value are explained. Lastly, the concept and pseudocode of the HVDM selection process are described.

HVDM was introduced by Wilson and Martinez [15] to compute the distance between two input vectors, and it was designed to overcome the weakness of the Euclidean distance function and the value difference matrix (VDM) distance function. The formula based on Euclidean distance works well when the attributes are linear (continuous or discrete). However, the Euclidean distance function is not suitable for nominal attributes because the values in nominal attributes are not necessarily

in any linear order. The values of some nominal attributes normally were converted from category name to a numeric value such as low (0), medium (1), and high (2) so they are not suitable for computing the numerical difference between the two values.

The VDM distance function focuses on providing an appropriate distance function for nominal attributes [16]. The conditional probability of each attribute value given a class is used to evaluate the distance between two input vectors. Each value of a nominal attribute usually appears many times among training instances. The probability calculation uses the number of occurrences of a value in each nominal attribute so the VDM is suitable for nominal attributes. However, the VDM is inappropriate for continuous attributes because of their large value range. Accordingly, the values of a continuous attribute can all potentially be unique.

The HVDM function returns the distance between two input instances x and y . It is defined as follows [15]:

$$\text{HVDM}(x, y) = \sqrt{\sum_{a=1}^m d_a^2(x_a, y_a)} \quad (34.2)$$

where m is the number of attributes. The function $d_a(x_a, y_a)$ returns the distance between x and y for attribute a and is defined as follows:

$$d_a^2(x_a, y_a) = \begin{cases} 1, & \text{if } x \text{ or } y \text{ is unknown;} \\ \text{normalized_vdm}_a(x_a, y_a), & \text{if } a \text{ is nominal;} \\ \text{normalized_diff}_a(x_a, y_a), & \text{if } a \text{ is linear} \end{cases} \quad (34.3)$$

The $d_a(x_a, y_a)$ function combines two distance functions for nominal and linear distance calculation. The HVDM function uses *normalized_diff* when the attribute is linear (discrete or continuous value). The function *normalized_diff* is shown in (34.4):

$$\text{normalized_diff}_a(x_a, y_a) = \frac{|x_a - y_a|}{4\sigma_a} \quad (34.4)$$

where σ_a is the standard deviation of the numeric values of attribute a .

The function *normalized_vdm* is used for nominal distance calculation. The function *normalized_vdm* uses the square of the difference that is similar to the Euclidean distance function. The function *normalized_vdm* is shown in (34.5):

$$\text{normalized_diff}_a(x_a, y_a) = \sqrt{\sum_{i=1}^C \left| \frac{N_{a,x,i}}{N_{a,x}} - \frac{N_{a,y,i}}{N_{a,y}} \right|^2} \quad (34.5)$$

where C is the number of classes, $N_{a,x}$ is the number of instances in the training set that have value x for attribute a , $N_{a,x,i}$ is the number of instances in the training set that have value x for attribute a , output class i , $N_{a,y}$ is the number of instances in the

training set that have value y for attribute a , and $N_{a,y,i}$ is the number of instances in the training set that have value y for attribute a and output class i .

After the training data have been divided into disjoint partitions, the HVDMS process focuses only on the selection of class representative in each partition. The method for finding the class representative is the HVDM distance value computation from a candidate to other data points in the same partition. A selected instance is a class representative that minimizes the sum of HVDM values to other instances in each partition. The pseudocode of HVDMS is shown as Algorithm 1.

There are three input parameters to the HVDMS process. First, TS is the list of instances in the current partition. Second, NS is the number of subsets in the current partition. Finally, R is the number of iterations, starting from 1 to R . In each partition, the selection process splits the training data into NS subsets. The *CreateInitialCandidates* function randomly selects a candidate in each subset. The *CalculateSumHVDM* function calculates the sum of the distance from the candidate to other points. After the set of selected candidates CS has been created, the *SelectRoundWinner* function selects a candidate that has the smallest sum of HVDM distance as RW for each round-winner. In the next iteration, the *GetNearestCandidatestoRW* function creates CS as a new set of candidates which is the nearest neighbors to the round-winner. The *CalculateSumHVDM* and *SelectRoundWinner* functions are called for finding RW , the new round-winner. The HVDMS process continuously repeats the above steps until the number of specified iterations is achieved. The result of the HVDMS process is S_{best} which is the instance minimizing the sum of HVDM distances to other instances.

Algorithm 1 Pseudocode of HVDM selection.

```

1: function HVDM SELECTION( $TS, NS, R$ )
2:    $CS \leftarrow$  CREATEINITIALCANDIDATES( $TS, NS$ )
3:    $Dist_{cs} \leftarrow$  CALCULATESUMHVDM( $CS, TS$ )
4:    $RW, Dist_{rw} \leftarrow$  SELECTROUNDWINNER( $CS, Dist_{cs}$ )
5:    $S_{best} \leftarrow RW$ 
6:    $Dist_{best} \leftarrow Dist_{rw}$ 
7:   for  $N = 1$  to  $R$  do
8:      $CS \leftarrow$  GETNEARESTCANDIDATESTORW( $RW, TS, NS$ )
9:      $Dist_{cs} \leftarrow$  CALCULATESUMHVDM( $CS, TS$ )
10:     $RW, Dist_{rw} \leftarrow$  SELECTROUNDWINNER( $CS, Dist_{cs}$ )
11:    if  $Dist_{rw} \leq Dist_{best}$  then
12:       $S_{best} \leftarrow RW$ 
13:       $Dist_{best} \leftarrow Dist_{rw}$ 
14:    end if
15:  end for
16:  return  $S_{best}$ 
17: end function

```

34.4 Experimental Materials and Methods

In this section, we describe the details of our experimental evaluation. Section 34.4.1 shows the list of benchmarking methods and datasets used in the experiment. The definitions of various performance measures and classification algorithms used in the evaluation of the HVDM-IS method are explained in Sect. 34.4.2.

34.4.1 Benchmarking Methods and Datasets

We evaluated the performance of HVDM-IS with three compared algorithms. The parameter settings of the compared methods are shown in Table 34.1.

The five datasets from the UCI data repository are shown in Table 34.2 [17]. In each test, 80% of the original dataset was used for training data, and the rest was used for test data. The compared algorithms were run over the training data to create a reduced training set.

Table 34.1 Parameter settings of compare methods

Compared methods	Parameters
Condensed nearest neighbor (CNN)	Mixed Euclidean distance
Instance base 3 (IB3)	Parameter $k = 3$, Upper interval = 0.9, Lower interval = 0.7, Mixed Euclidean distance
Hit miss networks (HMN-EI)	Epsilon = 0.1, Euclidean distance
HVDM-IS	Number of iterations = 5, Number of subsets = 10

Table 34.2 Description of dataset

Dataset	Number of samples	Number of attributes (Real/Integer/Nominal)	Number of classes
Fars	100,968	29 (5/0/24)	8
Census	299,284	41 (1/12/28)	3
KDD cup	494,020	41 (26/0/15)	23
Covertime	581,012	54 (0/54/0)	7
Poker	1,025,010	10 (0/10/0)	10

34.4.2 Performance Measures

The performance of HVDM-IS was compared with the other methods using three performance measures: reduction rate, accuracy, and Cohen's kappa. First, the reduction rate represents the reduction of storage capacity obtained by the method. A higher reduction rate shows that the algorithm can reduce the training data better. Second, accuracy shows the percentage of correctly classified instances. Accuracy is the most common performance indicator of classification methods. Finally, Cohen's kappa evaluates the ratio of correctly classified instances that can be attributed to a classifier itself by recompensing for random correctly classified instances. This measure is in the range from -1 to 1 . A larger kappa value shows that the rating of compliance between the predicted label and the actual label is higher. The formula of Cohen's kappa is shown in (34.6) [18]:

$$\text{Kappa} = \frac{N \sum_{i=1}^r x_{ii} - \sum_{i=1}^r (x_{i+} \times x_{+i})}{N^2 - \sum_{i=1}^r (x_{i+} \times x_{+i})} \quad (34.6)$$

where r is the number of rows or columns in the confusion matrix, x_{ii} is entry (i, i) of the confusion matrix, x_{i+} and x_{+i} are the marginal totals of row i and column i , respectively, and N is the total number of examples in the confusion matrix.

We used three classification algorithms to evaluate the performance of the HVDM-IS and the compared methods: decision tree (J48), neural net (NET), and support vector machine (SVM). First, J48 is a Java program of the C4.5 algorithm in the Weka data mining tool [19]. C4.5 is a widely used decision tree algorithm proposed by Quinlan [20]. C4.5 builds decision trees from a training set using information entropy. Second, the NET algorithm is a multilayer perceptron (MLP) that is a feed-forward artificial neural network model trained by back propagation algorithms [21, 22]. Lastly, we used the LIBSVM algorithm that is a popular open-source library for support vector machines [23, 24]. LIBSVM supports multiclass learning and probability estimation by Platt calibration for transforming the outputs of a classification model into confidence values of predicted classes [25].

34.5 Results

In this section, we show the details of our results. Section 34.5.1 shows the reduction rate of HVDM-IS and the benchmarking methods used in the study. The classification accuracy and Cohen's kappa are reported for the evaluation of the HVDM-IS method in Sect. 34.5.2. Finally, a trade-off between the reduction rate and classification performance is discussed in Sect. 34.5.3.

34.5.1 Reduction Rate

The reduction rate shows the reduction capability of the instance selection method. It is an important performance indicator in the data reduction. From Table 34.3, the IB3 method had the highest reduction rate (89.15%) because it includes two sample removal processes: one for noise and the other for poor classification score instances for 1NN performance. Otherwise, the reduction rate of the HVDM-IS was nearly 80% of the training data size. The reduction rate of HVDM-IS was higher than that of the CNN method. The selection process of HVDM-IS focuses on the quality of the training data, so the size of the selected instances from HVDM-IS is quite large.

34.5.2 Classification Accuracy

Table 34.4 shows the classification accuracy by J48 and the standard error of HVDM-IS, the full training model, CNN, and the IB3 method. The CNN method had the best average accuracy rate. However, the size of the reduced training sets of CNN was large compared with that of the other methods. The HVDM-IS method provided the best accuracy rate on two datasets. The accuracy rate of the HMN-EI method was approximately 2.5%, higher than that of the HVDM-IS method. However, the average accuracy rate of HVDM-IS was approximately 5% lower than that of the full training model.

In Table 34.5, the average classification accuracy by NET of HVDM-IS was higher than those of the CNN and IB3 methods by about 5%. The HVDM-IS and CNN methods provided the best accuracy on two datasets. The average accuracy of the HMN-EI method was nearly equal to that of the HVDM-IS method. However, the average accuracy of HVDM-IS was nearly equal to that of the full training model.

The classification accuracy achieved by LIBSVM of HVDM-IS was much higher than by the CNN, IB3, and HMN-EI methods, as shown in Table 34.6. The classifi-

Table 34.3 Reduction rate of HVDM-IS, CNN, IB3, and HMN-EI methods

Dataset	HVDM-IS	CNN	IB3	HMN-EI
Fars (%)	77.13	54.65	74.99	47.27
Census (%)	87.07	47.73	97.66	79.03
KDD cup (%)	97.80	99.68	99.85	90.18
Coverttype (%)	67.65	88.84	93.38	46.20
Poker (%)	67.49	48.52	78.89	13.25
Average (%)	79.43	67.88	89.15	55.19
Standard error	0.0584	0.1007	0.0496	0.1339

Table 34.4 Classification accuracy by J48 on five datasets

Dataset	Full training	HVDM-IS	CNN	IB3	HMN-EI
Fars (%)	79.88	78.87	79.39	78.79	79.73
Census (%)	95.65	94.64	94.56	86.64	94.20
KDD cup (%)	99.96	99.11	99.29	99.50	99.95
Coverttype (%)	96.80	90.80	79.77	77.68	91.82
Poker (%)	76.31	61.40	73.24	62.69	72.02
Average (%)	89.72	84.96	85.25	81.06	87.54
Standard error	0.0483	0.0678	0.0496	0.0602	0.0509

Table 34.5 Classification accuracy by NET on five datasets

Dataset	Full training	HVDM-IS	CNN	IB3	HMN-EI
Fars (%)	46.80	47.11	42.03	34.54	46.89
Census (%)	95.12	93.20	94.78	86.83	92.56
KDD cup (%)	99.89	99.55	99.60	99.63	99.87
Coverttype (%)	81.67	79.34	62.84	74.30	79.00
Poker (%)	52.92	54.03	54.54	51.47	53.24
Average (%)	75.28	74.65	70.76	69.35	74.31
Standard error	0.1084	0.1042	0.1131	0.1179	0.1050

Table 34.6 Classification accuracy by LIBSVM on five datasets

Dataset	Full training	HVDM-IS	CNN	IB3	HMN-EI
Fars (%)	45.05	45.12	41.37	40.27	44.90
Census (%)	94.84	94.24	94.53	89.14	93.04
KDD cup (%)	99.36	99.09	44.91	37.60	99.42
Coverttype (%)	82.69	86.10	86.96	86.49	84.02
Poker (%)	50.12	50.12	42.10	50.30	50.12
Average (%)	74.41	74.93	61.97	60.76	73.99
Standard error	0.1131	0.1137	0.1182	0.1125	0.1477

cation accuracy for LIBSVM of HVDM-IS was higher than that of the full training model. The average accuracy of the HVDM-IS method was nearly equal to that of the HMN-EI method as shown in Fig. 34.2.

34.5.3 Cohen's Kappa

Table 34.7 shows the values for Cohen's kappa and the standard error for J48, the full training model, the HVDM-IS model, and the other methods. The results indicated

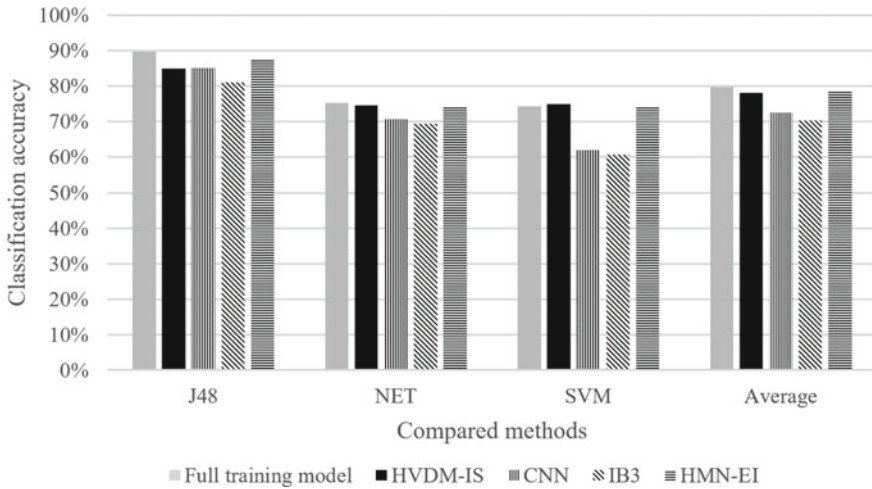


Fig. 34.2 Average classification accuracy of HVDM-IS and the compared methods

Table 34.7 Cohen’s kappa by J48 on five datasets

Dataset	Full training	HVDM-IS	CNN	IB3	HMN-EI
Fars	0.7260	0.7120	0.7190	0.7110	0.7230
Census	0.4960	0.4080	0.4660	0.3350	0.5250
KDD cup	0.9990	0.9850	0.9880	0.9920	0.9990
Coverttype	0.9350	0.8520	0.6770	0.6470	0.8680
Poker	0.5720	0.3050	0.5190	0.3690	0.4760
Average	0.7456	0.6524	0.6738	0.6108	0.7182
Standard error	0.0982	0.1293	0.0916	0.1207	0.0993

that CNN method had the best kappa value for three out of five datasets. The Cohen’s kappa of HVDM-IS was higher than that of the IB3 method. However, the Cohen’s kappa of HVDM-IS was lower than that of full training model.

In Table 34.8, NET of the HVDM-IS method yielded the best Cohen’s kappa that was higher than those of the CNN and IB3 methods by about 5%. Furthermore, the Cohen’s kappa of the HVDM-IS method was higher than that of full training model too. For LIBSVM, Cohen’s kappa of the HVDM-IS method was much higher than those of the CNN and IB3 methods as shown in Table 34.9. The average Cohen’s kappa of the HVDM-IS method was nearly equal to that of the HMN-EI method. Finally, Fig. 34.3 shows the average Cohen’s kappa values of the HVDM-IS method for the three classifier models were higher than those of the CNN and IB3 methods.

Table 34.8 Cohen’s kappa by NET on five datasets

Dataset	Full training	HVDM-IS	CNN	IB3	HMN-EI
Fars	0.2270	0.2500	0.2120	0.1530	0.2500
Census	0.3930	0.3880	0.4290	0.3000	0.4420
KDD cup	0.9980	0.9920	0.9930	0.9940	0.9980
Covertype	0.6290	0.6650	0.4320	0.5830	0.6540
Poker	0.1120	0.1180	0.1280	0.1030	0.0790
Average	0.4718	0.4826	0.4288	0.4266	0.4846
Standard error	0.1577	0.1563	0.1509	0.1646	0.1603

Table 34.9 Cohen’s kappa by LIBSVM on five datasets

Dataset	Full training	HVDM-IS	CNN	IB3	HMN-EI
Fars	0.2170	0.2330	0.2080	0.1980	0.2140
Census	0.2360	0.0780	0.1390	0.2000	0.3520
KDD cup	0.9890	0.9850	0.2990	0.2190	0.9900
Covertype	0.7160	0.7740	0.7870	0.7810	0.7390
Poker	0.0000	0.0000	0.0000	0.0000	0.0000
Average	0.4316	0.4140	0.2866	0.2796	0.4350
Standard error	0.1820	0.1966	0.1343	0.1316	0.1571

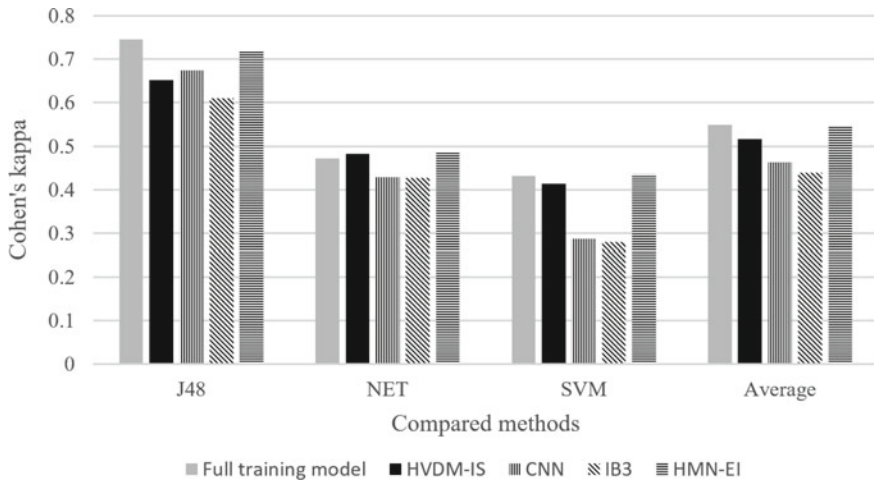


Fig. 34.3 Average Cohen’s kappa of HVDM-IS and the compared methods

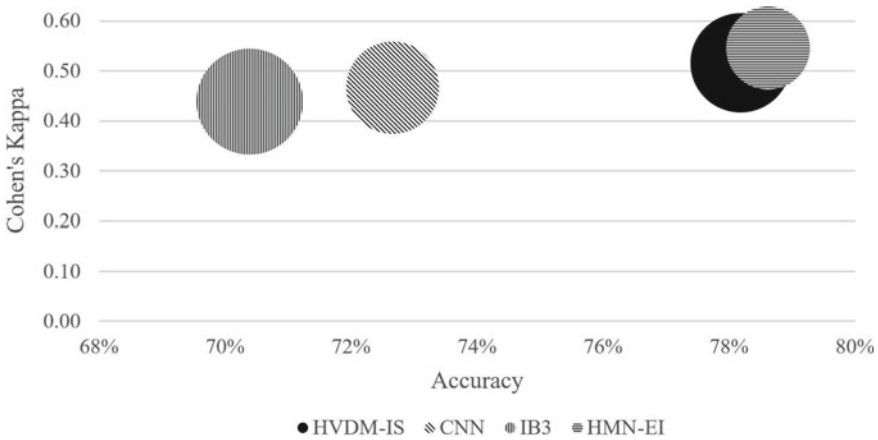


Fig. 34.4 Graphical comparison of accuracy (x-axis), Cohen’s kappa (y-axis), and reduction rate (size of bubble) of HVDM-IS and the compared methods

34.5.4 Trade-Off Between Accuracy and Reduction Rate

The average classification accuracy of the HVDM-IS method was higher than the CNN (72.66%) and IB3 (70.39%) methods. Figure 34.4 presents a bubble chart providing a graphical comparison of the accuracy, Cohen’s kappa, and reduction rates of the HVDM-IS and the compared methods. The HVDM-IS method had the second-highest reduction rate (79.43%). Even though the IB3 method had the highest reduction rate (89.15%), the classification accuracy of the IB3 method was the lowest. Moreover, the accuracy rate of HVDM-IS (78.18%) was nearly equal to that of the HMN-EI method (78.62%) but the HMN-EI method had a low reduction rate (55.19%).

The HMN-EI method had the highest average Cohen’s kappa value (0.55) while the HVDM-IS method had an average Cohen’s kappa value (0.52) which was higher than those of the CNN (0.47) and IB3 (0.44) methods. Although the HVDM-IS method had the second-highest reduction rate, it could handle the trade-off between accuracy, Cohen’s kappa, and the reduction rate reasonably well, as shown in Fig. 34.4.

34.6 Conclusion

A new method was proposed to select a subset of the training data to keep as a representative training set using the heterogeneous value difference matrix (HVDM) method, called HVDM-IS. The classifier model creation takes an overwhelming amount of time when a large-scale training dataset is being processed. The data

were split into independent partitions. In each partition, a class representative was selected which minimized the sum of HVDM values to other instances. The selected instances were used for the training classification model. The accuracy and kappa were measured from three classified models: decision tree (J48), neural net (NET), and support vector machine (LIBSVM).

The results of this experiment showed that the HVDM-IS method provided classification accuracy and Cohen's kappa values that were higher than those of the CNN and IB3 methods for the NET and LIBSVM classifier models. For the J48 classifier model, the accuracy and Cohen's kappa value of the HVDM-IS method were a little lower than those of the CNN method, while the number of selected instances from the CNN method was larger. Furthermore, the classification accuracy and Cohen's kappa of the HVDM-IS method were nearly equal to those of the full training model and the HMN-EI method. However, the HMN-EI method had a quite low reduction rate.

The results of this experiment showed that the HVDM distance can help the instance selection method to calculate the appropriate distance value because the distance function combination in HVDM is applicable to the characteristics of nominal and linear distance calculation.

In future work, the HVDM-IS method could be applied in a parallel and distributed processing system, which should improve the processing speed of the method. Furthermore, we would aim to reduce the large real-world datasets with HVDM-IS and show its scalability for the big data problem.

Acknowledgements This work was financially supported by the Faculty of Science at Sriracha, Kasetsart University, Sriracha campus, Thailand.

References

1. García, S., Luengo, J., Herrera, F.: Data Preprocessing in Data Mining. Springer, Heidelberg (2015). <https://doi.org/10.1007/978-3-319-10247-4>
2. García, S., Derrac, J., Cano, J.R., Herrera, F.: Prototype selection for nearest neighbor classification: taxonomy and empirical study. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 417–435 (2012). <https://doi.org/10.1109/TPAMI.2011.142>
3. Hart, P.E.: The condensed nearest neighbor rule. *IEEE Trans. Inf. Theory.* **14**, 515–516 (1968). <https://doi.org/10.1109/TIT.1968.1054155>
4. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Syst. Man. Cybern.* **SMC-2**, 408–421 (1972). <https://doi.org/10.1109/TSMC.1972.4309137>
5. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. *Mach. Learn.* **6**, 37–66 (1991). <https://doi.org/10.1007/BF00153759>
6. Brighton, H., Mellish, C.: Advances in instance selection for instance-based. *Data Min. Knowl. Discov.* **6**, 153–172 (2002). <https://doi.org/10.1023/A:1014043630878>
7. González, A.A., Pastor, D.F.J., Rodríguez, J.J., Osorio, G.C.: Local sets for multi-label instance selection. *Appl. Soft Comput.* **68**, 651–666 (2018). <https://doi.org/10.1016/j.asoc.2018.04.016>
8. Leyva, E., González, A., Pérez, R.: Three new instance selection methods based on local sets: a comparative study with several approaches from a bi-objective perspective. *Pattern Recognit.* **48**(4), 1523–1537 (2015). <https://doi.org/10.1016/j.patcog.2014.10.001>

9. Marchiori, E.: Hit miss networks with applications to instance selection. *J. Mach. Learn. Res.* **9**, 97–1017 (2008). <https://doi.org/10.5555/1390681.1390715>
10. Yu, G., Tian, J., Li, M.: Nearest neighbor-based instance selection for classification. In: 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 75–80. IEEE, New Jersey (2016). <https://doi.org/10.1109/FSKD.2016.7603154>
11. Carbonera, J.L., Abel, M.: A novel density-based approach for instance selection. In: IEEE 28th International Conference on Tools with Artificial Intelligence, pp. 549–556. IEEE, New Jersey (2016). <https://doi.org/10.1109/ICTAI.2016.0090>
12. Carbonera, J.L., Abel, M.: Efficient instance selection based on spatial abstraction. In: IEEE 30th International Conference on Tools with Artificial Intelligence, pp. 286–292. IEEE, New Jersey (2018). <https://doi.org/10.1109/ICTAI.2018.00053>
13. Max, Z.E., Marcacini, M.R., Matsubara, T.E.: Improving instance selection via metric learning. In: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, New Jersey (2018). <https://doi.org/10.1109/IJCNN.2018.8489322>
14. Yamane, T.: *Statistics: An Introductory Analysis*, 2nd edn. Harper and Row, New York, USA (1967)
15. Wilson, D.R., Martinez, T.R.: Improved heterogeneous distance functions. *J. Artif. Intell. Res.* **6**, 1–34 (1997)
16. Stanfill, C., Waltz, D.: Toward memory-based reasoning. *Commun. ACM.* **29**, 1213–1228 (1986). <https://doi.org/10.1145/7902.7906>. ACM, New York
17. UCI machine learning repository, <http://archive.ics.uci.edu/ml>
18. Triguero, I., Derrac, J., Garcia, S., Herrera, F.: A taxonomy and experimental study on prototype generation for nearest neighbor classification. *IEEE Trans. Syst. Man. Cybern.* **42**, 86–100 (2012). <https://doi.org/10.1109/TSMCC.2010.2103939>
19. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. *SIGKDD. Explor.* **11**(1), 10–18 (2009). <https://doi.org/10.1145/1656274.1656278>
20. Quinlan, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, MA (1993)
21. Rosenblatt, F.: *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington, DC (1961)
22. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, pp. 318–362. MIT Press, Cambridge (1986)
23. Corinna, C., Vladimir, V.: Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995). <https://doi.org/10.1023/A:1022627411411>
24. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**(3), 1–27 (2011). <https://doi.org/10.1145/1961189.1961199>
25. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Mar. Classif.* **10**(3), 61–74 (1999). <https://doi.org/10.1007/s10994-007-5018-6>

Part V
Intelligent Systems

Chapter 35

Bio-inspired Algorithms for Modeling and Control of Underwater Flexible Single-Link Manipulator



I. Z. Mat Darus and Ali A. M. Al-Khafaji

Abstract This research focuses on bio-inspired modeling and control system of an underwater flexible manipulator system (UFM). The dynamic behavior of the UFM was first modeled using system identification (SI) methods utilizing bio-inspired algorithms. The input–output data used for identification were acquired directly from a laboratory-sized UFM experimental rig developed earlier by the previous researcher. The models were developed using cuckoo search algorithm (CSA) and flower pollination algorithm (FPA) using parametric ARX model structured. For the controllers of the UFM, proportional-integral-derivative (PID) controllers were tuned using conventional heuristic and intelligent FPA methods. These algorithms were utilized to obtain the optimal values of controller parameters for trajectory tracking control of rigid-body motion of the UFM system. The PID controller is tuned offline based on the best identified SI model. The performance of these control schemes was analyzed via real-time PC-based control and observed in terms of trajectory tracking and error. The overall result of UFM described in this research revealed the superiority of the PID controllers tuned using bio-inspired flower pollination algorithm (FPA). It was found that the percentage of improvement achieved experimentally by the PID controller tuned by FPA indicates superiority compared to PID tuned heuristically with 45.6% improvement on overshoot and 66% improvement of MSE for negative pulse and 100% improvement on overshoot for positive pulse.

35.1 Introduction

Modern underwater applications make use of solid manipulators built from high stiffness material. This is because solid manipulators have certain advantages such as strong and heavy metal compositions that lead to stable performance. However, major drawbacks of solid manipulators lie in their need for high energy consumption and limitations in their speed of operation. Moreover, it is desirable in manufacturing of engineering systems to keep the weight as low as possible. There is a growing

I. Z. Mat Darus (✉) · A. A. M. Al-Khafaji
Universiti Teknologi Malaysia, Johor Bahru, 81310 Johor, Malaysia
e-mail: intan@utm.my

© Springer Nature Singapore Pte Ltd. 2021
S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation,
Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_35

483

trend in many applications to reduce the weight of mechanical structures to the barest minimum, especially in aircraft engineering and spacecraft. The utilization of weaker structures and/or lighter materials significantly reduces production cost. However, light materials can also lead to more flexibility which may limit the structure performance [1]. In recent years, the researchers focused on utilization of lightweight manipulators to build power-efficient robot manipulators in order to improve the industrial productivity [2]. Therefore, the use of lightweight manipulators is emerging in the field of space and various general-purpose industrial applications. The manipulators with rigid links are currently utilized in underwater applications because the heavy and strong metal leads to stable performance [3, 4]. However, major disadvantages of rigid link manipulators lie in their need for high energy consumption and limitation in speed of operation.

There is a motivation to use underwater manipulators with flexible link owing to the advantages offered by manipulator systems with flexible link compared with manipulator systems with rigid link such as fast response, lightweight, low inertia, cheap construction, less powerful actuators, longer reach, higher payload carrying capacity, and safer operation [5, 6]. Although several studies have been conducted on modeling and control schemes of land and space manipulator systems with flexible links, only very few literature discussed on underwater flexible manipulators [7–9].

Therefore, there is an open area of research to study and develop dynamic modeling and control strategies for underwater flexible manipulators. Thus, the main aim of this paper is to present a suitable computational comprehensive model governing the underwater flexible manipulator system (UFM) using system identification (SI) technique. The manipulator addressed in this study is restricted to move in horizontal plane. Also, there is a big challenge in controlling the UFM owing to the additional effects caused by underwater environment, namely disturbances by ocean currents, time variance and high nonlinearity [9]. Consequently, it is necessary to develop appropriate control approaches for this type of systems.

35.2 Underwater Manipulator Test Rig

An underwater manipulator test rig used in this research, as shown in Figs. 35.1 and 35.2, was designed and constructed in Faculty of Mechanical Engineering, Universiti Teknologi Malaysia, Johor in order to perform the underwater flexible manipulator experiments [9].

Several experimental testings have been conducted to check the similitude of the underwater and land manipulators [9]. The angular displacement was measured using encoder while the end-point vibration was measured by an accelerometer where the signals were transmitted to a data acquisition card for analog-to-digital conversion of the signal. Experimental work was conducted in order to acquire data to identify the model of the hub-angle of the UFM and demonstrate the practicality of the proposed control schemes. The dynamic model of the hub-angle of the UFM was developed using SI methods utilizing input–output data acquired experimentally. The modeling

Fig. 35.1. UFM experimental rig [9]

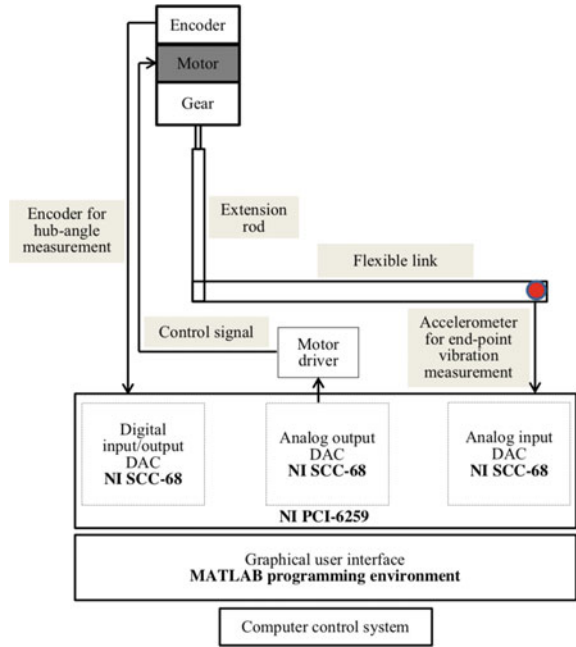
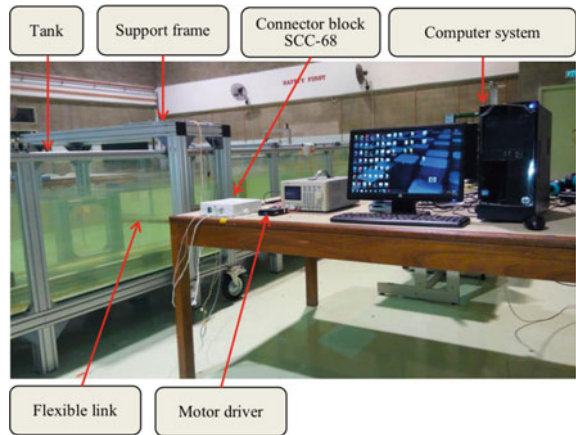


Fig. 35.2 Experimental setup [9]



was conducted within MATLAB programming environment using CSA and FPA. After validating the developed models of the hub-angle of the UFM, the best model among the models thus developed has been utilized for the development of control approaches for hub-angle of the UFM.

Later, PID control strategies were developed using heuristic and bio-inspired algorithm tuning methods. The control algorithms were computing the amount of motor voltage required for trajectory tracking of the UFM. The bio-inspired PID control

scheme tuned offline by using FPA and conventional tuning method using heuristic tuning. The performance of the intelligent PID control schemes was compared with a conventional PID control scheme. Both PID control schemes are implemented for trajectory tracking control of UFM via the developed experimental rig. The objective of the comparative study is to observe the differences in their performance simultaneously and to exploit the benefits of using the proposed strategies.

35.3 System Identification

System identification (SI) is one of the most fundamental requirements for several scientific and engineering applications. The aim of SI is to build exact or approximate model of a dynamic system based on measured data without knowledge of the actual system physics. After a system model is obtained, it can be utilized to predict the physical system behavior under different operating conditions or to control it [4].

Parametric modeling of UFM utilizing metaheuristic algorithm through cuckoo search algorithm (CSA) and flower pollination algorithm (FPA) is presented in this paper. The aim of the work is to represent the UFM behavior utilizing the applied voltage as input and hub-angle as output based on CSA and FPA. Model validations were also investigated using mean-squared error, one-step ahead prediction, and model residual analysis. The performances of the CSA and FPA were compared based on the validation mean-squared error, modeling mean-squared error, correlation tests, and stability. The aim of the identification process in this research is to allow for the design and implementation of controllers based on the identified model for trajectory tracking of UFM.

35.3.1 Cuckoo Search

CSA is a search algorithm developed by Yang and Deb [10]. The algorithm was inspired by the breeding behavior of cuckoos. Cuckoo birds lay their eggs in other birds' nests and rely on those birds for hosting the eggs. If some of the host birds discover that an egg is not their own, it might throw out the alien egg or move to a new location elsewhere. A cuckoo might emulate the shape, color, and size of the host eggs to protect their egg from being discovered. To increase the hatching probability of cuckoo birds own eggs, some of them might throw out other native eggs from the host nest. On the other hand, a hatched cuckoo chick will also throw other eggs out of the nest to improve its feeding share [10]. It can be noted from the literature that the efficiency of CSA has been demonstrated by solving several engineering problems. In control system problems area, the utilization of metaheuristic optimization approaches is widely and clearly appreciated. To date, CSA has been utilized in various control system problems successfully. These literatures show that CSA

has been utilized efficiently in tuning PID-based controller parameters in different control scheme.

The user-defined parameters required for optimization process using CSA are: number of generations, G , population size, N , problem dimension, D , flight step size, α , discovery rate of alien eggs, Pa , and boundary constraints, LHC. In this study, D refers to the number of unknown parameters in ARX model structure. The flight step size, $\alpha = 0.01$, and the fraction of eggs to be discarded, $Pa = 0.25$, were used, as suggested by Yang and Deb [10]. It is worth noting that other CSA's optimization parameters such as G , N , and LHC are difficult to choose in order to obtain promising results since there is no prior knowledge regarding the rules in selecting these parameters. Thus, these parameters were obtained by a trial-and-error method. The population needs to be initialized before the optimization starts and their fitness values need to be calculated. Therefore, (N) nests of dimension (D) are initialized randomly within the specified lower and upper range. Each nest in the initial population then updates the parameters of ARX model structure and its fitness value is calculated based on the error between the predicted and actual outputs. Among the nests in the initial population, the one with the minimum cost was considered as the best nest. Figure 35.3 shows the diagrammatic representation of initial population generation. The optimization process will run iteratively until the end of generations [9].

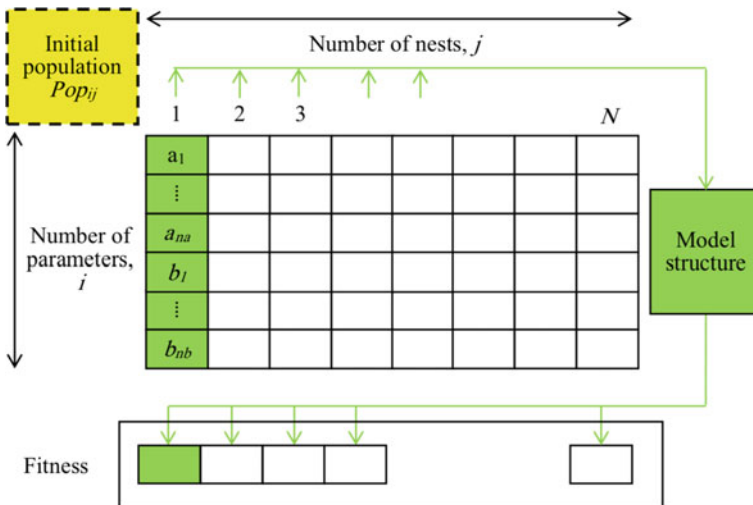


Fig. 35.3 Diagrammatic representation of the initial population generation [9]

35.3.2 Flower Pollination Algorithm

Flower pollination algorithm (FPA) is a biology-based algorithm inspired by the flow pollination process of flowering plants. Pollination can take two major forms: biotic and abiotic. Biotic (cross-pollination), means pollination can occur from pollen of a flower of a different plant, while abiotic (self-pollination) is the pollination of one flower from pollen of the same flower or different flowers of the same plant. Biotic is considered as global pollination process with pollen carrying pollinators performing Levy flights. Abiotic is considered as local pollination. Local pollination and global pollination are controlled by a switch probability $p \in [0, 1]$. That means, the probability will specify each of solutions to search the local area or global area [10].

The user-defined parameters required for optimization process using FPA are: number of generations, G , population size, N , problem dimension, D , flight step size, α , probability switch, P , and boundary constraints, LHC. In this study, D refers to the number of the unknown parameters in ARX model. The flight step size, $\alpha = 0.01$, and the probability switch, $P = 0.8$, were used as suggested by Yang and Deb [10]. It is worth noting that it is difficult to choose other FPA's optimization parameters such as G , N , and LHC in order to obtain promising results since there is no prior knowledge regarding the rules in selecting these parameters. Thus, these parameters were obtained by a trial-and-error method.

The population needs to be initialized before the optimization starts and their fitness values need to be calculated. Therefore, (N) pollens of dimension (D) are initialized randomly in the given upper and lower bounds [10]. Each pollen in the initial population then updated the parameters of ARX model structure and its fitness value based on the error between predicted and actual outputs. The best pollen in the initial population is corresponding to the pollen with minimum cost. Figure 35.4 shows the diagrammatic representation of the initial population generation. The optimization process will run iteratively until the end of generations. The pollen with lower fitness value is selected as the best pollen for the next generation [9]. Detail description of CSA and FPA identification process is described by Al-Khafaji in [9].

35.4 Tuning of PID Controller Using FPA

To achieve an appropriate control action, the overall effect of PID controller gains, K_P , K_I , and K_D should be in such a way optimum. Hence, the aim of this investigation is to tune the PID controller parameters offline utilizing two metaheuristic algorithm namely, CSA and FPA. Simulation study was conducted in order to highlight the performance of new metaheuristic optimization techniques to optimally tune the PID controller in the proposed control scheme. MATLAB/Simulink was used to tune the PID controller gains K_P , K_I , and K_D in offline mode. The schematic diagram of the closed-loop system utilizing PID controller with the identified hub-angle model is

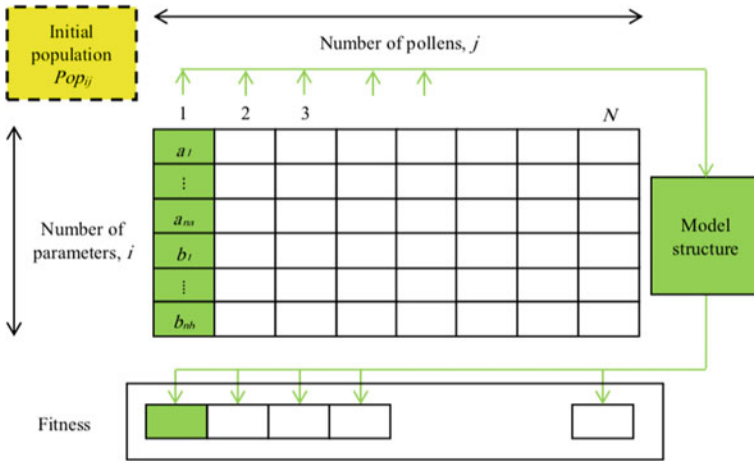


Fig. 35.4 Diagrammatic representation of generation the initial population [9]

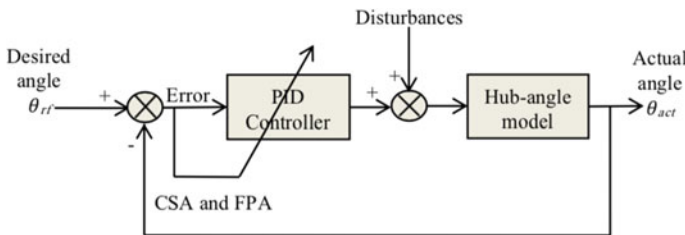


Fig. 35.5. Schematic diagram of PID controller tuning by bio-inspired methods [9]

shown in Fig. 35.5. Bang-bang signal was used as input reference with magnitude of ± 0.3 rad for 21 s. The hub-angle model was excited with a disturbance signal at amplitude of 0.2608 m/s. The performance of both tuning methods has been observed in terms of overshoots, M_{pi} and M_{pd} , and steady-state errors, E_{ssi} and E_{ssd} . Then, the tuned parameters achieved from the simulation were tested experimentally using the UFM test rig [10]. Detail description of bio-inspired PID-FPA controller is described by Al-Khafaji in [10].

35.5 Results and Discussion

35.5.1 System Identification Using CSA and FPA

In this study, CSA was used to determine the ARX model structure parameters which represent the hub-angle of the UFM. Same control parameters were optimized for

different arbitrarily selected model order to choose the best model order. The initial CSA optimization parameters used in modeling processes is as follows: number of generation, $G = 600$, population size, $N = 20$, problem dimension, D , depending on the model order, flight step size, $\alpha = 0.001$, fraction of eggs to be discarded, $P_a = 0.25$, and boundary constraints, $LHC = (3, 2)$. The ARX parameters are optimized via CSA by minimizing the fitness function, mean square error (MSE). The numerical results of the work carried out to select the best model order are summarized as shown in Table 35.1. The performances of CSA and FPA were compared based on the validation mean-squared error, modeling mean-squared error, stability, and correlation tests.

It can be noted from Table 35.1 that the best result was accomplished with third model order. Using third model order, the optimum values of $b_1, b_2, b_3, a_1, a_2,$ and a_3 are $0.0005882, 0.0003957, 1.537 \times 10^{-5}, -2.206, 1.484,$ and $0.2782,$ respectively. The optimal ARX parameters have been searched randomly utilizing CSA optimization technique in such a way that a global minimum of MSE is reached. The results of the hub-angle and error in both modeling and validation phases using CSA-MSE optimization method are shown in Figs. 35.6 and 35.7, respectively. It can be noted from Figs. 35.6 and 35.7 that a satisfactory response was attained and the output of CSA-based model could follow the actual output very well with modeling MSE of 1.24101×10^{-4} and validation MSE of 1.82360×10^{-4} .

Table 35.1 Performance of CSA and FPA with different numbers of model order

Model order	TUNING USING CSA				TUNING USING PFA			
	MSE of validation phase	MSE of modeling phase	Cor. tests	Stability	MSE of validation phase	MSE of modeling phase	Cor. tests	Stability
2	1.827×10^{-4}	1.242×10^{-4}	Biased	Stable	1.827×10^{-4}	1.247×10^{-4}	Biased	Stable
3	1.824×10^{-4}	1.241×10^{-4}	Biased	Stable	1.841×10^{-4}	1.253×10^{-4}	Biased	Stable
4	2.162×10^{-4}	1.461×10^{-4}	Biased	Stable	1.851×10^{-4}	1.293×10^{-4}	Biased	Stable
5	1.842×10^{-4}	1.255×10^{-4}	Biased	Stable	2.017×10^{-4}	1.368×10^{-4}	Biased	Stable
6	5.691×10^{-4}	4.257×10^{-4}	Biased	Stable	3.701×10^{-4}	1.307×10^{-4}	Biased	Stable
7	4.745×10^{-4}	3.915×10^{-4}	Biased	Stable	4.646×10^{-4}	4.981×10^{-4}	Biased	Stable
8	0.0035	0.0033	Biased	Unstable	5.180×10^{-4}	4.080×10^{-4}	Biased	Unstable
9	0.0053	0.0052	Biased	Unstable	5.336×10^{-4}	5.579×10^{-4}	Biased	Unstable
10	0.0077	0.0071	Biased	Unstable	0.0033	0.0036	Biased	Unstable

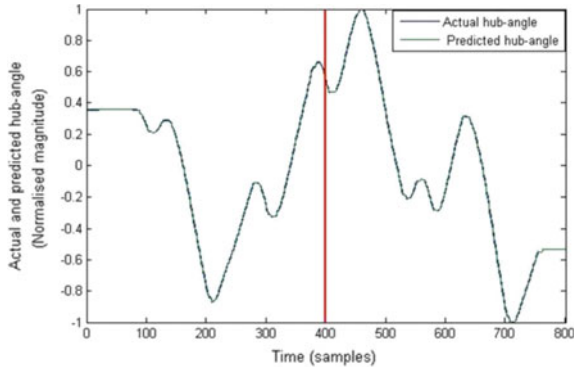


Fig. 35.6 Actual and estimated hub-angles using CSA-algorithm

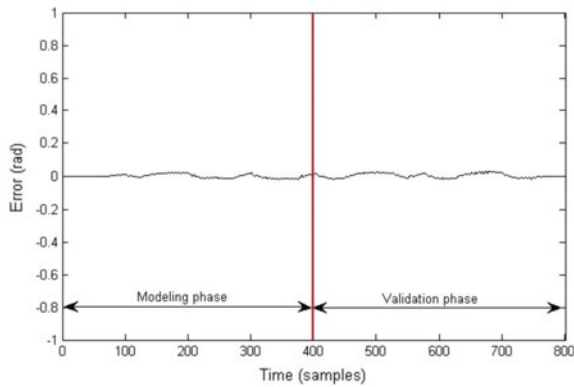


Fig. 35.7 Error between actual and estimated hub-angles using CSA-algorithm

Same control parameters of FPA were optimized for different arbitrarily selected model order to choose the best model order. Optimization of ARX model structure parameters utilizing FPA was achieved by initially initialized FPA control parameters as closed as possible to CSA control parameters. The population size and generation number are set to 20 and 600, respectively, while the flight step size, α and the probability switch, P are set 0.001 and 0.8, respectively. It is worth to know that the selection of flight step size and the probability switch is based on the guidelines from the previous literatures [10]. The numerical results of the work carried out to select the best model order are shown in Table 35.1.

It can be noted from Table 35.1 that the best result by FPA was accomplished with second order. After ARX optimization procedure was finished, the optimum values of ARX parameters are found as $b_1 = 0.0003741$, $b_2 = 0.0007799$, $a_1 = -1.925$, and $a_2 = 0.9251$. The optimal ARX parameters have been searched randomly utilizing

FPA in such a way that a global minimum of MSE is reached. The results of hub-angle and error in both modeling and validation phases using FPA-MSE optimization method are shown in Figs. 35.8 and 35.9, respectively, where the division between the modeling data and validation data is indicated as a vertical red line located at point 400. It can be noted from Figs. 35.8 and 35.9 that the predicted response using FPA method could follow the actual output very well with MSE of 1.24723×10^{-4} during training and MSE of 1.82754×10^{-4} for validation data. The pole-zero diagram was used to confirm the model was stable; all the poles of the transfer function were inside the unit circle. The correlation functions were carried out for 20 samples. It was found that the model is biased because the results are not within the 95% confidence bands. Parametric modeling of the hub-angle of the UFM has been performed utilizing two optimization algorithms, namely FPA and CSA. The overall comparative performance of optimization methods in terms of validation MSE, modeling MSE, stability, and correlation tests are summarized in Table 35.2.

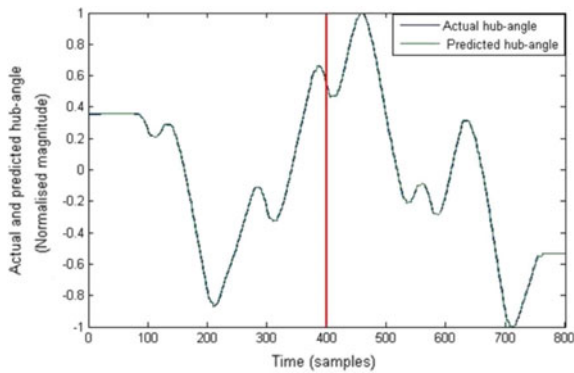


Fig. 35.8 Actual and estimated hub-angles using FPA algorithm

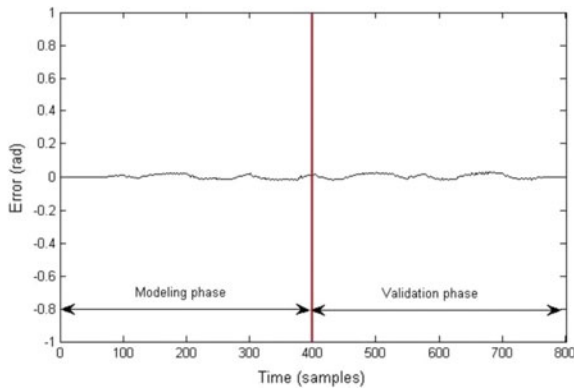


Fig. 35.9 Error between actual and estimated hub-angles using FPA algorithm

Table 35.2 Summary of the best performance achieved in parametric modeling

Methods	Validation MSE	Modeling MSE	Correlation test	Stability
CSA	1.82360×10^{-4}	1.24101×10^{-4}	Biased	Stable
FPA	1.82754×10^{-4}	1.24723×10^{-4}	Biased	Stable

Table 35.3 Numerical results of hub-angle responses using CSA, FPA, and heuristic methods

		PID-FPA	PID-heuristic	% of improvement
Positive pulse	M_{pi}	0	0.6	100
	E_{ssi}	0	0	0
Negative pulse	M_{pd}	12.06	22.16	45.6
	E_{ssd}	0.0002	0.0006	66
Controllers' parameters	K_P	6	3	–
	K_I	4.4613	3	–
	K_D	0.5	0.8	–

It can be noted from Table 35.3 that CSA parametric identification technique provides the best estimation of UFM model, as compared to FPA. The UFM model obtained using CSA will be utilized in subsequent studies for the development of control approaches for hub-angle of the UFM.

The results of all modeling methods were validated using MSE of unseen data, correlation tests, and stability. The performances of CSA, FPA models were assessed based on the validation MSE, modeling MSE, correlation tests, and stability. It can be seen that the CSA has achieved slightly better MSE value in both modeling and validation phases and has approximated the system response very well. The best model of the UFM thus developed is utilized for the development of control approaches for hub-angle of the UFM.

35.5.2 PID Tuning Using FPA

PID controller parameters tuning utilizing FPA were achieved by initially initializing FPA control parameters as closed as possible to CSA control parameters. The population size N and generation number G were set to 10 and 150, respectively, while the flight step size α and the probability switch P were set to 0.01 and 0.8, respectively. It is worthy to note that the selection of flight step size and the probability switch is based on the guidelines from previous literatures. Figure 35.10 shows the typical convergence of objective function for 150 generation. It is noted from Fig. 35.10 that FPA converges in about 77 generations. After controller tuning procedure was finished, the optimal values of PID parameters were found to be $K_P = 6$, $K_I = 4.4613$, and $K_D = 0.5$. Figure 35.11 shows the convergence profile of PID parameters. The

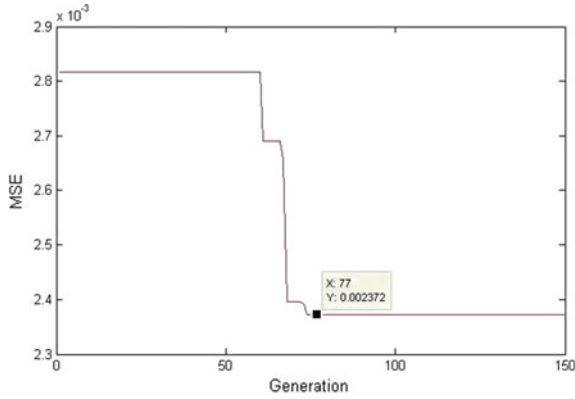


Fig. 35.10 FPA convergence

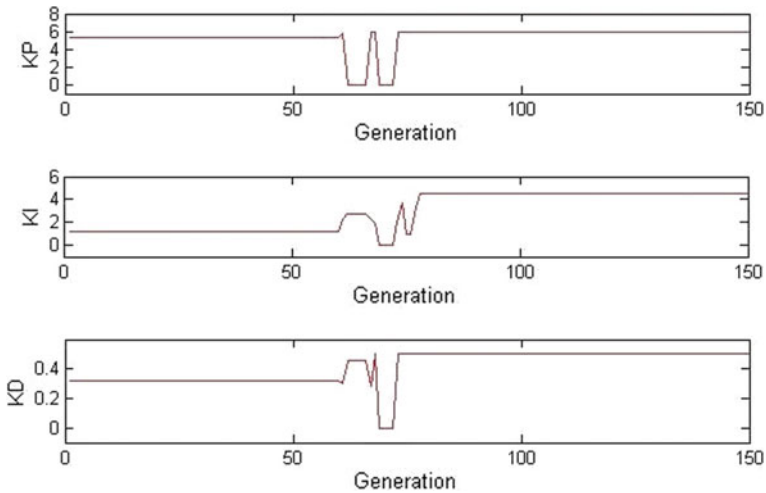
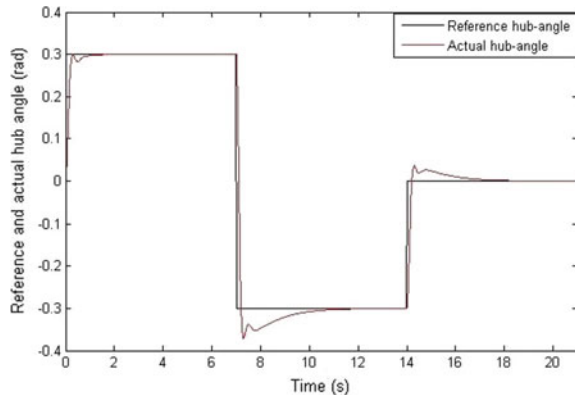


Fig. 35.11 FPA-PID parameter convergence

result of the closed-loop bang-bang response using FPA-MSE tuning method for hub-angle is shown in Fig. 35.12. It can be concluded from Fig. 35.12 that a satisfactory response was attained and the proposed controller is capable of tracking the desired hub-angle.

Hub-angle control of the UFM has been established utilizing PID control structure. The PID controller parameters were tuned offline via heuristic and bio-inspired algorithm based on the best identified model using system identification method. The best two sets of tuned controllers' parameters achieved from simulation were validated experimentally in real time using the UFM test rig where the manipulator is subjected to external disturbance. The performance of PID controller tuned by

Fig. 35.12 Simulation bang-bang response using PID controller tuned by FPA



FPA indicates superiority compared to PID tuned heuristically with 45.6% improvement on overshoot and 66% improvement of MSE for negative pulse and 100% improvement on overshoot for positive pulse.

References

1. Darus, I.Z.M., Zahidi Rahman, T.A., Mailah, M.: Experimental evaluation of active force vibration control of a flexible structure using smart material. *Int. Rev. Mech. Eng.* **5**(6), 1088–1094 (2011)
2. Jamali, A., Mat Darus, I.Z.M., Samin, P.M., Tokhi, M.O.: Intelligent modeling of double link flexible robotic manipulator using artificial neural network. *J. VibroEng.* **20**(2), 1021–1034 (2018)
3. Al-Khafaji, A.A.M., Mat Darus, I.Z.: Finite element method to dynamic modelling of an underwater flexible single-link manipulator. *J. VibroEng.* **16**(7), 3620–3636 (2014)
4. Mat Darus, I.Z., Tokhi, M.O.: Parametric and non-parametric identification of a two-dimensional flexible structure. *J. Low Freq. Noise Vibr. Active Control* **25**(2), 119–143 (2006)
5. Annisa, J., Mat Darus, I.Z., Tokhi, M.O., Mohamaddan, S.: Implementation of PID based controller tuned by evolutionary algorithm for double link flexible robotic manipulator. In: 2018 International Conference on Computational Approach in Smart Systems Design and Applications, ICASSDA 2018
6. Jamali, A., Mat Darus, I.Z., Tokhi, M.O., Abidin, A.S.Z.: Utilizing P-type ILA in tuning hybrid PID controller for double link flexible robotic manipulator. In: 2nd International Conference on Smart Sensors and Application, ICSSA 2018, pp. 141–146
7. Ng, B.C., Darus, I.Z.M., Jamaluddin, H., Kamar, H.M.: Application of adaptive neural predictive control for an automotive air conditioning system. *Appl. Therm. Eng.* **73**(1), 1244–1254 (2014)
8. Saad, M.S., Jamaluddin, H., Darus, I.Z.M.: Active vibration control of a flexible beam using system identification and controller tuning by evolutionary algorithm. *J. Vibr. Control (JVC)* **21**(10), 2027–2042 (2015)

9. Al-Khafaji, A.A.M.: Modeling and Control of Underwater Flexible Single-Link Manipulator Employing Bio-Inspired Algorithms and PID Based Control Schemes, Ph.D. Thesis, Universiti Teknologi Malaysia, Malaysia (2016)
10. Yang, X.-S., Deb, S.: Cuckoo search: Recent advances and applications. *Neural Comput. Appl.* **24**(1), 169–174 (2014)

Chapter 36

Application of Time Series Method to the Passenger Flow Prediction in the Intelligent Bus Transportation System with Big Data



Yinna Ye, Ruoxi Liu, and Feng Xue

Abstract Based on the real data collected from the bus IC card payment devices, first a time series plot on the daily passenger volume was obtained and then three kinds of time series models were proposed to do the prediction. The results show that the ARMA model with quadratic trend is the most suitable to the current data and performs the most effectively in the prediction.

36.1 Introduction

With the current development of the intelligent urban public transportation system in China, the investigation on the bus passenger flow has become a key research subject (see [1] for instance). In order to maintain the competitiveness in the transportation market and provide services with high-level quality to the passengers, the bus transportation companies need to grasp the change rules of the passenger demand sustainably [2]. However, the passenger flow in the bus system is influenced by many factors, including commuting, holiday, weather, temperature, etc. [2]. For example, the volume would experience a sudden increase during low temperature and snowy days, which would lead to the inability of bus transport capacity to meet passenger demand and brings tremendous pressure to the bus transportation management. Considering the limited bus resources, some popular routes are often in short supply, which might result in the problems of passenger flow detention and reduced-quality service. The bus companies might thus lose competitiveness in the transportation market. Therefore, it is necessary to find an effective solution to the problems

Y. Ye (✉) · R. Liu

Department of Mathematical Sciences, Xi'an Jiaotong—Liverpool University, Suzhou, China
e-mail: yinna.ye@xjtlu.edu.cn

R. Liu

e-mail: ruoxiliu15@outlook.com

F. Xue

Xiongdi Shenzhen Emperor Technology Company, Shenzhen, China
e-mail: xuefeng@xiongdi.cn

caused by such burst of passenger flow and adjust the current management policies in support to the optimal bus resource allocation, line planning and bus scheduling. The solution is of great importance to improve both the service capability and the working efficiency in the public transportation system.

The driving motivation of this work is to find a reliable method to solve the problems above. Undoubtedly, this piece of work is socially significant and important since the urban transport plan and policy could be well designed or adjusted with adapting the market demand. The implementation of this work involves a combination of big data processing, time series modeling and analysis. The primary objective of the work is to apply the time series models and data analytics to explore the passenger demand based on the real data and then to predict the daily passenger volume in a given bus line. The study will mainly focus on the following two aspects:

- Descriptive statistics on the trip characteristics of passengers, including riding date and time, and on the volume and variation characteristics of transit passenger flow at different stations in a given bus line.
- Time series parameters estimation and passenger volume prediction are based on the bus tick sale records.

In this work, SAS (version 9.4) (see [3] for instance) will be used to obtain the descriptive statistics, to do time series analysis and predictions.

The rest of the paper is organized as follows. Section 36.2 reviews the development of time series analysis and recent works on the application of the time series to the public transportation systems. Section 36.3 presents time series related concepts and methods, as well as our data analysis process. Section 36.4 summarizes and evaluates the empirical results. And finally the conclusion is discussed in Sect. 36.5 and certain open questions and some future improvements are proposed in Sect. 36.6.

36.2 Literature Review

Prior to 1920, the time series was limited to drawing lines through a mass of data. In 1927, Yule [4] first introduced the concept of ‘autoregressive’ that the variables are time related and time is not a causal factor, and pioneered the autoregressive (AR) Model of order two when studying the number of sunspots and exploring the period of the disturbed sequence. The autoregressive model he established is a special kind of stationary time series. In 1931, Walker [5] expanded and generalized the AR model to higher orders. While, Slutsky [6] was interested in the randomness of the time series, regarding them as the perturbations and then the moving average (MA) model was proposed. In 1938, Wold [7] proved that the discrete stationary process consists of implicit periodicity and linear regression. The hidden cycle is a deterministic component, while the linear regression part consists of a moving average and an autoregressive process, which are non-deterministic components of random perturbations. Any stationary time series, whose deterministic components are eliminated, can be reduced to a linear combination of random perturbations. This

well-known time series decomposition idea is the theoretical basis for the idea of the autoregressive moving average (ARMA) model. By taking non-stationary into consideration, the autoregressive integrated moving average (ARIMA) model was proposed in the landmark work [2]. The book provided a systematic approach to analyze and forecast the time series and discussed how to identify, estimate and diagnose the ARIMA model.

The application of time series models in the modern society has rapidly widespread, as the application was extended to non-stationary process (see for instance [8]). A large number of empirical results show that most time series established based on the socio-economic phenomena are non-stationary and have a trend (see for instance [9]). According to Xia [9], there are two types of time trend, one is deterministic and another one is random. Deterministic time trend is the one that can be characterized by a function of the time. The commonly used trend functions are linear functions, quadratic parabola functions, exponential functions and logarithmic functions. By contrast, the time series with stochastic trend cannot be expressed by the deterministic functions of time. In this case, multiple differences are operated to the original process and then the ARIMA model is used to fit the data.

In the literature, the existing researches suggest that the time series analysis has been properly utilized in studying different public transportation systems. For the subway systems in Shanghai, Zhu [10] constructed an ARIMA model for the daily passenger flow by comparing the change rate of daily volume with that of '7-day' average volume. For the airport terminal departure passenger traffic, Li et al. [11] took daily periodicity of the process into consideration and proposed a seasonal autoregressive integrated moving average (SARIMA) model to predict the passenger flow in Kunming Changshui International Airport. For the railway passenger flow forecast, a time series model was established in [12] with the combination of the long-term trend, the seasonal and the weather factors. To achieve an accurate real-time taxi passenger hotspot prediction, Jamil and Akbar [13] proposed an automatic ARIMA model to determine the value of the model order automatically. The algorithm designed by them overcame the common obstacle, subjectivity and complexity. All these applications make use of the knowledge of passenger flow and provide instructive insight to the management of the public transportation system, which has a referential significance for our investigation.

36.3 Methodology

36.3.1 *Stationary Time Series Models*

The time series analysis aims to reveal the underlying dynamics and structures that affect observable data, thus establishing a suitable theoretical model for monitoring and predicting data. For the definition of stationary time series (or simply called 'time series'), one can refer for instance to the Definition 1.3.2 in [14]. In this book, the daily

passenger flow volumes $\{Z_t\}$ at any unit of time t will be regarded as a discrete-time stochastic process. Roughly speaking, assuming that $\{Z_t\}$ is a stationary time series with mean 0 and Z_t depends only on its historical records Z_{t-1}, Z_{t-2}, \dots then we can use the observed historical data to estimate the dynamic properties, create optimal models and then use these models to do the prediction. In this project, we construct discretely sampled time series based on the actual daily records of passenger volume in a given bus line. The detailed description about the database can be found in Sect. 36.4.1. In the rest of this subsection, some related fundamental concepts will be introduced. One may refer to [8] for the details.

Autoregressive Model: AR (p). The autoregressive (AR) model is a very common time series. The general p -order autoregressive model, denoted as AR(p), is given by:

$$Z_t = \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \dots + \varphi_p Z_{t-p} + a_t, \quad (36.1)$$

where the parameters $\varphi_1, \varphi_2, \dots, \varphi_p$ are called **autoregressive coefficients** and they are to be estimated. The random error terms $\{a_t\}$ is the white noise, i.e., a sequence of i.i.d. random variables, $a_t \sim N(0, \sigma_a^2)$ and $\{a_t\}$ is mutually independent with $Z_{t-1}, Z_{t-2}, \dots, Z_{t-p}$.

Moving Average Model: MA (q). The general q -order moving average model, denoted as MA (q), is given by:

$$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q}, \quad (36.2)$$

where $\theta_1, \theta_2, \dots, \theta_q$ are called **moving average coefficients** and they are to be estimated.

Autoregressive Moving Average Model: ARMA (p, q). The autoregressive moving average (ARMA) combines an AR model with a MA model to produce a new process that simulates the time series. The general ARMA model, denoted as ARMA (p, q), is given by

$$\begin{aligned} Z_t = & \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \dots + \varphi_p Z_{t-p} \\ & + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q}. \end{aligned} \quad (36.3)$$

Autoregressive Integrated Moving Average Model: ARIMA(p, d, q). Notice that the AR, MA, and ARMA models are stationary time series. However, sometimes the time series are not necessarily stationary. It may have a linear trend component. For non-stationary time series, it is necessary to transform it into a stationary one through the backward shift operator. Such a non-stationary time series is called **ARIMA process**, denoted as ARIMA (p, d, q), and is given by

$$(1 - \varphi_1 B - \dots - \varphi_p B^p)(1 - B)^d Z_t = (1 - \theta_1 B - \dots - \theta_q B^q) a_t, \quad (36.4)$$

where B is the backward shift operator (lag) defined as $(1 - B)Z_t = Z_t - Z_{t-1}$ and d is the number (order) of the difference to make the process stationary.

ARMA Model with a Quadratic Function Trend. Indeed, besides considering a linear trend component in the time series, some other trend forms may also be taken into account. If the trend of a time series has a shape as a quadratic function, then it can be fitted by a quadratic function. The ARMA model with a quadratic function trend is given by

$$\begin{aligned} Z_t &= \text{quadratic function} + \text{ARMA process} \\ &= at + bt^2 + \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \dots + \varphi_p Z_{t-p} + a_t \\ &\quad - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q} \end{aligned} \quad (36.5)$$

In the rest of this section, the application of time series method to the passenger flow prediction will be introduced. This can be achieved by the descriptive and inferential studies on the current data.

36.3.2 Time Series Analysis

According to [8], the main steps of time series analysis and modeling are:

1. Stationarity and white noise test
2. Model identification (i.e., specifying the lag order)
3. Model selection and parameter estimation
4. Diagnostic checking
5. Prediction based on the optimal model.

Stationary Test. The first step of time series analysis is to verify whether the series is stationary. There are two main methods: one is the graph test, which illustrates the features shown in the time series plots and autocorrelation diagrams, while the other one is the unit root test.

Graph Test.

1. Time series plot

According to the property that mean and variance of a stationary time series are constant, the time series plot should show that the process fluctuates randomly near a constant value and the ranges of fluctuation are similar. The time series is usually not stationary if there exists a significant trend or periodicity.

2. Autocorrelation function (ACF) plot

ACF is used to describe the degree of linear correlation between different observations in time series. It is proven that the stationary time series usually have short-term correlation. The time series is stationary if the autocorrelation function declines rapidly to zero and all the values fall into the confidence interval by lag 3. In contrast, the autocorrelation of a non-stationary series declines slowly.

Unit Root Test. The unit root test is used to check whether a time series needs to be differenced. The procedure is described in [15]. Among the unit root tests, the most widely used one is the Dickey–Fuller (DF) test, which is applicable to the AR(1) model:

$$Z_t = \varphi_1 Z_{t-1} + a_t = (1 - \varphi_1 B)^{-1} a_t = \sum_{k=0}^{\infty} \varphi_1^k a_{t-k}, \tag{36.6}$$

where $|\varphi_1| < 1$. Since the root of the characteristic equation $1 - \varphi_1 B = 0$ is φ_1^{-1} , another equivalent statement of the stationary form is that the root must be outside the unit circle. So it suffices to test whether the root of the characteristic equation is outside the unit circle, with, respectively, null and alternative hypothesis:

- H_0 : $\{Z_t\}$ is non - stationary, $|\varphi_1| = 1$, a regular difference is needed
- H_1 : $\{Z_t\}$ is stationary, $|\varphi_1| < 1$, the series donot need to be deferenced

The DF test is only applicable to the AR(1) model. In order to generalize the DF test and make it widely applicable to AR(p) processes, an augmented Dickey–Fuller (ADF) test was proposed in [16] with the same hypothesis and the decision rules and includes two other new terms: drift and trend.

White Noise Test. In order to verify whether a process is worth further time series modeling and analysis, it is needed to perform the white noise test. From the definition of the white noise, for any lag k , its autocorrelation coefficient is given by $\rho_k = 0$. It should be noted that this is the ideal situation. While in practice, most of the autocorrelation coefficients $\hat{\rho}_k$ are not equal to zero due to the finiteness of the sample sequence, but they fluctuate randomly around a value of 0 with a small float. According to the methods summarized by Wei [17], instead of considering each autocorrelation individually, the first m autocorrelation coefficients as a whole are considered and an index to determine whether a sequence is white noise or whether there exists a correlation between observations is constructed. The null and alternative hypotheses for the white noise test are, respectively:

- H_0 : $\rho_1 = \rho_2 = \dots = \rho_m = 0, \forall m \geq 1$, so $\{Z_t\}$ is a white noise sequence
- H_1 : for $\forall m \geq 1, \exists k \leq m$ and $k \neq 0$ that $\rho_k \neq 0$, so $\{Z_t\}$ is not a white noise sequence

This is an approximate statistical hypothesis test that none of the autocorrelations of the series up to a given lag are significantly different from 0. If this is true for all m lags, then there is no information in the series to model and no ARIMA model is needed.

Methods of Order Specification. To determine the order (p, q) of ARMA models, SAS provides a list of the order combinations, which is mainly referred to ESACF, SCAN and MINIC methods.

The extended sample autocorrelation function (ESACF) method. Since the ACFs and PACFs of ARMA (p, q) model are all trailing, these two functions cannot be jointly used to determine the order (p, q) . Considering this situation, Tsay and Tiao [18] proposed a general iterative regression method, and used the ESACF to estimate the order of the model. The method is applicable if the time series y_t belongs to ARMA (p, q) process, then by fitting AR (p) model to it, the estimate of the autocorrelation regression coefficients $\hat{\varphi}_i, i = 1, 2, \dots, p$ will be inconsistent. Therefore, the residual error of regression must be introduced into the model as an explanatory variable, and when such process goes on until the q times the estimated model is as follows:

$$Z_t = \sum_{i=1}^p \varphi_i^{(q)} Z_{t-i} + \sum_{i=1}^q \alpha_i^{(q)} \hat{e}_{t-i}^{(q-1)} + e_t^{(q)}. \tag{36.7}$$

Now the estimator $\hat{\varphi}_i^{(q)}$ will be consistent. Based on this idea, let $m = 0, 1, 2, \dots$, $\hat{\varphi}_i^{(j)}$ is the j th iteration estimated autoregressive coefficient of the AR (m) model, then $\hat{\rho}_i^{(m)}$ is defined as the sample autocorrelation function of the following model:

$$y_t = \left(1 - \hat{\varphi}_1^{(j)} B - \hat{\varphi}_2^{(j)} B^2 - \dots - \hat{\varphi}_m^{(j)} B^m\right) z_t. \tag{36.8}$$

Regarding the ESACF, there exists the following probabilistic convergence:

$$\hat{\rho}_j^{(m)} \xrightarrow{p} \begin{cases} 0, & 0 \leq m - p \leq j - q; \\ X \neq 0, & \text{otherwise} \end{cases}. \tag{36.9}$$

Because of this property, the distribution of the ESACF for ARMA $(1,1)$ model can be displayed as in Table 36.1, which is characterized by the fact that all zeroes form a triangle with the vertex $(1,1)$. Similarly for the general ARMA (p, q) , the vertex of all zeroes is located at (p, q) , which is the rule of identifying the order of the model. In fact, SAS provides two tables, one is for the estimate of ESACF and the other one is for the significance test.

The smallest canonical correlation coefficient (SCAN) method. Tsay and Tiao [19] firstly put forward this idea, and Choi [20] gave the concrete method of solving and judging ARMA (p, q) model. Only the conclusion of this method is given here. First, the SCAN of each model with different order combination is calculated, and then the table of SCAN similar to that of ESACF is formed. The only difference is

Table 36.1 ESACF for ARMA (1, 1) model, where X is a non-zero number

MA	0	1	2	3	...
AR					
0	X	X	X	X	...
1	X	0	0	0	...
2	X	X	0	0	...
3	X	X	X	0	...
...

that the judgment is based on the rectangle with zeroes being vertices so that the corresponding vertex position is the order of the model. In our project, SAS gives two tables, one for the estimate of SCAN coefficients and the other for chi-square test results of the coefficient significance.

The minimum information criterion (MINIC) method. The minimum information criterion (MINIC) method, proposed by Hannan and Rissanen [21], can tentatively identify the order of a stationary and invertible ARMA process. The MINIC table is constructed by computing Bayesian information criterion (BIC) for various autoregressive and moving average orders. Suppose L is the value of the likelihood function evaluated at the parameter estimates of ARMA(p, q), N is the number of observations, and k is the number of estimated parameters, the BIC of ARMA(p, q) model can be calculated as:

$$BIC(p, q) = k \ln(N) - 2 \ln(L) \tag{36.10}$$

Values of $BIC(p, q)$ that cannot be computed are set to missing. For large autoregressive and moving average test orders with relatively few observations, a nearly perfect fit can result. This condition can be identified by a large $BIC(p, q)$ negative value. The MINIC table can be in the form in Table 36.2. The model with the minimum BIC value is chosen as the best fitted one.

Methods of Parameters Estimation. There are various ways to estimate the parameters, such as moment estimation, least squares estimation, maximum likelihood

Table 36.2 MINIC table

MA	0	1	2	3	...
AR					
0	BIC(0,0)	BIC(0,1)	BIC(0,2)	BIC(0,3)	...
1	BIC(1,0)	BIC(1,1)	BIC(1,2)	BIC(1,3)	...
2	BIC(2,0)	BIC(2,1)	BIC(2,2)	BIC(2,3)	...
3	BIC(3,0)	BIC(3,1)	BIC(3,2)	BIC(3,3)	...
...

estimation and so on. In this work, method of maximum likelihood estimation is adopted, which is recommended by most experts using SAS for prediction.

Maximum likelihood method. According to the maximum likelihood method of time series analysis discussed by Guidolin and Pedio [22], under the maximum likelihood criterion, it is considered that the sample comes from the population with the highest probability of occurrence of this sample. Therefore, the maximum likelihood method for the unknown parameter's estimation is to make the likelihood function $L(\varphi_1, \dots, \varphi_p, \theta_1, \dots, \theta_q)$ reach the maximum, suppose $p(z_1, z_2, \dots, z_n, \varphi_1, \dots, \varphi_p, \theta_1, \dots, \theta_q)$ is the joint density function, L can be written as:

$$L(\varphi_1, \dots, \varphi_p, \theta_1, \dots, \theta_q) = p(z_1, z_2, \dots, z_n, \varphi_1, \dots, \varphi_p, \theta_1, \dots, \theta_q) \quad (36.11)$$

The distribution function of the population must be known to use the maximum likelihood. However, in the time series analysis, the distribution of population is often unknown. In order to facilitate calculation and analysis, it is usually assumed that the sequence follows multivariate normal distribution:

$$\begin{aligned} Z_t &= \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \dots + \varphi_p Z_{t-p} + a_t - \theta_1 a_{t-1} \\ &\quad - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q}, \tilde{z} = (z_1, z_2, \dots, z_n)', \end{aligned} \quad (36.12)$$

$$\tilde{\beta} = (\varphi_1, \dots, \varphi_p, \theta_1, \dots, \theta_q)', \quad (36.13)$$

$$\sum_n = E(\tilde{z}'\tilde{z}) = \Omega\sigma_a^2. \quad (36.14)$$

The likelihood function of \tilde{z} is

$$L(\tilde{\beta}) = p\tilde{\beta} = (2\pi)^{-n/2} \left| \sum_n \right|^{-1/2} \exp \left\{ -\frac{\tilde{z}' \sum_n^{-1} \tilde{z}}{2} \right\}. \quad (36.15)$$

The log likelihood function is

$$l(\tilde{\beta}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma_a^2) - \frac{1}{2} \ln|\Omega| - \frac{1}{2\sigma_a^2} [\tilde{z}' \Omega^{-1} \tilde{z}]. \quad (36.16)$$

The system of likelihood equations can be obtained by computing the partial derivatives of the unknown parameters of the logarithmic likelihood function.

Theoretically, solving the likelihood equations yields the maximum likelihood of the unknown parameter. However, since $\tilde{z}' \Omega^{-1} \tilde{z}$ and $\ln|\Omega|$ is not an explicit expression of the parameter, the likelihood equations are actually composed of $p + q + 1$ transcendental equations, which usually requires a complex iterative algorithm to find the maximum likelihood of the unknown parameter.

The maximum likelihood method makes full use of the information provided by each observation, so its estimation accuracy is high, and it also has good statistical properties such as consistency and progressive validity.

Diagnostic test. In this test, the goodness of fit and the accuracy of the model are measured and the correlation test and the normality test on the residual series are performed. The following two kinds of criterion will be used to measure the goodness of fit for a model:

Akaike's information criterion (AIC). Akaike [23] defined AIC as

$$\text{AIC} = -2 \ln(L) + 2k, \quad (36.17)$$

where L is the value of the likelihood function evaluated at the parameter estimates, N is the number of observations and k is the number of estimated parameters. The first term of the AIC measures the goodness of fit of the ARMA model to the data, and the second term is called the penalty function of the criterion because it penalizes a candidate model by the number of parameters used. Therefore, the model with the minimum AIC value should be chosen.

Schwarz's Bayesian information criterion (SBC). Schwarz [24] defined AIC as

$$\text{SBC} = -2 \ln(L) + \ln(N)k, \quad (36.18)$$

Similarly, the model with the minimum SBC value should be chosen. The penalty for each parameter is 2 for AIC and $\ln(N)$ for SBC, so compared to AIC, SBC tends to select a lower-order model when sample size is moderate or large.

There are other two kinds of criterion to measure the accuracy of a model's predictions will be used. One can refer to [25] for the detailed description.

Mean absolute percentage error (MAPE). The MAPE is a common measure of forecast error in time series analysis. It usually expresses accuracy as a percentage and is defined by the formula:

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n |Z_t - F_t| / Z_t, \quad (36.19)$$

where Z_t is the actual value and F_t is the forecast value.

Mean square error (MSE). The MSE is measure of the differences between prediction values and the actual values. It is defined by:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Z_t - F_t)^2, \quad (36.20)$$

where Z_t is the actual value and F_t is the forecast value.

36.4 Empirical Results

36.4.1 Introduction to Database

In this work, the original data was provided by the bus companies in the city of Jiaozuo in China, including the bus IC card payment records among the six bus lines in the city, during the period of time January 01, 2018 to March 31, 2018. Table 36.3 shows a part of the originally collected data. The whole dataset consists of 2,874,878 rows (records), each with eight component variables. The meaning of the variables used in this work is shown in Table 36.4.

36.4.2 Data Preprocessing

The first phase of the data analysis is to process the data in order to construct a time series. The steps to obtain the descriptive statistics and the time series plot are the follows.

- Step 1: **Standardization of the raw data.** The variables which are not straightforward numeric or character, such as 'SITE_TIME', need to be standardized in the format that the SAS can recognize and interpret.
- Step 2: **Data extraction.** The daily passenger volume from the original database is extracted. For this, the 'DATA' and 'PROC' procedures are mainly used in SAS to create the datasets.
- Step 3: **Construct the time series.** After the datasets, including daily passenger flow volume in each bus line, are constructed, the graphical procedure in SAS is used to plot the time series for each bus line. In this work, the line No. 18 is chosen for case study. The time series plot of the daily passenger volume in the line No. 18 during the period January 01, 2018 to March 31, 2018 is shown in Fig. 36.1.

36.4.3 Model Building

Case 1: ARMA modeling with the original time series. According to the results of ADF test shown in Table 36.5, the p -value is less than 0.05 for a lag of 0, indicating that the null hypothesis can be rejected and the sequence is stationary, so the ARMA model is suitable to the original data. After calculating the BIC of the models with different order combinations, SAS shows the optimal order for the order selection by the ascending order of BIC value. And three candidate models with minimum BIC values, namely AR(3), ARMA(1, 1) and ARMA(1, 3) are chosen. The results of the parameter estimation and fitting statistics for each candidate models are summarized in Tables 36.6 and 36.7, respectively.

Table 36.3 Part of raw data in the database

LINE_NO	BUS_NO	MACH_NO	IS_UP_DOWN	LABEL_NO	UP_PASSENGER	DOWN_PASSENGER	SITE_TIME
28750	2420	1649572	1	7	8	0	02JAN2018:19:02:51
28751	2420	1649572	1	8	2	0	02JAN2018:19:04:12
28752	2420	1649572	1	9	2	0	02JAN2018:19:05:51
28753	2420	1649572	1	10	0	0	02JAN2018:19:06:45
28754	2420	1649572	1	11	1	0	02JAN2018:19:08:30
28755	2420	1649572	1	12	1	0	02JAN2018:19:09:49
28756	2420	1649572	1	13	0	0	02JAN2018:19:12:35
28757	2420	1649572	1	14	0	2	02JAN2018:19:14:42
28758	2420	1649572	1	15	0	0	02JAN2018:19:16:38
28759	2420	1649572	1	16	2	0	02JAN2018:19:18:05
28760	2420	1649572	1	17	0	0	02JAN2018:19:19:38
28761	2420	1649572	1	18	0	0	02JAN2018:19:20:36
28762	2420	1649572	1	19	0	0	02JAN2018:19:22:00

Table 36.6 Parameter estimation results of three candidate models (in Case 1)

(p, q)	MU	φ_1	φ_2	φ_3	θ_1	θ_2	θ_3
(3, 0)	2227.2	0.65563	0.05421	0.27748			
(1, 1)	2412.4	0.98886	–	–	0.37098	–	–
(1, 3)	2154.8	0.98111			0.35491	0.08107	–0.30409

Table 36.7 Fitting statistics of three candidate models (in Case 1)

(p, q)	AIC	SBC	MAPE
(3,0)	1445.263	1455.262	13.57413
(1,1)	1446.71	1454.209	13.82645
(1,3)	1445.193	1457.693	13.81004

Although ARMA(1,1) and ARMA(1,3) have the smallest SBC and the smallest AIC value, respectively, their MAPE values are much higher than that of AR(3). In comparison, AR(3) model has the smallest MAPE value indicating the highest prediction precision, and either its AIC or BIC value is slightly higher than the minimum indicating a relatively good goodness of fit. Therefore, AR(3) is chosen as the optimal model in this case. After implementing the parameter estimation by goodness of fit test for AR(3) model in SAS, all the AR coefficients are significant, so the optimal model is determined as below.

$$Z_t = 0.65563Z_{t-1} + 0.05421Z_{t-2} + 0.27748Z_{t-3} + a_t \tag{36.21}$$

As shown in Fig. 36.2, the residual diagnostics in SAS shows ACF value of the residual sequence is almost 0, and the white noise probability is greater than 0.05, which indicates that there is no dependence between the residuals and the AR(3) model has extracted all the useful information from the historical time series. Besides, the histogram and QQ-plot of residuals (Fig. 36.3) show that the residual sequence follows normal distribution, indicating the model is adequate. The ten-step ahead prediction of the passenger flow volume by using the model (36.21) and the comparison between actual and prediction are shown, respectively, in Table 36.8; Fig. 36.4.

Although the diagnostic results show that AR(3) is adequate for the sequence fitting, 95% confidence interval of the prediction is very wide. Since the wider the confidence region is, the lower the prediction accuracy is, the prediction especially in the long term may not be accurate.

Case 2: ARMA modeling with the first-order differenced time series.

According to the ACF plot shown in Fig. 36.5, even the autocorrelation decreases exponentially, it does not fall into the confidence interval until lag 5. Considering that the ACF decays gradually, not rapidly to zero, the time series is regarded as non-stationary and needs to be differenced, so the ARIMA model is applied to fit the data. And two models with minimum BIC value are chosen as candidate models,

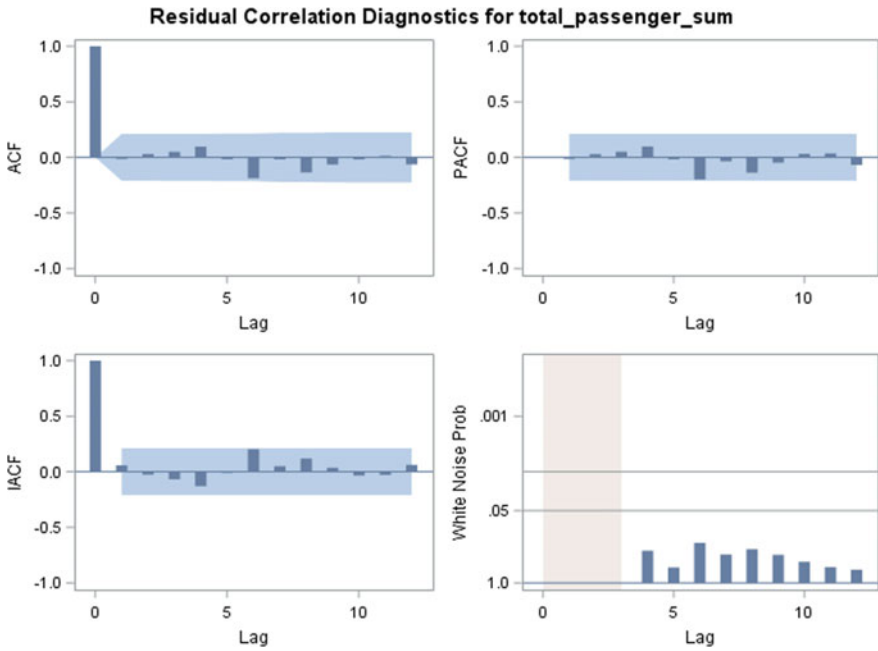


Fig. 36.2 Residual correlation diagnostics for AR(3) model

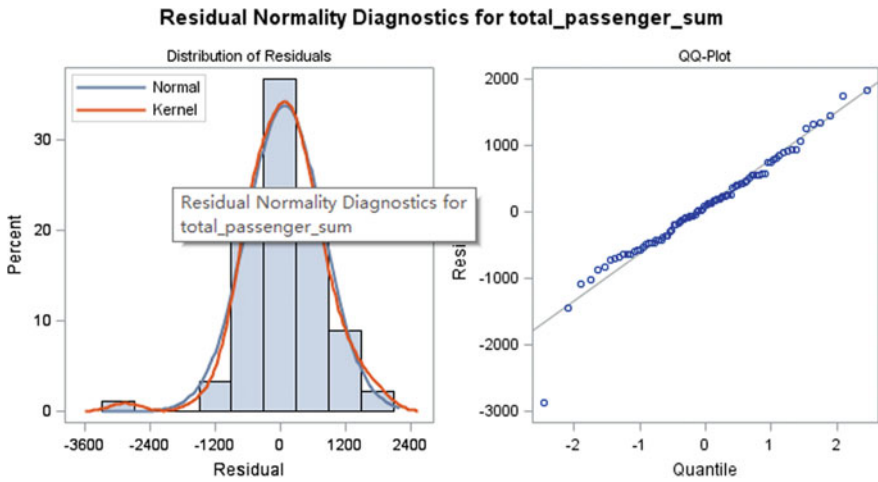


Fig. 36.3 Residual normality diagnostics for AR(3) model

Table 36.8 Prediction based on AR(3) model

Date	Forecast for the daily passenger volume	Std error of forecast	Lower 95% confidence limit	Upper 95% confidence limit
20180401	4910.496318	726.7519712	3486.088629	6334.904007
20180402	5093.005425	869.0218861	3389.753826	6796.257023
20180403	4937.714098	937.5256512	3100.197587	6775.230609
20180404	4904.203801	1043.474524	2859.031316	6949.376287
20180405	4924.457867	1137.024275	2695.931238	7152.984497
20180406	4892.829962	1211.618031	2518.102259	7267.557665
20180407	4863.893423	1284.335326	2346.642439	7381.144406
20180408	4848.827417	1353.560756	2195.897083	7501.75775
20180409	4828.60494	1417.048994	2051.239948	7605.969932
20180410	4806.500485	1476.948875	1911.733884	7701.267087

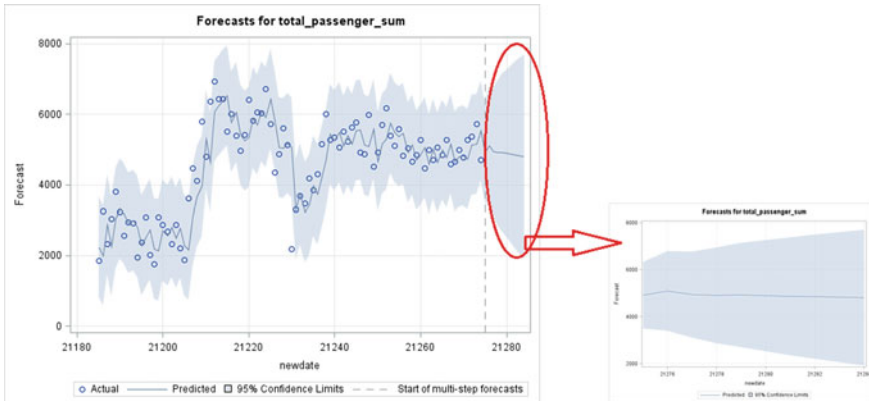


Fig. 36.4 Actual values against prediction based on AR(3)_{ij} model

namely ARIMA(0, 1, 1) and ARIMA(2, 1, 0). The results of the parameter estimation and fitting statistics for each candidate models are summarized in Tables 36.9 and 36.10, respectively.

It can be seen ARIMA(2, 1, 0) has a smaller AIC value indicating a higher goodness of fit, and a smaller MAPE value indicating a higher prediction precision, while its SBC value is slightly higher than ARIMA(0, 1, 1). Therefore, ARIMA(2, 1, 0) model is chosen as the optimal one for the first-order differenced time series. After implementing the parameter estimation by goodness of fit test, all the coefficients are significant. So the optimal model is determined as below.

$$(1 - B)Z_t = -0.33412(1 - B)Z_{t-1} - 0.28956(1 - B)Z_{t-3} + a_t \quad (36.22)$$

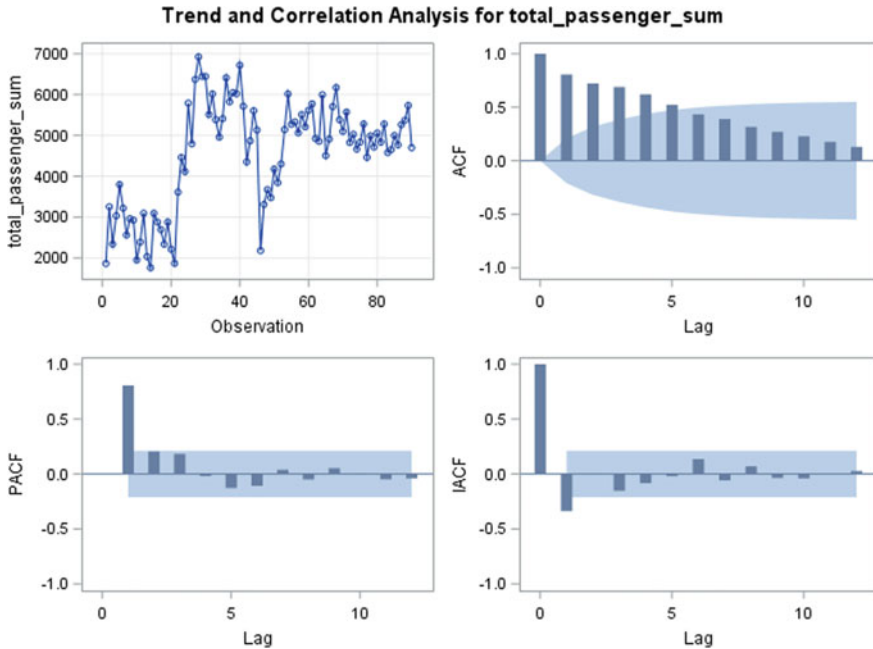


Fig. 36.5 Trend and ACF plots for the original time series

Table 36.9 Parameter estimation results of two candidate models (in Case 2)

(p, d, q)	MU	φ_1	φ_2	φ_3	θ_1	θ_2	θ_3
(0, 1, 1)	32.28610				0.37507		
(2, 1, 0)	32.81746	-0.33412	-0.28956	-		-	-

Table 36.10 Fitting statistics of three candidate models (in Case 2)

(p, d, q)	AIC	SBC	MAPE
(0, 1, 1)	1429.956	1434.934	13.9703
(2, 1, 0)	1428.162	1435.628	13.64128

By performing the residual diagnostics (similar to Case 1), it is observed that there is no dependence between the residuals and the ARIMA(2, 1, 0) model has extracted all the useful information from the time series. Besides, the residual sequence follows normal distribution, indicating the model is adequate. The ten-step ahead prediction of the passenger flow volume by using the model (36.22) and the comparison between actual and prediction are shown, respectively, in Table 36.11; Fig. 36.6.

Although the diagnostic results show that ARIMA(1, 2, 0) is adequate for the sequence fitting, its 95% confidence region of the prediction from the model is still very wide.

Table 36.11 Fitting statistics of three candidate models (in Case 2)

Date	Forecast for the daily passenger volume	Std error of forecast	Lower 95% confidence limit	Upper 95% confidence limit
20180401	5073.267	737.6109	3627.576	6518.958
20180402	5105.553	869.797	3400.783	6810.324
20180403	5137.84	984.39	3208.471	7067.208
20180404	5170.126	1086.969	3039.706	7300.545
20180405	5202.412	1180.668	2888.344	7516.479
20180406	5234.698	1267.46	2750.522	7718.873
20180407	5266.984	1348.678	2623.624	7910.343
20180408	5299.27	1425.275	2505.783	8092.757
20180409	5331.556	1497.96	2395.608	8267.504
20180410	5363.842	1567.279	2292.033	8435.652

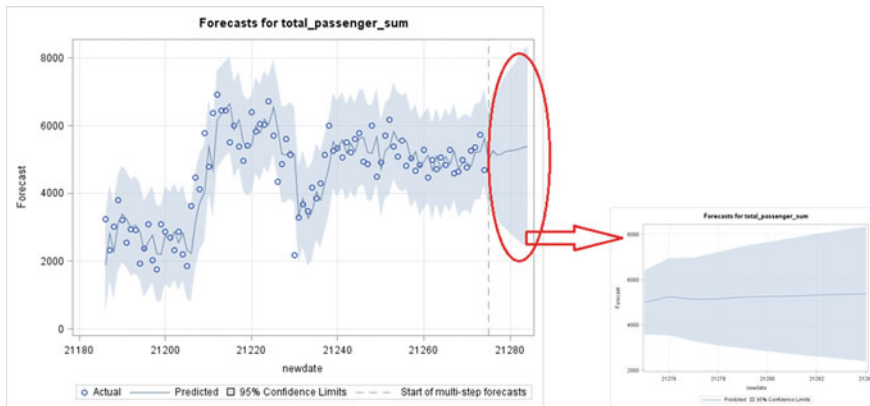


Fig. 36.6 Actual values against forecasts based on ARIMA(2, 1, 0) model

Case 3: ARMA modeling with quadratic function trend. According to the original time series plot in Fig. 36.1, it is observed that the passenger flow time series may have a quadratic trend. The two trend variables, `_LINEAR_` and `_SQUARE_` (as shown in Table 36.12), representing linear and quadratic relationships, respectively, are pre-generated.

Then, the same steps as in the previous two cases are followed to build a quadratic ARMA model. And three models with minimum BIC value as candidate models are chosen, namely Quadratic + AR(3), Quadratic + ARMA(1, 1) and Quadratic + ARMA(1, 3). The results of the parameter estimation and fitting statistics for each candidate models are summarized, respectively, in Tables 36.13 and 36.14.

Compared with the other two models, the Quadratic + ARMA(1, 3) model has the smallest AIC and MAPE indicating the highest goodness of fit and the highest

Table 36.12 Part of dataset with two trend variables

	Date	total_passenger_sum	_LINEAR_	_SQUARE_
1	180101	1863	1	1
2	180102	3251	2	4
3	180103	2339	3	9
4	180104	3033	4	16
5	180105	3793	5	25
6	180106	3219	6	36
7	180107	2558	7	49
8	180108	2957	8	64
9	180109	2922	9	81
10	180110	1950	10	100

prediction precision. Although Quadratic + ARMA(1,1) has the smallest SBC, its MAPE is the highest indicating the lowest prediction accuracy. Therefore, the Quadratic + ARMA(1, 3) is chosen as the optimal model with quadratic trend.

After implementing the parameter estimation by goodness of fit test, it can be seen that not only the MA and AR coefficients, but also the coefficients of linear and quadratic trend variables are significant. So the optimal model is determined as below.

$$\begin{aligned}
 Z_t = & 113.79672t - 0.92369t^2 + 0.76222Z_{t-1} \\
 & + a_t - 0.26487a_{t-1} - 0.05254a_{t-2} + 0.35486a_{t-3}
 \end{aligned}
 \tag{36.23}$$

By performing the residual diagnostics (similar to Case 1), it is observed that there is no dependence between the residuals and the Quadratic + ARMA(1, 3) model has extracted all the useful information from the time series. Besides, the histogram and QQ-plot (obtained by using the same method in Case 1) show that the residual sequence follows normal distribution, indicating the model is adequate.

Using the model (36.23), the ten-step ahead prediction and the comparison between actual and prediction are shown, respectively, in Table 36.15; Fig. 36.7.

The 95% confidence region width is significantly narrower, but the prediction does not describe the rapid growth at the end of the sequence, so probably it is caused by some external factors such as weather and holiday policies. If further improvements are needed, the external influences must be included in the model.

36.5 Conclusion

In this work, the prediction on the passenger flow volume in the bus transpiration system is performed, by using three kinds of time series models: AR, ARIMA and

Table 36.13 Parameter estimation results of three candidate models (in Case 3)

Quadratic + (p, q)	MU	φ_1	φ_2	φ_3	θ_1	θ_2	θ_3	a	b
Quadratic + (3,0)	2009.0	0.57459	0.006668	0.20647	-	-	-	107.78168	-0.86291
Quadratic + (1,1)	2075.2	0.82535	-	-	0.26871	-	-	106.02820	-0.85541
Quadratic + (1,3)	1895.5	0.76222	-	-	0.26487	0.05254	-0.35486	113.79672	-0.92369

Table 36.14 Fitting statistics of three candidate models (in Case 3)

Quadratic + (p, q)	AIC	SBC	MAPE
Quadratic + (3, 0)	1442.351	1457.35	13.73232
Quadratic + (1, 1)	1443.143	1455.642	14.00114
Quadratic + (1, 3)	1439.571	1457.07	13.43795

Table 36.15 Forecasts based on Quadratic + ARMA(1, 3) model

Date	Forecast for the daily passenger volume	Std error of forecast	Lower 95% confidence limit	Upper 95% confidence limit
20180401	4934.188	693.2455	3575.452	6292.924
20180402	5074.388	774.2533	3556.879	6591.896
20180403	4629.014	806.6707	3047.968	6210.059
20180404	4536.912	908.7964	2755.704	6318.121
20180405	4450.85	963.1695	2563.073	6338.628
20180406	4368.952	993.3936	2421.936	6315.967
20180407	4289.787	1010.538	2309.169	6270.406
20180408	4212.268	1020.367	2212.386	6212.15
20180409	4135.564	1026.034	2124.575	6146.553
20180410	4059.041	1029.312	2041.627	6076.455

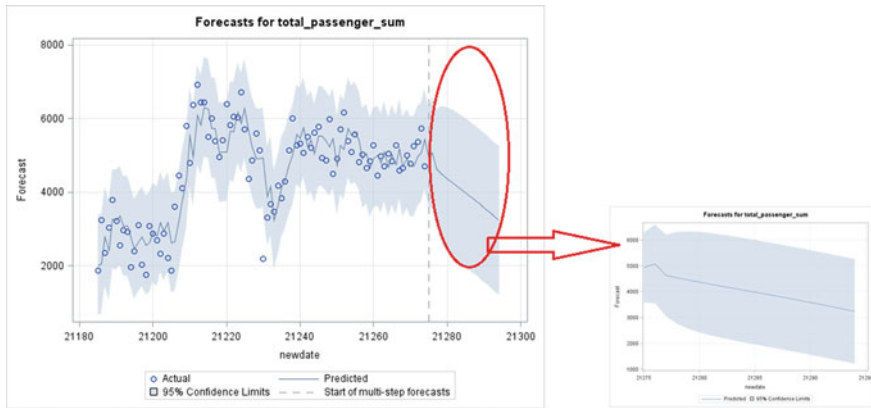


Fig. 36.7 Actual values against forecasts based on Quadratic + ARMA(1, 3) model

Table 36.16 Fitting statistics of three optimal models

Model	AIC	SBC	MAPE	MSE
AR(3)	1445.263	1455.262	13.57413	504694.3
ARIMA (2,1,0)	1428.162	1435.628	13.64128	531843.6
Quadratic + ARMA (1,3)	1439.571	1457.07	13.43795	443210.1

quadratic ARMA. At first, the bus IC card payment records were transformed into a time series, which represents the daily passenger volume in line No. 18. Then, the time series analysis was used and two optimal models, AR(3) and ARIMA(2, 1, 0), were found. Both models performed well in terms of goodness of fit but failed to attain accurate predictions. In order to achieve a higher prediction accuracy, the ARMA model with the quadratic trend was further explored and combined, and a Quadratic + ARMA(1, 3) model was established for the time series, which achieves a better balance between fitting and forecasting. The fitting statistics of those models are shown in Table 36.7.

Each model has its own advantages and disadvantages. They are discussed as follows one by one.

- **AR(3)**: The AR(3) model has no obvious advantages and disadvantages, because its performance is not outstanding either in goodness of fit or prediction accuracy. Its SBC, MAPE and MSE value are all in the middle level, except its AIC value is slightly higher than the other two models. The only advantage worth mentioning is that since differencing the process is not needed, this model is the simplest and the most straightforward one, and the cost is thus the lowest for the application.
- **ARIMA(2, 1, 0)**: In terms of fitting effect, the ARMA(2, 1, 0) model has the lowest AIC and SBC values indicating the highest goodness of fit. However, it owns the highest MAPE and MSE among the three models, indicating the greatest deviation between the predicted value and the true value. Moreover, its prediction confidence region widens over time, so it may perform poorly in the long-term prediction. But since it has the best fitting effect, it can accurately describe the surge trend at the end of the original time series, so the prediction result will be reliable when the model is used to predict the most recent value.
- **Quadratic + ARMA(1, 3)**: Compared with the other two models, Quadratic + ARMA(1, 3) model has the smallest MAPE and MSE value, so it achieves the highest prediction accuracy. Most importantly, this model has a unique advantage over the others, and it has a narrower prediction confidence interval of a constant width over time, so it will perform more effectively with high prediction accuracy.

The initial objective of this project and the main demand from the traffic management is to improve the forecast accuracy. Due to this, the accuracy of the prediction is the most important factor for the solution performance evaluation. So it can be concluded that the Quadratic + ARMA(1, 3) model is the most appropriate, compared to the other two models. Although ARIMA(2, 1, 0) model fits the current data the best and its short-term prediction shows relatively higher volatility, it may be more useful for short-term prediction.

36.6 Open Questions and Potential Improvements

Although the ARMA model with quadratic function trend performs best in our case, its application range is limited, because the time sequence must show a quadratic

trend. In the reality, only the short-term change of passenger flow may show such a trend. For the long-term daily passenger flow, if the data span is more than one year, it usually fluctuates within a limited range near a fixed value. So the stationary time series model may be more suitable for such kind of data. In addition, in view of the change of daily passenger flow in certain city, a seasonal factor with week cycle may be considered because of the difference of the commuting time between weekdays and weekends. In this case, a seasonal ARIMA model may be built to fit the series.

As mentioned in the end of Sect. 36.4, the time series method has limitations. When the prediction time span is long, only a rough future trend line can be obtained, but not the specific volatility. In order to accurately describe the future fluctuations, more external factors, such as weather, temperature, holidays and events, might be introduced into the model. When the historical data is updated continuously and the sample size is increasing, the algorithm should be updated and adjusted accordingly.

Acknowledgements The authors would like to acknowledge the support of Xiongdì Shenzhen Emperor Technology Company for the research fund (RDS10120190006) and the data used in this work.

References

1. Chen, Y., Wang, D.: Intelligent Traffic Information Collection, Analysis and Application. China Communication Press (2011)
2. Zhou, C., Zhang, Z., Tang, W.: System and methods of passenger demand prediction on bus network. *Comput. Sci.* **45**, 527–535 (2018)
3. Delwiche, L.D., Slaughter, S.J.: The little SAS book: A Primer, 5th edn. SAS Institute (2012)
4. Yule, G.U.: On a method of investigating periodicities in disturbed series, with Special Reference to Wolfer's sunspot numbers. *Philos. Trans. R. Soc. London, Ser. A* **226**, 267–298 (1927)
5. Walker, G.: On periodicity in series of related terms. *Proc. R. Soc. London, Ser. A.* **131**(818) (1931)
6. Slutsky, E.: The summation of random causes as the source of cyclic processes. *Econometr. Soc.* **5**(2), 105–146 (1937)
7. Wold, H., Kendall, M.: A study in the analysis of stationary time series. *J. Roy. Stat. Soc.* **102**(2), 295–298 (1939)
8. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control, 5th edn. Wiley, London (2015)
9. Xia, K.: Deep Analysis of SAS: Data Processing, Analytical Optimization and Business Applications. China Machine Press (2015)
10. Zhu, H.: N-day average volume based time-series analysis for passenger flow of metro. In: 2010 International Conference on Multimedia Information Networking and Security (2010)
11. Li, Z., Bi, J., Li, Z.: Passenger flow forecasting research for airport terminal based on SARIMA time series model. *IOP Conf. Ser. Earth Environ. Sci.* **100**(1), 1–7 (2017)
12. Xu, X., Dou, Y., Zhou, Z., Liao, T., Lu, Y., Tan, Y.: Railway passenger flow forecasting based on time series analysis with big data. In: Chinese Control and Decision Conference, pp. 3584–3590 (2018)
13. Jamil, M.S., Akbar, S.: Taxi passenger hotspot prediction using automatic ARIMA model. In: 2017 3rd International Conference on Science in Information Technology (2017)

14. Brockwell, P.J., Davis, R.A.: *Time Series: Theory and Methods*, 2nd edn. Springer, New York (2006)
15. Harris, R., Sollis, R.: *Applied Time Series Modelling and Forecasting*. Wiley, London (2003)
16. Dickey, D.A., Fuller, W.A.: Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.* **74**(36), 427–431 (1979)
17. Wei, W.S.W.: *Time Series Analysis: Univariate and Multivariate Methods*, 2nd edn. Pearson Addison Wesley (2006)
18. Tsay, R.S., Tiao, G.C.: Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models. *J. Am. Stat. Assoc.* **79**(385), 84–96 (1984)
19. Tsay, R.S., Tiao, G.C.: Use of canonical analysis in time series model identification. *Oxford Univ. Press* **72**(2), 299–315 (1985)
20. Choi, B.: *ARMA Model Identification*. Springer, New York (1992)
21. Hannan, E.J., Rissanen, J.: Recursive estimation of mixed autoregressive moving-average order. *Oxford Univ. Press* **69**(1), 81–94 (1982)
22. Guidolin, M., Pedio, M.: *Essentials of Time Series for Financial Applications*. Academic Press, London (2018). (an imprint of Elsevier)
23. Akaike, H.: A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**(6), 716–723 (1974)
24. Schwarz, G.: Estimating the dimension of a model. *Ann. Stat.* **6**(2), 461–464 (1978)
25. Xu, G.: *Statistical Forecasting and Decision-making*. Shanghai University of Finance & Economics Press (2016)

Chapter 37

Application of Sample Entropy to Analyze Consciousness in CLIS Patients



Shang-Ju Wu and Martin Bogdan

Abstract In this paper, an approach using sample entropy in order to detect the consciousness in two complete locked-in syndrome (CLIS) patients is presented. The typical symptom of CLIS patients is complete paralysis, but internal brain activities are supposed to be still available. On the other hand, there is no certainty about the actual state of consciousness in CLIS patients. For communication reasons and thus the patient's quality of life, it is an important problem to investigate consciousness in CLIS patients. A brain computer interface (BCI) potentially provides the family members a method to communicate with CLIS patients. There are arguments whether the CLIS patients are conscious or not. As consciousness is required to use BCI correctly, this study proposes to use sample entropy to uncover awareness from electroencephalography signals in CLIS patients. In a first proof of concept, data from two patients have been analyzed. The results for these two patients indicate that the use of sample entropy might be helpful to uncover awareness and thus to detect consciousness in CLIS patients.

37.1 Introduction

Locked-in syndrome (LIS) is a state perhaps caused by a stroke, car accident, or motor neuron diseases, such as amyotrophic lateral sclerosis (ALS) and unresponsive wakefulness syndrome (UWS). LIS patients with this medical condition have no movements of limbs and most of facial muscles, but consciousness is supposed to remain. Those patients are frequently misdiagnosed as having no consciousness. Nevertheless, there is one UWS case proving that the patient is awake after 20 years [1]. LIS patients often communicate only by eye or eyebrow movements before slipping into completely locked-in state. Patients enter the complete locked-in syndrome (CLIS) status while the last remaining eye movements and anal sphincter control

S.-J. Wu (✉) · M. Bogdan
Leipzig University, Augustusplatz 10, 04109 Leipzig, Germany
e-mail: shanglu@informatik.uni-leipzig.de

M. Bogdan
e-mail: bogdan@informatik.uni-leipzig.de

© Springer Nature Singapore Pte Ltd. 2021
S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation, Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_37

disappear [2, 3]. Chaudhary et al. [4] attempted electroencephalography (EEG) and near-infrared spectroscopy (NIRS) to interact with CLIS patients, but there are still many controversies [5]. The biggest challenge regarding the proof of consciousness in CLIS patients is that the CLIS patients cannot express themselves explicitly anymore; thus, it is difficult to prove the results with last evidence.

In order to detect the state of consciousness, this study proposes to analyze the continuously recorded electroencephalography (EEG) signals using sample entropy. The main objective of this research is to provide a possible method to confirm that consciousness which is present in these patients during a defined time period. The final goal of this investigation is to restore the communication channel between CLIS patients and the outside world via BCI.

The sections of this paper are organized as follows: First, information about the modus operandi, the used analysis software, and the dataset is presented. Then, the method of sample entropy is described. Third, the preliminary results are presented, and the possible influencing factors are discussed. Finally, future developments are considered.

37.2 Method

The flowchart of the proposed data processing approach is shown in Fig. 37.1. First, in order to reduce the computation time, the original EEG signals are down-sampled to 100 Hz. Second, the down-sampled signals were band pass filtered by a sixth-order 1–45 Hz Butterworth filter since thoughtfulness and awareness, which is closely related with consciousness, are considered to be in the beta band (13–30 Hz) [6] and interference of the 50 Hz power-line frequency is avoided. Finally, the sample entropy algorithm was applied to obtain a level of consciousness. All the data were analyzed using MATLAB R2018b.

37.2.1 Dataset

The dataset in question comprises the signals of electroencephalography (EEG) and electrooculography (EOG). The dataset provider published the analysis results from four patients [4]. The same codename is used to facilitate comparison for readers. Two of these patients completed more than 130 sessions over several weeks: patient B completed 56 sessions, patient F completed 80 sessions (we kept the denomination



Fig. 37.1 Data processing flowchart

as in [4]). Because of the high number of completed sessions from two patients, we decided to use the data from these two patients. The EEG data was recorded with an EEG amplifier (Brain Amp DC, Brain Products, Germany) using a sampling rate of 200 Hz.

Figure 37.2 shows the channel positions which were used to acquire EEG signals and four electrodes which were used to acquire the vertical and horizontal EOGs. Table 37.1 shows the electrodes used to record the EEG signals for different patients on different days.

Beside the effect that in LIS the EOG signals that are measured to reject its influence in the recorded EEG data, during the course of disease in ALS toward CLIS, the patients gradually lose the ability to control muscles and even eye movements, and thus, the EOG disappears. Therefore, in this paper, we focus only on the analysis of the source of brain waves, EEG signals.

Patient B is a 61 year old CLIS patient. He was diagnosed with ALS in May 2011. From April 2012 to December 2013, he was able to communicate with the MyTobii eye-tracking device. His family members attempted to train him to move his eyes to different sides to express “yes” and “no,” but the response was unstable.

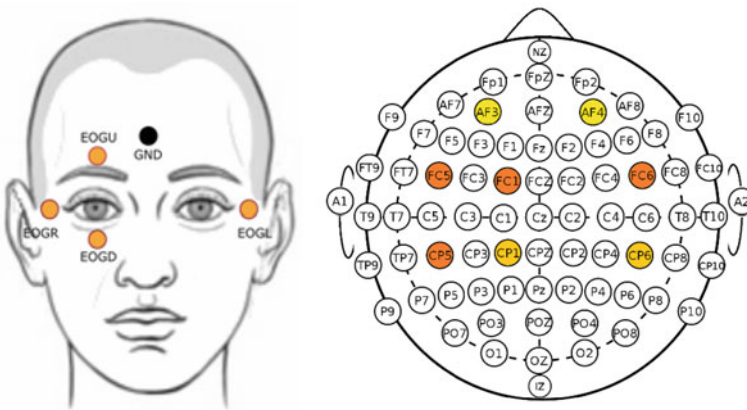


Fig. 37.2 Channels' positions. Left: the channels positions of electrooculographic (EOG). Right: the channels positions of electroencephalography (EEG)

Table 37.1 Electrodes recording the EEG signals of the two patients on different days

Codename	Day	EEG channels
Patient B	Day1	FC5, FC1, FC6, CP5, CP1, CP6, AF3, AF4
	Day2	FC5, FC1, FC6, CP5, CP1, CP6, AF3, AF4
Patient F	Day1	FC5, FC1, FC6, CP5, CP1, CP6
	Day2	FC5, FC1, FC6, CP5

Since August 2014, a communication was no longer possible. Patient F is a 68-year-old completely locked-in state patient. She was diagnosed with ALS in May 2007, locked-in syndrome (LIS) in 2009, and CLIS in May 2010. No communication channel was realized since 2010. Gallegos-Ayala et al. [7] described the details of this patient.

In total, over 22 h (130 sessions) of auditory experiments like [2] were recorded including trigger marks, the states of baseline, presentation, last word, and response. Beforehand of the study, the investigators discussed with family members in order to compile 200 personal questions known by the patients for sure and 40 open questions.

First, the investigators trained the patients ahead the experiments by asking the known questions, for example: “Berlin is the capital of Germany?”/“Berlin is the capital of France?,” in which the patient was expected to answer these paired “yes” or “no” questions.

During the experiments, the investigators asked the patients personal questions as well, such as “Is your husband’s name Joachim?” and also open questions the like “You feel good today?”/“You feel bad today?” related to the topic around the quality of life and compare the answers with the actual physiological status reported by the caretakers.

37.2.2 *Sample Entropy*

The entropy family is used frequently in nonlinear dynamic analysis to estimate the variability in time and frequency domain [8, 9]. This investigation utilizes sample entropy to analyze noninvasive electroencephalography (EEG) physiological signals as proposed in [10–12].

Richman and Moorman [13] improved the approximate entropy algorithm by developing sample entropy in order to apply it in short-term time series and to be more sensitive. The parameters of sample entropy are the same as approximate entropy, but sample entropy reduced the effect of self-matching. The entropy family is widely used in neuroscience, including evaluating consciousness approaches when patients are in anesthesia during surgery [14–16].

SampEn (m, r, N) indicates the sample entropy where m is the dimension selected in advance, r is the range of the tolerance coefficient selected in advance, and N is the number of data points respective to the data length. Pincus [17, 18] suggested that the appropriate number N of the corresponding data length should be in-between the number 10^m and 30^m . Thus, for 1000 data points in our case, we set the dimension m to 3 (and therefore, the length of the observed pattern). The range of the tolerance coefficient r is 0.2, which means 20% of the standard deviation of time series. Figure 37.3 shows a time series $X = x[1], \dots, x[i], \dots, x[N]$. The color band around the data point $x[1]$, $x[2]$, and $x[3]$ represents point $x[1] \pm r$, $x[2] \pm r$, and $x[3] \pm r$, respectively. All data points in the red band match the data point $x[1]$, and similarly, all the data points in the orange and yellow bands match the data points $x[2]$ and $x[3]$.

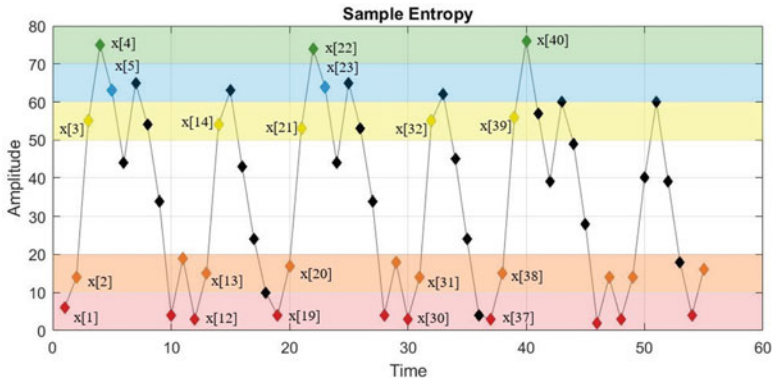


Fig. 37.3 A diagram where each point represents one data point in time domain to explain the operation of sample entropy depending on the row of occurrence over time

Consider the three components red-orange-yellow as consecutive sequence pattern $(x[1], x[2], x[3])$ and the four components red-orange-yellow-green as consecutive sequence pattern $(x[1], x[2], x[3], x[4])$. In this example, there are four red-orange-yellow sequences, $(x[12], x[13], x[14])$, $(x[19], x[20], x[21])$, $(x[30], x[31], x[32])$, and $(x[37], x[38], x[39])$, they match $x[1], x[2], x[3]$ on the same color bands, but only two red-orange-yellow-green sequences that match $x[1], x[2], x[3], x[4]$. Continuing that way with the next three-component sequence (orange-yellow-green) and the four-component sequence pattern (orange-yellow-green-blue), in this case, the number of matches of three-component pattern matches is two, and only one match for a four-component pattern. These numbers of matches are added to the previous numbers, the total number of three-component matches is six, and the total number of four-component matches is three. Now, repeat all possible sequence patterns, $(x[3], x[4], x[5], x[6]), \dots, (x[N - 3], x[N - 2], x[N - 1], x[N])$ to determine the ratio of all three-component pattern matches and four-component pattern matches. Then, the sample entropy is computing as follows:

$$\text{SampEn}(N, m, r) = -\log \frac{A^m(r)}{B^m(r)} \tag{37.1}$$

$$B^m(r) = (N - m)^{-1} \sum_{i=1}^{N-m} B_i^m(r) \tag{37.2}$$

$$A^m(r) = (N - m - 1)^{-1} \sum_{i=1}^{N-m} A_i^m(r) \tag{37.3}$$

where $B_i^m(r)$ is the match number of $x(j)$ with $x(i)$ according to the following conditions, $A_i^m(r)$ is for the situation of $m + 1$:

$$d[u_m(i), u_m(j)] = \max\{|x(i+k) - x(j+k)|\} < r \times SD \quad (37.4)$$

which must be less than a threshold, $R = r * SD$, where SD is the standard deviation of the time series $X = [x(1), x(2), \dots, x(N)]$ and r is the tolerance. X is divided into several selected sequences $u_m(i) = [x(i), x(i + 1), \dots, x(i + m - 1)]$ ($i = 1 \dots N - m + 1, k \in [0, m - 1], i \neq j$). Thus, the higher the value of sample entropy, the lower the self-similarity of the series, the higher the probability of producing a new signal, and finally, the more complicated is the data series. Otherwise, the smaller the value of sample entropy, the higher the self-similarity of the series, and thus, the lower the probability of producing a new signal, and consequently, the simpler is the data series.

37.3 Results

As proposed in Sect. 37.2.2, we have applied sample entropy to the dataset as shown in Fig. 37.3. We interpret therefore a higher value of sample entropy as higher brain activity and thus hypothetically more consciousness. This interpretation is based on the results shown in [4] where during the corresponding time slots (see Figs. 37.4, 37.5, 37.6, and 37.7 trigger marks), the experimenter received a good number of correct answers, thus indicating the consciousness of the patient. Figures 37.4, 37.5, 37.6, and 37.7 are showing the results for two patients over two days for each patient.

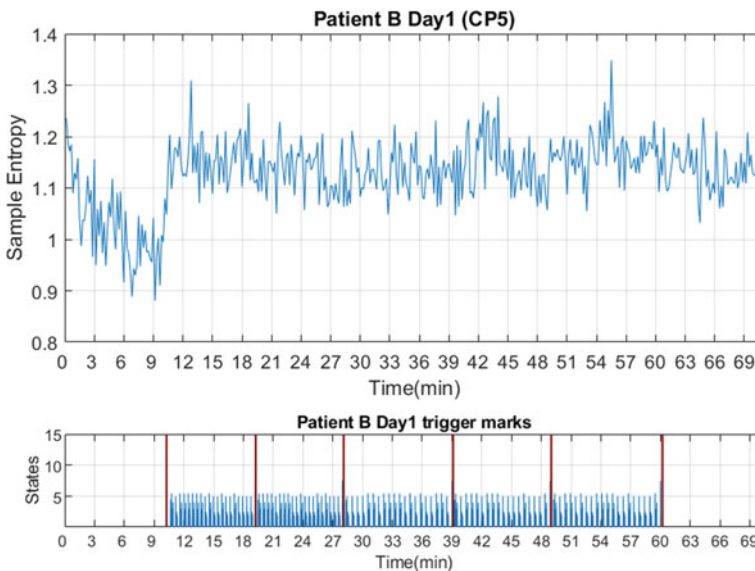


Fig. 37.4 Result of sample entropy (Patient B/Day 1)

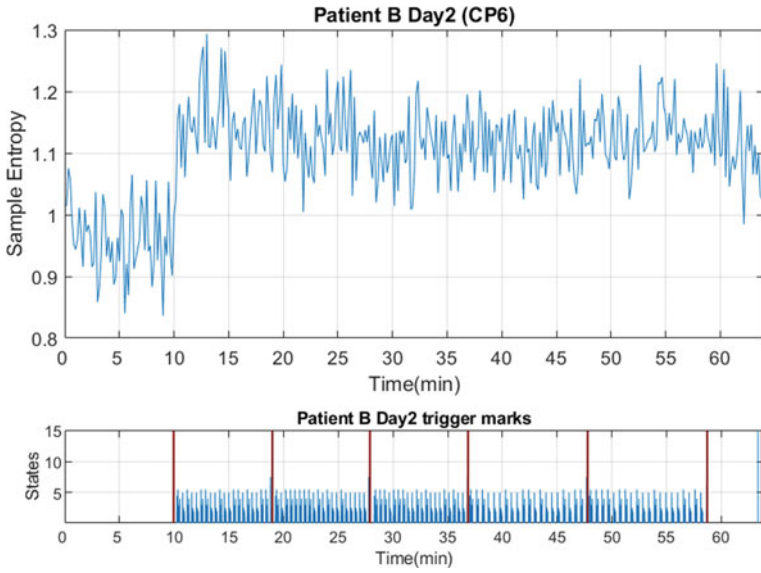


Fig. 37.5 Result of sample entropy (Patient B/Day 2)

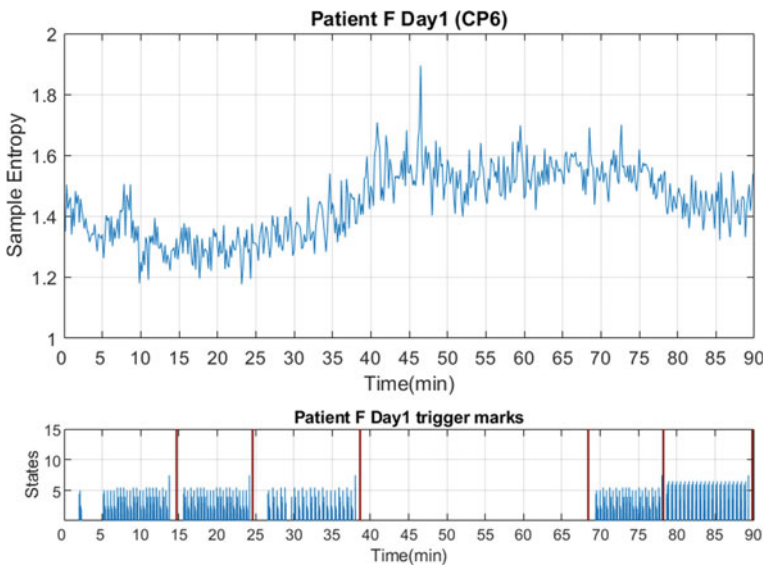


Fig. 37.6 Result of sample entropy (Patient F/Day 1)

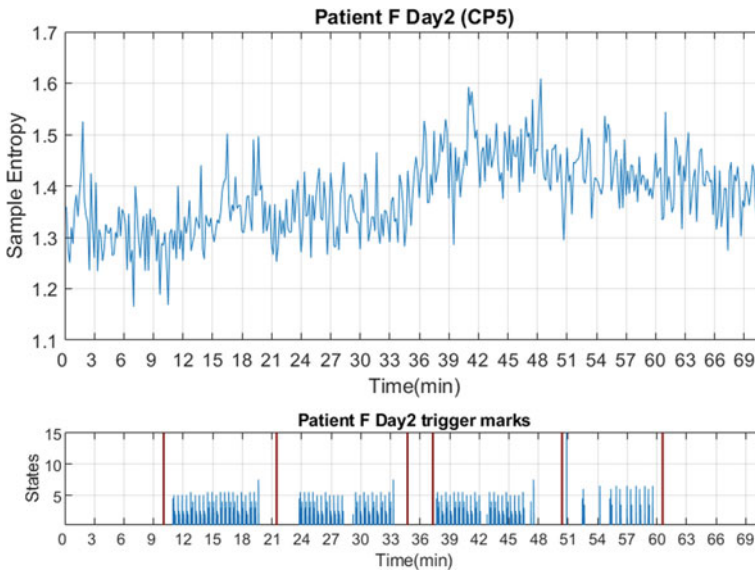


Fig. 37.7 Result of sample entropy (Patient F/Day 2)

These data were selected, since during the two days results to the questions, and therefore, consciousness was reported as being better than during the other days of experiments. In order to obtain relatively clear results in term of consciousness detection, the treated data should potentially contain corresponding information. For this reason, we preselected the two days as argued above.

In Fig. 37.4, we combined seven consecutive sessions over 69 min during day one. The figure below shows the time window between 10 and 60 min in which the investigators asked the questions. The time windows from 0 to 10 min and 60–69 min are the rest states. The top diagram shows the result of the sample entropy: We assume that the higher the value, the higher the relative consciousness of the patient. After 9 min, the value rises obviously, clearly showing the difference between rest and experience state as at that point the questions were started.

In Fig. 37.5, we combined seven consecutive sessions during day 2 all in all over 64 min. The figure below shows the time window between 10 and 59 min in which the investigator asked the questions. The periods from 0 to 10 min and 59–64 min are the rest states. After 10th min, the value of sample entropy rises obviously as well, what we would interpret again as a higher level of consciousness. There is a slow decline after 59 min. The results of patient B in these two days are consistent with the time window of trigger marks. The results for patient B in all other channels (FC5, FC1, FC6, CP5, CP1, CP6, AF3, AF4) have a similar trend.

In Fig. 37.6, a combination of six consecutive sessions during one day with over 90 min is shown for patient F, who performed less good as patient B in general. The figure below shows the time windows between 0 and 39 min and 68–90 min in which the investigator asked the questions. The period of 39–68 min is a rest state.

After the 38 min, the value of sample entropy rises slowly. After the 74 min, the value declines slowly. But the result is the opposite of the result of patient F. The trend of symmetrically positioned electrodes (CP5 vs. CP6, FC5 vs. FC6) is similar, whereas not symmetrical positioned electrodes differ. So, depending on the area over which the electrode is placed (e.g., over the Broca area or the Wernicke area), the task-related signal must be different and thus will indicate different corresponding aspects and show different trends on the related electrodes. We speculate that the level of consciousness depends on how difficult the questions were. Perhaps the questions asked by the investigators can promote the patients' thinking?

In Fig. 37.7, we show the consecutive combination of seven sessions during another day over all in all 70 min during day 2. The figure below shows the time windows between 10 and 35 min and 37–60 min in which the investigator asked the questions. The time windows of 0–10 min and 60–70 min are the rest states. After the 36 min, the value of sample entropy rises slowly. The value declines slowly after the 48th minute. The correct response rate is around 70% by functional near-infrared spectroscopy (fNIRS) and support vector machine (SVM) to ensure that patients are awake [2]. There is a similar result between channel CP5 and the other channels (FC5, FC1, FC6) for patient F on day 2. In the future, we will correlate the different types of questions and related feedback in order to refer to the value of sample entropy and to obtain further results.

37.4 Discussion

The results of patient B shows a relatively higher value of sample entropy while the time window is consistent with the communication period. But not all the results of patient F are in line with this trend. Perhaps, there is an effect related to the difficulty of the questions as well. Therefore, we need to classify with the correct response rate and compare with the type of questions for further analysis. Nevertheless, globally, the obtained results are correlating with the observations in [4]. Even though this does not prove the correctness of the approach in terms of detecting correctly the consciousness, it indicates that the approach might be correct. Remember that in CLIS patients, the final proof of consciousness cannot yet determined without any doubts since the patient cannot tell any more by any means, if she/he was conscious at the moment anymore.

37.5 Conclusion

In this study, sample entropy was proposed to demonstrate consciousness in two complete locked-in state patients. Preliminary results show that it can be hypothetically possible to use sample entropy in the time domain to detect the awareness respective to consciousness in the case of CLIS patients; still, it has to be mentioned

that the response rate and the type of questions can be important factors to explain the result presented here. At least these results provide an approach for the detection of the level of consciousness and give the possibility to interpret some meanings from EEG signals instead of supposing unconsciously suspecting them and may finally prove that CLIS patients can recognize the external stimulus and answer questions through a brain computer interface.

Remark that the presented results still state an indication toward consciousness detection in CLIS patients since yet no one other than the patient himself can state about his consciousness and thus proof the correctness of the obtained results at 100%—but cannot communicate it in a manner we understand due to his/her complete inability to communicate by any means we are used to. Nevertheless, we presented an approach that provides a good option to obtain a marker indicating the consciousness of a CLIS patient with a good probability. If this method is combined with competing methods as proposed in [19], it may significantly advance the solution of the consciousness detection problem in CLIS patients.

Acknowledgements Data was kindly provided by Prof. Dr. hc. mult. Niels Birbaumer and Dr. Ujwal Chaudhary from the Institute for Medical Psychology and Behavioral Neurobiology, University of Tübingen. We are grateful for providing the data to us.

References

1. Vanhaudenhuyse, A., Charland-Verville, V., Thibaut, A., Chatelle, C., Tshibanda, J.-F.L., Maudoux, A., Faymonville, M.-E., Laureys, S., Gosseries, O.: Conscious while being considered in an unresponsive wakefulness syndrome for 20 years. *Front. Neurol.* **9**, 671 (2018). <https://doi.org/10.3389/fneur.2018.00671>
2. Murguialday, A.R., Hill, J., Bensch, M., Martens, S., Halder, S., Nijboer, F., Schoelkopf, B., Birbaumer, N., Gharabaghi, A.: Transition from the locked in to the completely locked-in state: a physiological analysis. *Clin. Neurophysiol.* **122**(5), 925–933 (2011). <https://doi.org/10.1016/j.clinph.2010.08.019>
3. Soekadar, S.R., Born, J., Birbaumer, N., Bensch, M., Halder, S., Murguialday, A.R., Gharabaghi, A., Nijboer, F., Schölkopf, B., Martens, S.: Fragmentation of slow wave sleep after onset of complete locked-in state. *J. Clin. Sleep Med. JCSM Off. Publ. Am. Acad. Sleep Med.* **9**(9), 951–953 (2013). <https://doi.org/10.5664/jcsm.3002>
4. Chaudhary, U., Xia, B., Silvoni, S., Cohen, L.G., Birbaumer, N.: Brain-computer interface-based communication in the completely locked-in state. *PLoS Biol.* **15**(1), e1002593 (2017). <https://doi.org/10.1371/journal.pbio.1002593>
5. Spüler, M.: Questioning the evidence for BCI-based communication in the complete locked-in state. *PLoS Biol.* **17**(4), e2004750 (2019). <https://doi.org/10.1371/journal.pbio.2004750>
6. Schwender, D., Daunerer, M., Mulzer, S., Klasing, S., Finsterer, U., Peter, K.: Spectral edge frequency of the electroencephalogram to monitor “depth” of anaesthesia with isoflurane or propofol. *Br. J. Anaesth.* **77**(2), 179–184 (1996). <https://doi.org/10.1093/bja/77.2.179>
7. Gallegos-Ayala, G., Furdea, A., Takano, K., Ruf, C.A., Flor, H., Birbaumer, N.: Brain communication in a completely locked-in patient using bedside near-infrared spectroscopy. *Neurology* **82**(21), 1930–1932 (2014). <https://doi.org/10.1212/wnl.0000000000000449>
8. Costa, M., Goldberger, A.L., Peng, C.-K.: Multiscale entropy analysis of biological signals. *Phys. Rev. E* **71**, 21906 (2005). <https://doi.org/10.1103/PhysRevE.71.021906>

9. Costa, M., Goldberger, A.L., Peng, C.-K.: Multiscale entropy analysis of complex physiologic time series. *Phys. Rev. Lett.* **89**(6), 68102 (2002)
10. Yeragani, V.K., Pohl, R., Mallavarapu, M., Balon, R.: Approximate entropy of symptoms of mood: an effective technique to quantify regularity of mood. *Bipolar Disord.* 279–286 (2003)
11. Diambra, L., Bastos de Figueiredo, J.C., Malta, C.P.: Epileptic activity recognition in EEG recording. *Phys. A* **273**, 495–505 (1999)
12. Courtiol, J., Perdikis, D., Petkoski, S., Müller, V., Huys, R., Sleimen-Malkoun, R., Jirsa, V.K.: The multiscale entropy: Guidelines for use and interpretation in brain signal analysis. *J. Neurosci. Methods* **273**, 175–190 (2016)
13. Richman, J.S., Moorman, J.R.: Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* **278**(6), H2039–H2049 (2000). <https://doi.org/10.1152/ajpheart.2000.278.6.h2039>
14. Wu, S.-J., Chen, N.-T., Jen, K.-K., Fan, S.-Z.: Analysis of the level of consciousness with sample entropy a comparative study with bispectral index. *Eur. J. Anaesthesiol.* **32**, 11 (2015)
15. Wu, S.-J., Chen, N.-T., Jen, K.-K.: Application of improving sample entropy to measure the depth of Anesthesia. In: 2014 International Conference on Advanced Manufacturing (ICAM), Chiayi (2014)
16. Wu, S.-J., Chen, N.-T., Jen, K.-K., Shieh, J.-S., Fan, S.-Z.: The physiological signals EEG, ECG, and SpO2 are Applied to analyze the consciousness and anesthesia depth. In: 2013 Advances in Materials & Processing Technologies (AMPT), Taipei (2013)
17. Pincus, S.M.: Approximate entropy: a complexity measure for biological time series data. In: *Bioengineering Conference*, pp. 35–36 (1991)
18. Pincus, S.M., Goldberger, A.L.: Physiological time-series analysis: what does regularity quantify? *Am. J. Physiol. Heart Circ. Physiol.* **266**, H1643–H1656 (1994)
19. Adama, V.S., Wu, S.-J., Nicolaou, N., Bogdan, M.: Extendable Hybrid Approach to Detect Conscious States in a CLIS Patient Using Machine Learning. *EUROSIM, Logroño* (2019)

Chapter 38

Intelligent Tuning of PID Controller for Double-Link Flexible Robotic Arm Manipulator by Artificial Bee Colony Algorithm



A. Jamali, I. Z. Mat Darus, M. H. A. Talib, H. M. Yatim, M. S. Hadi, and M. O. Tokhi

Abstract Robotics system particularly robotic arm has received tremendous demand in various fields especially manufacturing industries. Robotic arm is highly needed to enhance production, to improve output, and reduce human error. The current robotics arm not only they are expensive and required specialist for maintenance, but they are also bulky and very heavy. Thus, the option is employing lightweight, stronger, and more flexible robotics arm. However, the lightweight robotic arm can be easily influenced by unwanted vibration which may lead to problems including fatigue, instability, and performance reduction. These problems may eventually cause damage to the highly stressed structure. This research focuses on the development of intelligent controller utilizing artificial bee colony (ABC) algorithm to tune proportional integral derivative (PID) parameters for controlling two-link flexible manipulator (TLFRM). The essential objective of the designing the controller is to improve the performance of desired position and vibration suppression of TLFRM. The MATLAB environment is utilized to verify the accomplishment of the recommended control system. An assessment is conducted to illustrate the efficiency of PID-ABC controller in terms of input tracking and vibration suppression. The results show that the system with embedded new proposed controller is capable to achieve preferred angle at decrease overshoot and the settling time is exceptionally much quicker. The vibration reduction demonstrated substantial improvement as compared to manual tuning method. Overall, the proposed controller for two-link flexible manipulator that is intelligent PID-ABC was successfully control the system to the preferred position with vibration suppression in the entire system.

A. Jamali (✉)
Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
e-mail: jannisa@unimas.my

I. Z. Mat Darus · M. H. A. Talib · H. M. Yatim
Universiti Teknologi Malaysia, 81310 Skudai, Johor Bahru, Malaysia

M. S. Hadi
Universiti Teknologi Mara, 40450 Shah Alam, Selangor, Malaysia

M. O. Tokhi
London South University, London, UK

38.1 Introduction

Robotics manipulator arm system has progressively become popular not only among the heavy industries but also among small to medium enterprise industries. This is meant for simple and repetitive tasks to increase their productivity. The demand led to the changes of physical configurations of a robot such that the link structure is longer and thinner and the material used is lightweight. The developed robotic arm by using the lightweight material introduces flexibility to the system. Due to that, the flexible arm manipulator (FRM) motion tracking control is considered as a challenging problem due to the system dynamic replicated a highly coupled nonlinear and time-varying. Despite the problem, the flexible robotic manipulators have several potential applications such as in space exploration, military, medical field, automotive, oil and gas and other industrial applications. In manufacturing industries, the demand for flexible robotic arm manipulator is more imperative in order to fulfill the needs of current industrial such as higher maneuverability, superior transportability, quicker response times, and lower power consumption.

Researchers proposed various control strategies for FRM such as passivity-based velocity feedback and strain feedback schemes [1], hybrid collocated and non-collocated PID controller [2], global terminal sliding mode [3], a genetic algorithm (GA)-based hybrid fuzzy logic control strategy [4], decoupling controller based on the cloud model [5], adaptive distributed control strategy [6] and decentralized controller based on linear matrix inequalities [7]. Besides, researchers also proposed various controller strategies to design multivariable (MIMO) systems for multi-link FLM, ranging from intelligent control [8–10] to adaptive control [11], sliding mode controller [12, 13], adaptive iterative learning control scheme [14], torque controller [15, 16], optimal nonlinear controller [17], and PDE-based controller [18]. Most of the listed control schemes incorporate both conventional and intelligent control strategies to compensate the drawback of each controller.

Despite various advance control strategies proposed for the industrial environment, simple controller in which employing decentralized control scheme is preferable particularly for MIMO frameworks. The decentralized control scheme has raised interest among researchers. This is due to the fact they are frequently successful of imparting an extraordinary overall accomplishment despite their handy structure and intuitiveness. Though decentralized controller structure constraints bring about certain performance deterioration if compared with centralized full controller systems, it still gains popularity due to hardware simplicity and employ failure tolerant structure. Subsequently, it is easy to implement and maintain by plant personnel apart from delivering an adequate performance. There are few examples that showcase the decentralized control which have been implemented to two-link flexible robotic manipulator (TLFRM) system. The work in [19, 20] has proposed the decentralized PI-PID controller for TLFRM through manual tuning. Then, the overall performance has been elevated by adding ILC which have been verified in the simulation. The linear matrix inequalities (LMI)-based PID control of a nonlinear two-link flexible robotic manipulator (TLFRM) incorporating payload have been reported in [21].

In [7], decentralized proportional integral derivative (PID) controller by incorporating bounding parameters of interconnection terms in LMI formulation for an n-link robotic manipulator system was proposed. Finally, another decentralized control strategy utilized neural network (NN) to approximate the ZN-PID for every link of TLFRM in [22].

Apart from that, Alam et al. [23] applied hybrid PD-PD/ILA tune by multi-objective genetic algorithm optimization for single-link flexible manipulator (SLFM). Tijani carried out a multi-objective optimization the use of differential evolution (MODE) for PID controller of SLFM [24]. Another researcher has proposed an expanded bacterial foraging algorithms (BFA) to fine-tune the PID controller of SLFM [25]. Bee algorithm has been successful to optimize the hierarchical PID parameter of SLFM in [26]. Finally, PSO is used to tune one of PID parameters of the hybrid PID-PID controller of SLFM [27].

The literatures disclose that the application of intelligent tuning is considered in both TLFRM and SLFM. However, the survey confirms that the unique type of evolutionary algorithm such as DE, BFA, and ABC provides an effective method in optimizing the PID controller confine only in SLFM. Thus, there are relatively few PID controllers and have been used in TLFRM compared to their SLFM. The reason can be associated with the problem in the tuning coupled system. Besides, most of the time, the tuning methods showed sluggish responses when applied to a non-minimum phase system like flexible manipulator.

This paper therefore proposed hybrid PID-ABC for TLFRM, whereby ABC is used to optimize the parameters' of PID controllers. The system is modeled via system identification in which NARX model structure is utilized and the nonlinear part is predicted by neural network (Sect. 38.2). Section 38.3 presents the ABC algorithm used for optimization. The proposed control schemes that are the collocated PID controller for position tracking and the non-collocated PID controller for endpoint vibration suppression are then described in Sect. 38.4. Section 38.5 discusses the simulation results in which include the assessment of the recommended controllers in terms of reference tracking and endpoint acceleration. The conclusion remarks are presented in Sect. 38.6.

38.2 Experimental Setup and System Identification

38.2.1 Robotics Manipulator Test Rig

The planar TLFRM is constructed as shown in Fig. 38.1. The developed rig was executed to mimic the actual angular motion of manipulator. There are four outputs acquired from the sensors that are encoders and accelerometer. The outputs characterize the hub angles and endpoint acceleration of every link, respectively. The test is conducted in 9 s for every individual movement and repeated for similar angle. In order to match the mechanical system with software, the sampling time of 0.01 s



Fig. 38.1 Setup of two-link flexible robotic manipulator rig

was applied. Figure 38.2 indicates the schematic layout to illustrate the integration among all devices.

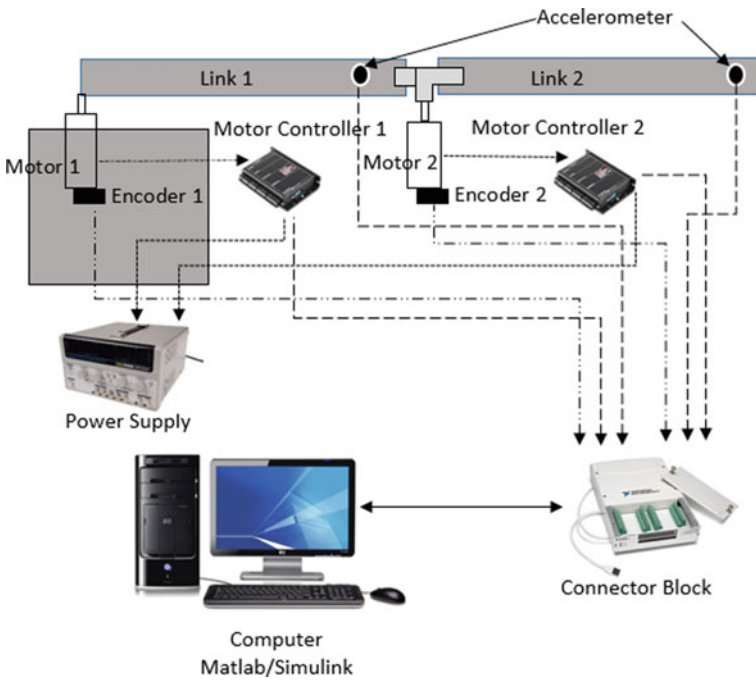


Fig. 38.2 Schematic diagram of TLFRM

38.2.2 System Identification

The TLFRM is classified under distinctly nonlinear. Thus, the development of non-parametric modeling is favored for this study utilizing neural network. NARX is chosen as model structure in this study because it has the simplest structure among nonparametric model. The research makes use of backpropagation for multilayer perceptron (MLP) neural network and Elman neural networks (ENN) for modeling the TLFRM system. All the developed models are validated via mean squared error (MSE). They are further validated via correlation test. The details can be found in [28].

38.3 Optimization Algorithm

After a system model is obtained, it can be utilized to predict the physical system behavior under different operating conditions or to control it. In this work, artificial bee colony (ABC) is employed to tune the PID parameters. In the bees' nature, they are classified into employed bees, onlooker bees, and scout bees. ABC system engaged both neighborhood search methods and global search methods. ABC algorithm contains the first half of employed bees and the second half comprises of the onlooker bees. The preliminary meals sources are randomly produced. Each employed bee generates a new candidate solution in the neighborhood of its present position. The neighbor food source v_{mi} is chosen. The fitness is determined. Then, a greedy decision is utilized between x_m and v_m . The quantity of a food source is evaluated by its profitability and the profitability of all food sources. After all the employed bees have finished the search processes, they share the information of their food sources with the onlooker bees through waggle dances. An onlooker bee evaluates the nectar data taken from all employed bees and chooses a food supply with a likelihood associated with its nectar amount. The procedure of ABC algorithm is illustrated in the diagram in Fig. 38.3.

38.4 Controller Development

The recommended control structure using ABC was incorporated to tune the PID controllers. Figures 38.4 and 38.5 present a block diagram of the closed-loop system for rigid body and flexible motion control, respectively.

Step input was used as input reference. The performance of PID controllers for hub angle models was observed in terms of t_r , t_s , M_p , and Ess . Meanwhile, the performances of vibration suppression were observed in terms of the attenuation of the first three mode of vibration. The objective functions of optimization are expressed based on the MSE of the hub angle error and endpoint vibration concealment.

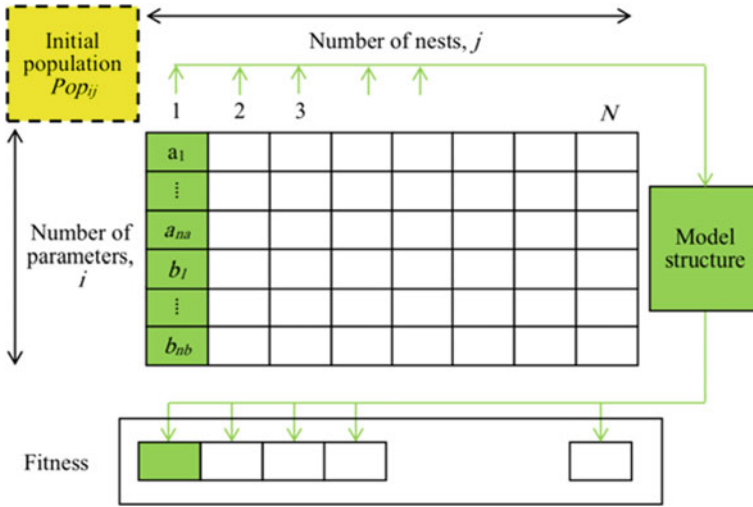


Fig. 38.3 Diagrammatic representation of generation the initial population

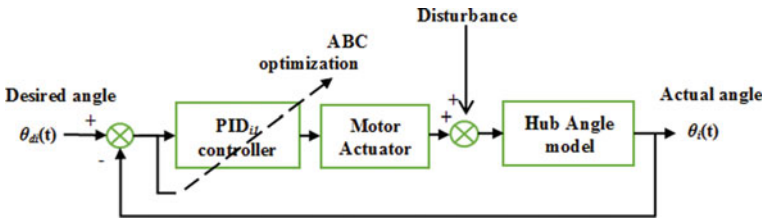


Fig. 38.4 PID control structure for hub angles 1 and 2

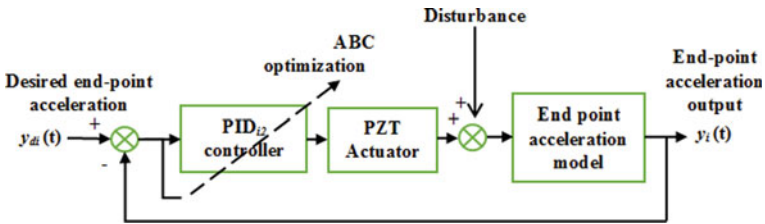


Fig. 38.5 PID control structure for endpoint accelerations 1 and 2

The collocated PIDi1 and non-collocated PIDi2 controller are applied for hub angle motion and flexible body motion, respectively. The two loops of each link ($i = 1, 2$) are consolidated to allow control inputs to the two-link flexible robotic manipulator framework.

38.4.1 Intelligent Collocated PID Controller

The details of hub angle motion controller can be described by referring to Fig. 38.6. The closed-loop signal of U_{mi} can be written as:

$$U_{mi}(t) = A_{mi}[(C_{mi}(t)e_{mi}(t))] \quad i = 1, 2 \tag{38.1}$$

Therefore, the closed-loop transfer function acquired as in Eq. (38.2);

$$\frac{\theta_i}{\theta_{di}} = \frac{[C_{mi}]A_{mi}H_{mi}}{1 + [C_{mi}]A_{mi}G_{mi}H_{mi}} \tag{38.2}$$

where θ_{di} and $\theta_i(t)$ represent reference hub angle and actual hub angle. U_{mi} is PID control input, A_{mi} is motor gain, and C_{mi} is PID controller. The controller gains are K_{Pi} , K_{Ii} , and K_{Di} .

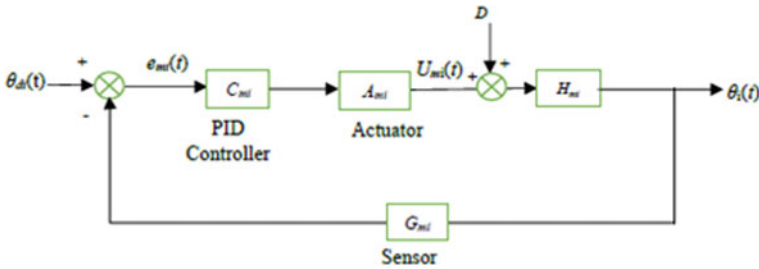


Fig. 38.6 Block diagram of control rigid body motion

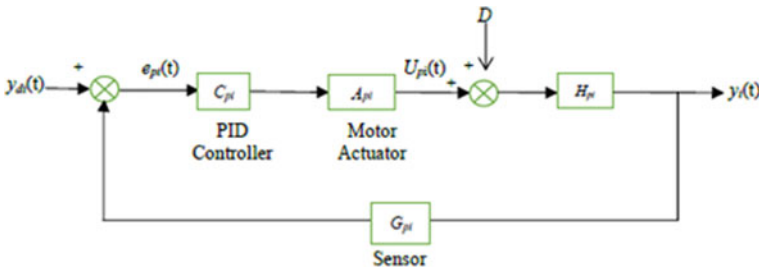


Fig. 38.7 Block diagram of control flexible body motion

38.4.2 Intelligent Non-collocated PID Controller

In Fig. 38.7, the block diagram for flexible body motion is presented to explain the details of the controller. The control input is given by;

$$U_{pi}(t) = A_{pi} [C_{pi}(t)e_{pi}(t)] \quad i = 1, 2 \quad (38.3)$$

where U_{pi} is PID control input, A_{pi} is piezoelectric gain, C_{pi} is PID controller. The controller gains are K_{pi} , K_{li} , and K_{Di} . The deflection output represents by y_i , and the desired deflection y_{di} is set to zero. Therefore, the closed-loop transfer function obtained as;

$$\frac{y_i}{y_{di}} = \frac{[C_{pi}]A_{pi}H_{pi}}{1 + [C_{pi}]A_{pi}G_{pi}H_{pi}} \quad (38.4)$$

The parameters of PID controller, K_{pi} , K_{li} , and K_{Di} were tuned accordingly to be fed into the U_{mi} and U_{pi} , thus grant satisfactory accomplishment of TLFRM. The accomplishment of the PID controller was evaluated by minimizing the MSE value.

38.5 Results and Discussion

TLFRM was modeled with the nonparametric identification approaches of neural network particularly MLP and ENN algorithm using NARX modeled structure. The best-obtained model system is then used in the control structure of TLFRM.

38.5.1 Modeling Results

Table 38.1 presents the achievement in modeling the TLFRM. The results reveal that all models predicted by ENN are one-sided. Thus, the TLFRM model obtained using MLP will be utilized in developing of control for hub angle and endpoint acceleration of the TLFRM.

38.5.2 Control Results

The recommended control strategies are applied on TLFRM system and executed through MATLAB/Simulink environment. The responses of the system are analyzed to optimize the performance of the recommended controllers.

Table 38.1 Summary of the performance achieved in modeling

	Model	Spec.	T (s)	MSE	Correlation test
MLP	Hub1	MS: [2 2 1], Ite: 150	3	0.0000685	Unbiased
	Hub2	MS: [2 2 1], Ite: 150	3	0.000752	Unbiased
	E.P. Acc1	MS: [2 2 1], Ite: 150	3	0.0025	Unbiased
	E.P. Acc2	MS: [2 2 1], Ite: 150	3	0.0049	Unbiased
ENN	Hub1	MS: [8 8 1], Ite: 150	2	0.0047	Biased
	Hub2	MS: [8 8 1], Ite: 150	2	0.0023	Biased
	E.P. Acc1	MS: [8 8 1], Ite: 150	3	0.018	Biased
	E.P. Acc2	MS: [8 8 1], Ite: 150	3	0.015	Biased

Hub Angle Motion The hub angles were controlled by the collocated PID-ABC controller individually. The TLFRM system is required to comply with a step input of 1 rad to test the hub tracking input of link 1 and link 2. The parameters of PID controllers are obtained via ABC algorithm. The tuning is initialized by setting the number of iterations to 15 and varying the number of colony size from 10 to 50. The same procedure was repeated for 50 maximum iterations. It was found that the satisfactory result was obtained with 50 colony sizes at 15th iteration for both hub angles. Figure 38.8a, b exhibits the 15 iterations of MSE convergence of ABC for hub angle.

The fitness function of ABC optimization is formed in such a way to reduce the tracking error via MSE values. The convergence MSE values with regard to the PID parameters obtained are organized in Table 38.2. Numbers of the simulation were repeated with different colony sizes.

The results were compared with manual tuning method to examine the significant of using ABC algorithm. The controller performances are presented in Table 38.3.

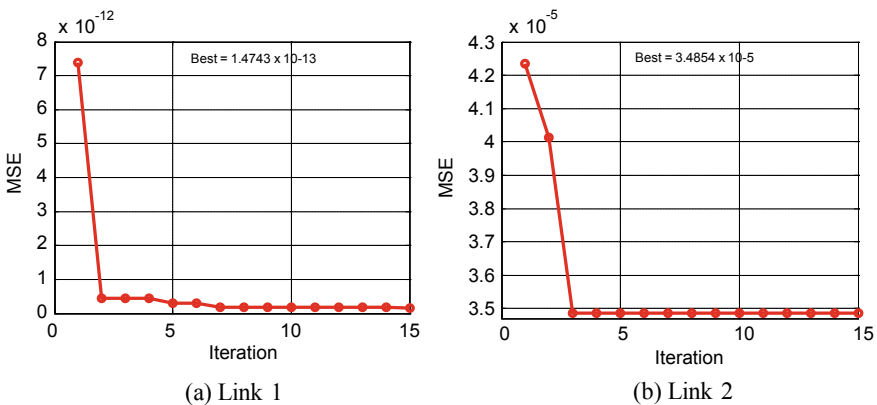


Fig. 38.8 ABC convergence for hub angle

Table 38.2 Convergence of MSE for ABC algorithm

Parameter	MSE	PID parameters		
		K_P	K_I	K_D
Hub angle 1	1.4743×10^{-13}	6.54	20.5	49.43
Hub angle 2	3.4854×10^{-05}	5.48	28.3	13.72

Table 38.3 Parameters and performance of hub input tracking

	Link	PID parameters			Tracking capabilities		
		K_p	K_i	K_d	Rise Time	OS (%)	T_s (s)
PID	Hub 1	2	25	3	0.123	3.247	5.163
	Hub 2	2	57	3	0.066	2.012	2.164
PID-ABC	Hub 1	6.54	20.5	49.43	0.044	1.061	1.078
	Hub 2	5.48	28.3	13.72	0.028	0.869	1.049

The response of the hub angle for both links is shown in Figs. 38.9 and 38.10. The proposed PID-ABC controller achieved an acceptable hub angle response. It is exceptionally important enhancement in terms of rate overshoot and settling time. The TLFRM system reached the required angle at reduce overshoot by employing the recommended approach that is 67 and 56% improvement as compared to the conventional method and faster settling time that is from 5.1633 to 1.0783 s for hub 1 and from 2.1635 to 1.0499 s for hub 2.

Flexible Body Motion The PID-ABC controllers were also executed to TLFRM system to effectively stifle the vibration at the endpoint of link 1 and link 2 independently. The desired output is set to zero to minimize the vibration in the system. The parameters of PID controllers are also acquired via ABC algorithm. It was found that

Fig. 38.9 Input tracking of hub 1

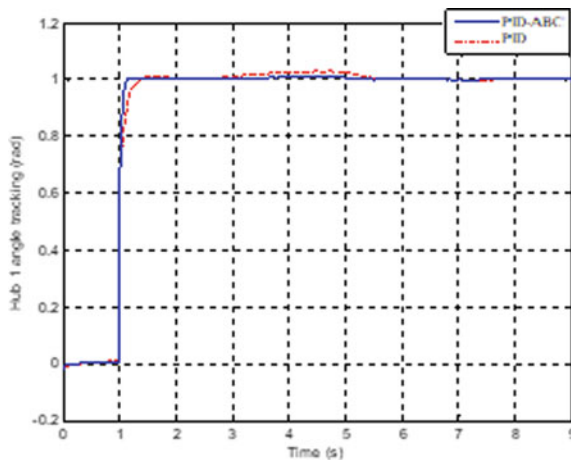
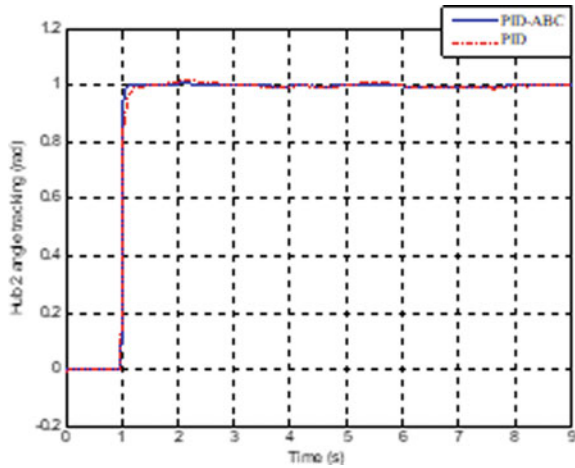


Fig. 38.10 Input tracking of hub 2



the satisfactory result was obtained with 50 colony sizes at 25th iteration for end-point acceleration suppression. Figure 38.11a, b displays the 25 iterations of MSE convergence of ABC for endpoint acceleration. It reveals that ABC optimization merges very fast and yields a small value of MSE for all the controllers. Besides, it was discovered that when the number of iterations higher or the number of colony sizes were set to greater values, there were no noteworthy improvement of MSE.

The results were compared with manual tuning method of PID controllers to assess the noteworthy of utilizing the ABC algorithm. The controller parameters obtained, and their performances are organized in Table 38.4.

The table displays that the PID-ABC controller accomplished better MSE level as compared to manual tuning method for controlling flexible body motion of both link 1 and link 2. This is portrayed in the simulation results of vibration suppression as

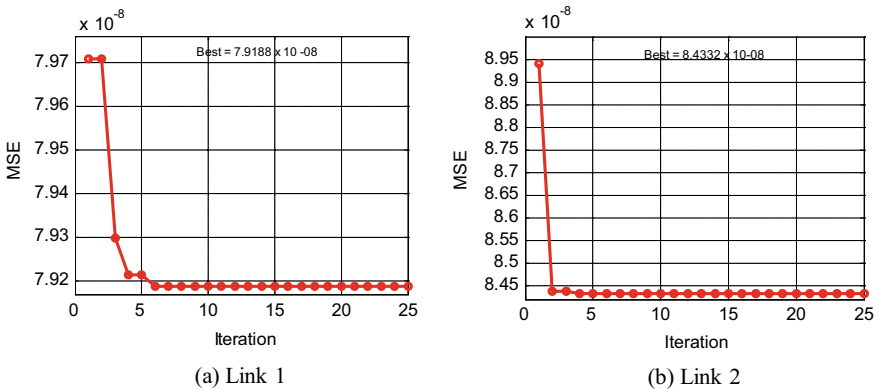


Fig. 38.11 ABC convergence for endpoint acceleration

Table 38.4 Parameters and performance of endpoint acceleration

Controller	Link	PID parameters			MSE
		K_p	K_i	K_d	
PID	1	4	9	1	1.708×10^{-6}
	2	5	1	2	8.469×10^{-6}
PID-ABC	1	30.03	56.07	88.95	7.9188×10^{-08}
	2	50.1	46.96	23.62	8.4332×10^{-08}

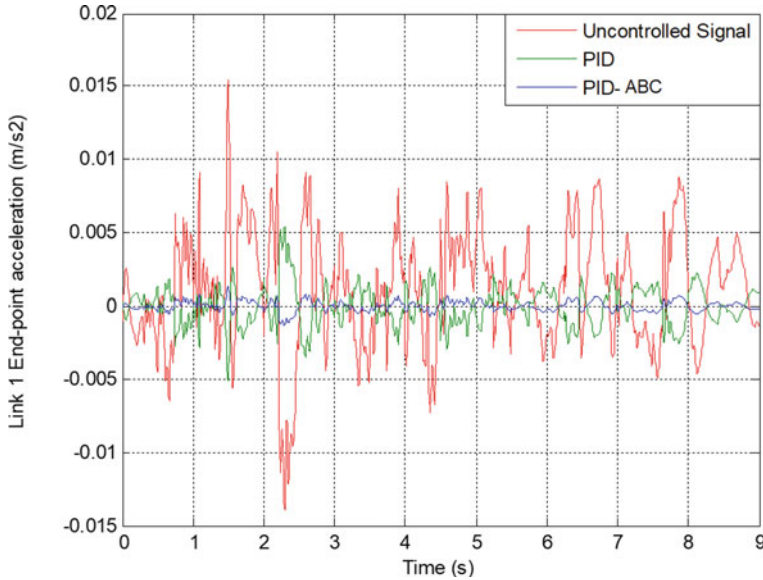


Fig. 38.12 Endpoint vibration suppression of link 1

shown in Figs. 38.12 and 38.13. The manual tuning of PID controller applied to the system undoubtedly aids to reduce the vibration in the system. However, the process is tedious and time-consuming. The vibration can be easily and additionally suppressed by utilizing the PID-ABC controller. This implies that, the ABC algorithm is very effective in optimizing the PID parameters.

38.6 Conclusion

This paper has presented the optimum PID controller using ABC for controlling TLFRM. The experimental test was carried out to obtain the input–output of the real system to characterize the dynamic behavior of TLFRM was first developed. Subsequently, TLFRM was modeled using NARX model structure in which predicted

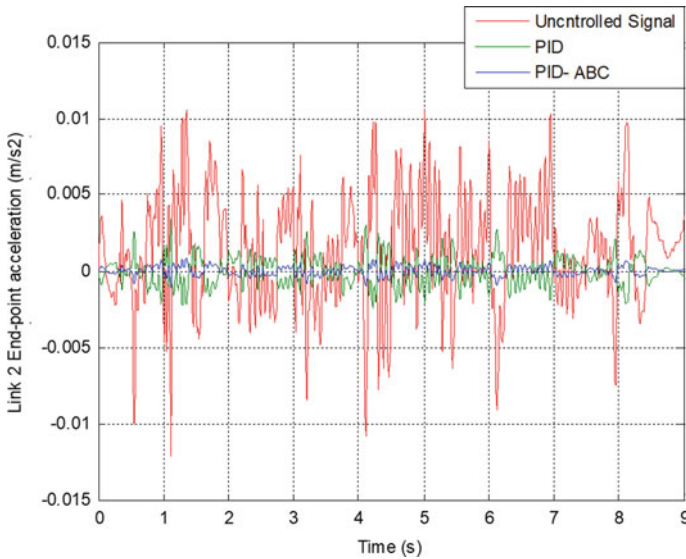


Fig. 38.13 Endpoint vibration suppression of link 2

by neural network. Then, hybrid PID controller is developed to control the hub motion and endpoint vibration suppression of each link, respectively. The optimum gains acquired through global search has been tested on the control structure. The system responses inclusive of input tracking and vibration suppression at the endpoint were evaluated. The results were compared to the heuristic methods. Though the simulation results portrayed that the manual tuning of PID controllers was able to control the system in terms of input tracking and reduce the vibration in the system, the process is tedious and time-consuming. On the other hand, the PID parameters tuned by ABC is easily obtained with less time. Besides, the results exhibit that the recommended controller is more effective to move the two-link flexible at lower overshoot with the improvement of 67 and 56% compared with the heuristic method and faster time that is from 5.1633 to 1.0783 s for hub 1 and from 2.1635 to 1.0499 s for hub 2. The vibration suppression shows 93.53% and 90.47% improvement, respectively.

Acknowledgements The authors would like to express credits to Universiti Malaysia Sarawak (UNIMAS) and Universiti Teknologi Malaysia (UTM) for financing and offering facilities to carry out this research.

References

1. Peza-Solís, J.F., Silva-Navarro, G., Castro-Linares, R.: Control of a rigid-flexible two-link robot using passivity-based and strain-feedback approaches. In: International Conference on Electrical Engineering, Computing Science and Automatic Control, pp. 476–481 (2010)
2. Mahamood, R.M., Pedro, J.O.: Hybrid PD/PID controller design for two-link flexible manipulators. In: Proceedings of 2011 8th Asian Control Conference, pp. 1358–1363 (2011)

3. Chu, M., Jia, Q.X., Sun, H.X.: Global terminal sliding mode robust control for trajectory tracking and vibration suppression of two-link flexible space manipulator. In: Proceedings of 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems, pp. 353–357 (2009)
4. Zebin, T., Alam, M.S.: Dynamic modeling and fuzzy logic control of a two-link flexible manipulator using genetic optimization techniques. In: Proceedings of 13th International Conference on Computer and Information Technology, pp. 418–423 (2010)
5. Lingbo, Z., Funchun, S., Zengqi, S.: Decoupling control of the two-link flexible manipulator. In: Multiconference on Computational Engineering in Systems Applications, pp. 2045–2049 (2006)
6. Raouf, F., Mohamad, S., Maarouf, S., Maamar, B.: Distributed adaptive control strategy for flexible link manipulators. *Robotica* **35**(7), 1562–1584 (2017)
7. Leena, G., Ray, G.: A set of decentralized PID controllers for an n-link robot manipulator. *Sadhana Acad. Proc. Eng. Sci.* **37**, 405–423 (2012)
8. Guangzheng, P.G.P., Xuesong, W.X.W., Yang, X.Y.X.: Study on fuzzy PD control of planar two-link flexible manipulator. In: 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering. TENCOM'02, vol. 3, pp. 1542–1545 (2002)
9. Green, A., Sasiadek, J.Z.: Dynamics and trajectory tracking control of a two-link robot manipulator. *J. Vib. Control*, 1415–1440 (2004)
10. Nguyen, V.B., Morris, A.S.: Using a genetic algorithm to fully optimise a fuzzy logic controller for a two-link-flexible robot arm. *Robotica* (2009)
11. Lee, T.H., Ge, S.S., Wang, Z.P.: Adaptive robust controller design for multi-link flexible robots. *Mechatronics* **11**, 951–967 (2001)
12. Rigatos, G.G.: A robust nonlinear control approach for flexible-link robots using Kalman filtering. *Cybern. Phys.* **1**(2), 134–143 (2012)
13. Shin, H.-C., Choi, S.-B.: Position control of a two-link flexible manipulator featuring piezoelectric actuators and sensors. *Mechatronics* **11**, 707–729 (2001)
14. Zhang, L., Liu, S.: Basis function based adaptive iterative learning control for non-minimum phase systems. In: World Congress on Intelligent Control and Automation, pp. 828–833 (2014)
15. Vakil, M., Fotouhi, R., Nikiforuk, P.N.: Maneuver control of the multilink flexible manipulators. *Int. J. Non-Linear Mech.* **44**, 831–845 (2009)
16. Sawada, M., Itamiya, K.: A position control of 2 DOF flexible link robot arms based on computed Torque method. *IEEJ Trans. Electron. Inf. Syst.*, 547–552 (2013)
17. Dogan, M., Istefanopulos, Y.: Optimal nonlinear controller design for flexible robot manipulators with adaptive internal model. *IET Control Theory Appl.* **1**(3), 770–778 (2007)
18. Zhang, X., Xu, W., Nair, S.S., Chellaboina, V.S.: PDE modeling and control of a flexible two-link manipulator. *IEEE Trans. Control Syst. Technol.* **13**(2), 3796–3801 (2005)
19. Mahamood, R.M.: Direct adaptive hybrid PD-PID controller for two-link flexible robotic manipulator. In: Proceedings of World Congress on Engineering and Computer Science, vol. 2 (2012)
20. Mahamood, R.M., Pedro, J.O.: Hybrid PD-PID with iterative learning control for two-link flexible manipulator. In: Proceedings of World Congress on Engineering and Computer Science, vol. 2 (2011)
21. Khairudin, M., Mohamed, Z., Husain, A.R.: System Identification and LMI based robust PID control of a two-link flexible manipulator. *Telkomnika* **12**(4), 829–838 (2014)
22. Khairudin, M., Arifin, F.: NN robust based-PID control of a two-link flexible robot manipulator. *Int. J. Adv. Sci. Eng. Inf. Technol.* **2**(1) (2012)
23. Alam, M.S., Md Zain, M.Z., Tokhi, M.O., Aldebraz, F.: Design of hybrid learning control for flexible manipulators: a multi-objective optimisation approach. In: Proceedings of the 8th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines, CLAWAR 2005, pp. 599–606
24. Tijani, I.B., Akmeliawati, R., Muthalif, A.G.A., Legowo, A.: Optimization of PID controller for flexible link system using a pareto-based multi-objective differential (PMODE) evolution. In: 2011 4th International Conference on Mechatronics (2011)

25. Supriyono, H., Tokhi, M.O., Md Zain, B.A.: Control of a single-link flexible manipulator using improved bacterial foraging algorithm. In: ICOS 2010—2010 IEEE Conference on Open Systems, pp. 68–73 (2010)
26. Fahmy, A.A., Kalyoncu, M., Castellani, M.: Automatic design of control systems for robot manipulators using the bees algorithm. *J. Syst. Control Eng.* **1**, 1–12 (2011)
27. Yatim, H., Mat Darus, I.Z.: Self-tuning active vibration controller using particle swarm optimization for flexible manipulator system. *WSEAS Trans. Syst. Control* **9**, 55–66 (2014)
28. Annisa, J., Mat Darus, I.Z., Tokhi, M.O., Mohd Samin, P.P.: Intelligent modeling of double link flexible robotic manipulator using artificial neural network. *J. Vibroeng.* **20**(3) (2018). ISSN Print 1392-8716

Chapter 39

MG-CMF: A Multi-granularity Capture Matching Features Model for Text Matching



Liang Jin and Xiaopeng Cao

Abstract There is a problem of lacking attention to the matching information between texts in text matching. In order to improve the accuracy, we propose multi-granularity capture matching features (MG-CMF) model to capture matching features from multiple granularities. The model uses convolution operations to construct the representation of text under multiple granularities, uses max pooling operations to filter more reasonable text representations, and builds matching matrices at different granularity. We use a convolution neural network (CNN) to capture matching features at the different granularity and input the captured matching features into the fully connected neural network to obtain the matching similarity. By making some experiments, we can get that the accuracy and F_1 values of the optimal experimental results on MSRP corpus are 73.60 and 82.20%, the mean average precision (MAP) and mean reciprocal rank (MRR) of the optimal experimental results on the WIKIQA corpus are 0.6821 and 0.6924. We draw a conclusion that the MG-CMF model is better than that of the other text matching models on accuracy.

39.1 Introduction

Natural language processing is a cross-cutting field of computer science, artificial intelligence, and linguistics. The goal is to enable computers to understand natural language and perform tasks such as language translation and question answering instead of humans. Natural language processing is inspired by deep neural networks [1]. Neural networks are widely used in sentence modeling. Deep learning models can represent sentences as vector matrices in semantic space and more accurately describe two sentences by using the distance between vectors. CNN is good at extracting abstract features from sentences, and recurrent neural network (RNN) is good at maintaining and utilizing long-distance information [2].

Text matching research is a key problem in natural language processing. For paraphrase identification (PI) task, it can be attributed to whether a sentence pairs

L. Jin · X. Cao (✉)

Xi'an University of Posts and Telecommunications, 710121 Xi'an, Shanxi, China
e-mail: cxp2760@163.com

© Springer Nature Singapore Pte Ltd. 2021

S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation, Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_39

549

or a text pair matches. For answer selection (AS) task, it can be attributed to the matching of the question and the candidate's answer [3]. Natural language sentences have complex structures, both sequential and hierarchical, which are essential for understanding natural language sentences. Therefore, a successful sentence matching algorithm not only needs to capture the internal structure of the sentence but also captures the rich interaction patterns between sentences. Taking the task of PI for example, given the following two texts:

T_1 : *A woman is slicing a tomato.*

T_2 : *A woman is eating a tomato.*

We can see that T_1 and T_2 have different semantics. We divide the sentences into different granularities (word, phrase, ..., sentence) to analyze T_1 and T_2 . For word granularity, there are many identical words in T_1 and T_2 , only "slicing" in T_1 and "eating" in T_2 are different, thus causing high similarity of word granularity matching interactions and affecting the final result of matching. The granularity of phrase is higher than that of the word, and phrases are composed of adjacent words. The features of low similarity between T_1 and T_2 are obviously more than word granularity in phrase granularity, such as "woman is slicing" in T_1 and "woman is eating" in T_2 . From the word, granularity to form a higher level of granularity finally reaches the sentence granularity. We need to consider the matching similarity between sentences from each granularity. For those problems, we propose the MG-CMF model for text matching.

The MG-CMF model constructs the text expression of sentences under multiple granularities and then constructs the similarity matching matrix under each granularity. We capture the important matching features under each granularity through CNN and calculate the text matching degree by using these matching features.

39.2 Related Works

With the rapid development of deep learning in recent years, much recent research uses neural network modeling to solve the text matching problem.

In 2014, Kim [4] uses CNN to model sentence representation. Hu [5] proposes ARC-I and ARC-II model by referring to Kim's idea in the same year. The ARC-I model fills the length of two sentences to the same length and constructs two vector matrices of the same size. Then, the two vector matrices, respectively, use convolution pooling to extract features and input these features into a fully connected neural network model to obtain the final matching score. The ARC-I model is simple in structure and fast in matching speed, but the ARC-I model only focuses on the extraction of low-level information, compresses the whole sentence into a low-dimensional vector, and loses the description of the detail matching. The ARC-II model lets two sentences interact before their respective high-level representations mature and retains the individual development space of each sentence. But the ARC-II model is fuzzy in the definition of fine-grained matching. It is defined by a weighted average

of two fine-grained representations. This definition does not reflect the degree of matching on fine grain, so the overall performance is not very good.

In 2014, Kalchbrenner [6] proposes a DCNN model. The bottom layer gradually transmits upward by combining adjacent word information, and the upper layer combines new phrase information, thus forming a semantic relationship between sentences. Based on Kalchbrenner's idea, Yin [7] proposes a MultiGranCNN model in 2015. Yin uses convolution neural networks to extract semantic information at different levels and splices different levels of semantic information together to obtain the final matching value. By extracting the text information between different levels, the MultiGranCNN model preserves the detailed information of the sentence better and improves the accuracy of text matching effectively. But the interaction between the sentence information is insufficient, and the matching information between the sentences causes a certain degree loss.

In 2016, Pang [8] presents MatchPyramid model on the basis of the ARC-II model and redefines the interaction between two sentences. The matching matrix defined by the MatchPyramid model is based on the degree of matching between words in two sentences with the finest granularity. The model uses the XNOR relation between the word vectors of two words, cosine similarity or dot product to define the similarity between words. According to the spatial position of the word in the sentence, the similarity between the two words can be calculated, and a two-dimensional matching matrix can be constructed. MatchPyramid model shows a good effect on text matching, but there is a lack of matching information between phrases when adjacent words are combined to form the phrase.

In 2018, Kim [9] proposes a densely connected co-attentive RNN on the basis of DenseNet [10]. Each layer of the model uses concatenated information of attentive features as well as hidden features of all the preceding recurrent layers. The model also uses an autoencoder after dense concatenation.

In 2019, Lai presents a model that uses bidirectional long short-term memory (Bi-LSTM) with pairwise comparisons and attention-pooling [11]. By using Bi-LSTM to capture the interaction among the first sentence words conditioned on the second sentence after the soft alignment attention layer and adopts three pooling mechanisms to extract features.

39.3 MG-CMF Model

The structure of the MG-CMF model is shown in Fig. 39.1.

1. Text preprocessing: The MG-CMF model preprocesses text, converts the text into a word vector matrix, and extracts the feature vector from the word vector matrix.
2. Feature extraction: The MG-CMF model obtains the text representations at multiple granularities through the continuous feature extraction operation and constructs matching matrices of each granularity.

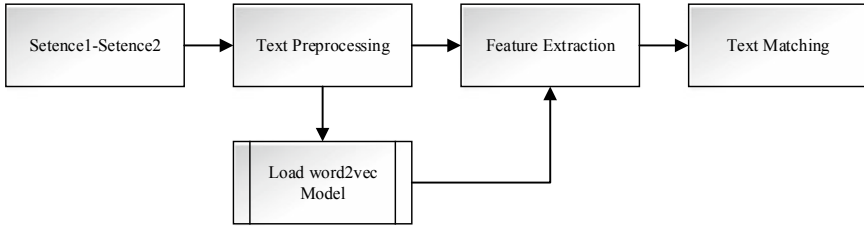


Fig. 39.1 MG-CMF model structure

- 3. Text matching: The MG-CMF model captures the matching features between the texts at multiple granularities and finally inputs the captured features into a fully connected neural network to obtain a final matching degree.

39.3.1 Text Preprocessing

In this paper, we use the pretrained Word2vec [12] model to transform the text into a word vector matrix. Word2vec is a software tool for training word vectors created by Tomas Mikolov’s research team at Google. By using a large number of corpus training, the vector representation of words is obtained.

Firstly, we divide the text into separate words. Then, we remove the meaningless stop words in the text and retain the key text information. Finally, we need to count the maximum length of all sentences in the corpus to ensure all sentences are filled to the same length during embedding.

39.3.2 Feature Extraction

The traditional feature TF-IDF [13] mainly extracts features from the perspective of word and sentence structure. TF-IDF lacks the semantic relationship between the two sentences. Therefore, we build the semantic relationship between text pairs by using the word vector technique and the deep matching model.

The feature extraction structure is composed of the embedding layer and several convolution layers (see Fig. 39.2). This paper uses the Word2vec model that has been pretrained. The trained word vector model has a dimension of 300. We load the Word2vec model when the MG-CMF model starts to train and use the word vector to construct the word vector matrix. In the embedding layer, the sentence consists of n words such as v_1, v_2, \dots, v_n and each word passes through the word2vec model to obtain vectors $\omega_1, \omega_2, \dots, \omega_n$. We obtain the vector matrix A by splicing these vectors. The vector matrix A represents the representation of a sentence transformed into a vector matrix by loading the word2vec model. The row of the matrix represents

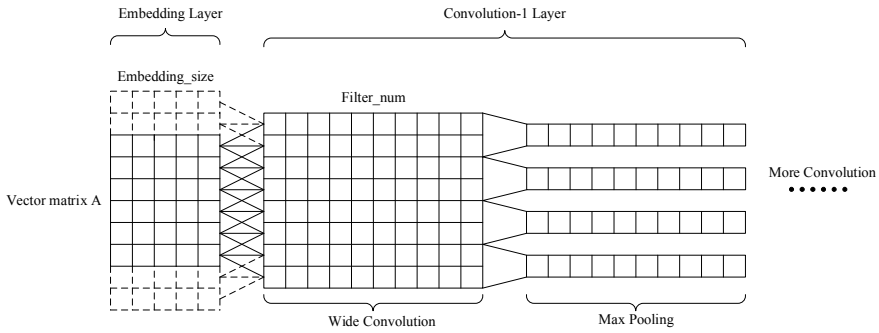


Fig. 39.2 Model feature extraction structure

the length of sentences, and the column of the matrix represents the embedding dimension.

Each of the convolution layers consists of convolution and max pooling. In this paper, the convolution layer uses wide convolution [14]. The size of the convolution kernel W is $w \times \text{embedding_size}$, the number of the convolution kernel is filter_num , and the convolution formula is as follows.

$$Z_i = \sigma(W \cdot c_i + b) \tag{39.1}$$

where σ is the activation function and b is the bias. The activation function uses the hyperbolic tangent function, which is better than other activation functions for text matching. The bias is initialized to a constant of 10^{-4} .

The implementation of wide convolution is to zero padding both above and below the dimension of the sentence length before convolution. The filling size is $w - 1$. The rest of the operation is consistent with convolution.

In this paper, we use max pooling after each wide convolution and have two effects for the MG-CMF model: (i) Max pooling reduces the dimension of the network, reduces the calculation amount, and accelerates the training speed of the network; (ii) max pooling filters out some unreliable and invalid phrases.

A sentence uses a wide convolution of a sliding window of fixed length 3 to combine words and words (see Fig. 39.3). Different feature maps provide different combinations of words and words. Each phrase is a vector. Max pooling selects the phrase vector with higher confidence after wide convolution. The gray section in Fig. 39.3 represents a lower confidence phrase. After max pooling, we discard lower confidence phrases and leave more confidence phrases as the input of the next layer of convolution.

After the first convolution, the information of the sentence is from the level of word to the level of the phrase. After several convolutions, it reaches the level of the whole sentence. Therefore, we obtain the representation of a sentence at different granularity through the feature extraction structure.

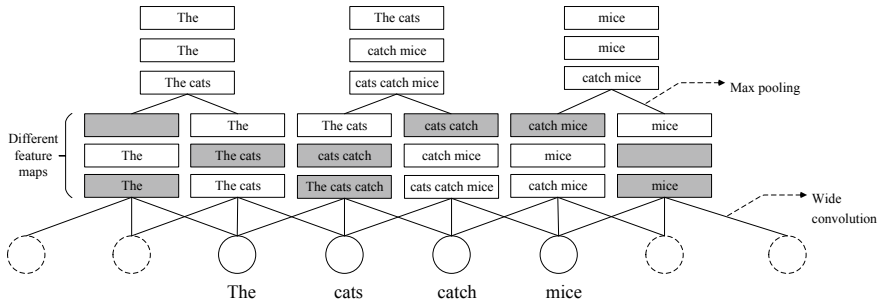


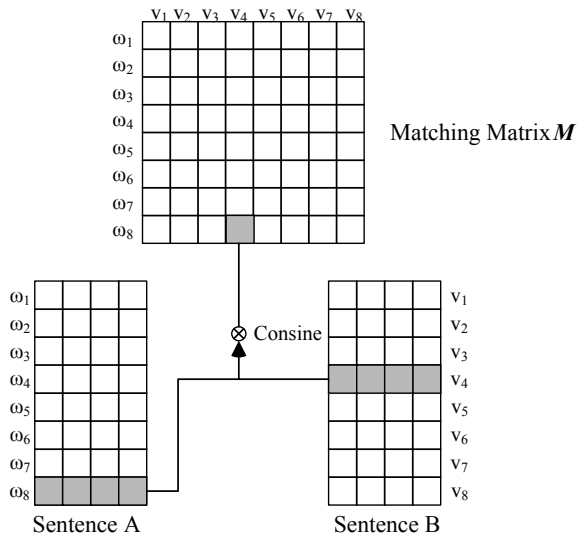
Fig. 39.3 Wide convolution and max pooling

39.3.3 Text Matching

The essential problem of matching two texts is that calculate the semantic similarity between two texts. The MG-CMF model is different from the deep learning model based on single semantic document representation and multi-semantic document representation. The MG-CMF model uses the feature extraction module to get representations of a sentence in different granularity, establishes matching matrices on multiple granularities, and shifts the attention from how to better represent text information with vectors to the capture matching feature.

In this paper, we obtain multiple granularity representations of sentences from the feature extraction module and establish a matching matrix M in each granularity (see Fig. 39.4). By calculating the similarity between two vectors, the MG-CMF model

Fig. 39.4 Matching matrix



constructs a two-dimensional matching matrix. Each node in the matching matrix has a number M_{ij}^L which represents the degree of matching between text pairs.

M_{ij}^L represents the interaction between the i th segment information in sentence A and the j th segment information in sentence B under granularity L , often using cosine similarity and the dot product to measure the similarity between two vectors.

$$M_{ij}^L = \frac{\alpha_{ij}^{LT} \beta_j^L}{\alpha_i^L \cdot \beta_j^L}, \quad (39.2)$$

$$M_{ij}^L = \alpha_i^{LT} \beta_j^L \quad (39.3)$$

Using different similarity calculation methods has a slight effect on the experimental results. This paper uses cosine similarity to calculate the similarity of text matching.

In order to keep the length of the two sentences in the embedded layer consistent, we use zero padding for sentences with insufficient length. When calculating the cosine similarity, the loss of the entire model is NAN due to the filling parts α_i and β_j are zero, so we give a minimum value c . The calculation formula of cosine similarity in this paper is as follows.

$$M_{ij}^L = \frac{\alpha_i^{LT} \beta_j^L}{\alpha_i^L + c) \cdot (\beta_j^L + c)} \quad (39.4)$$

Assuming that the module of the convolution layer of feature extraction has k layer, the matching matrix M^0, M^1, \dots, M^k which represents the matching matrix under different granularity will be obtained.

The MG-CMF model then uses the convolution neural network to capture the matching features of the text in image recognition and finally gets the matching features at different granularity (see Fig. 39.5). The MG-CMF model inputs the matching features from different granularity into the fully connected neural network to obtain the final matching degree (see Fig. 39.6).

39.4 Experiment and Analysis

We make some experiments based on PI and AS task to validate the superiority of the model over other models.

For all tasks, words are embedded by the pretrained 300-dimensional word2vec model, which is not changed during training. The training uses the Adagrad [15] optimizer and uses L_2 regularization to tune the model.

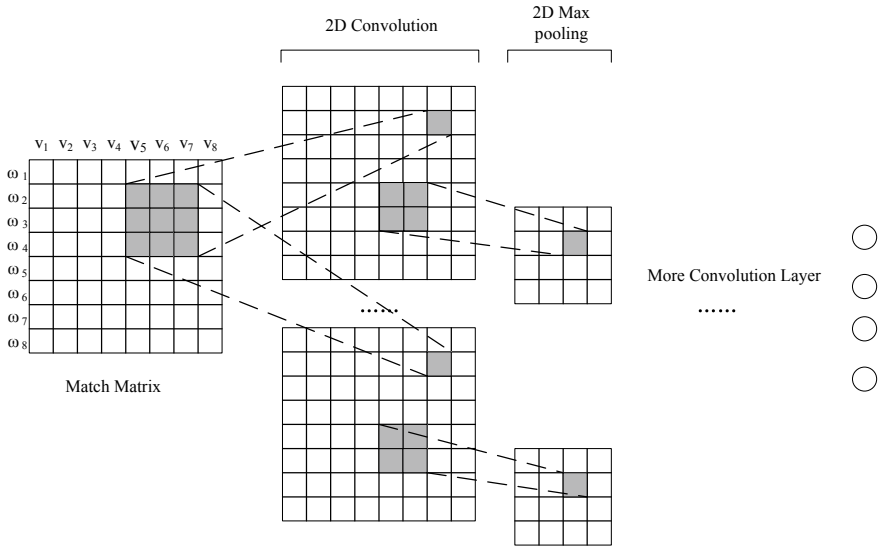


Fig. 39.5 Capture matching feature

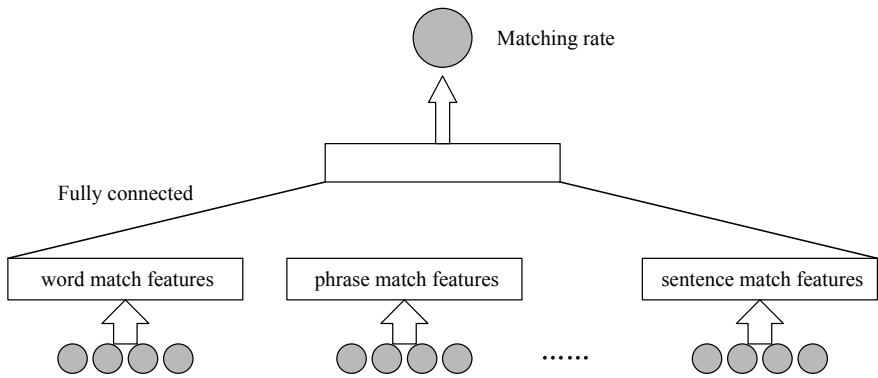


Fig. 39.6 Calculation of matching similarity by fully connected neural network

39.4.1 Paraphrase Identification Task

The purpose of the PI task is to determine whether another sentence has the same meaning. This paper uses Microsoft Research Paraphrase Corpus (MSRP) [16]. MSRP is a classic public corpus for this task. The MSRP corpus contains a total of 5801 pairs of texts. A pair of matching text has a label of 1, otherwise 0. The training set contains 2753 pairs of positive examples and 1323 pairs of false examples. The test sets contain 1147 pairs of positive examples and 578 pairs of false examples. The evaluation of the MSRP corpus is based on a classification problem,

and the evaluation indicators are the accuracy rate and the F_1 score. Accuracy rate refers to the proportion of correctly classified text pairs to the total. F_1 score refers to the geometric mean of accuracy and recall rate for matching categories (label 1). The F_1 formula is as follows.

$$F_1 = \frac{2PR}{P + R} \quad (39.5)$$

where P is the accuracy and R is the recall rate.

We compare our model with representative text matching approaches: (i) TF-IDF [8]; (ii) DSSM/CDSSM [17]; (iii) ARC-I/ARC-II [5]; and (iv) MultiGranCNN [7]. These experiments use the same MSRP corpus. (See Table 39.1).

It can be seen from Table 39.1, model 2 achieves a high accuracy of about 70%. In this paper, we use different granularity to observe the effect of the model. The experimental results show that the MG-CMF model works best when we use two convolution layers. When using three convolution layers, the accuracy and F_1 score of the model decline. The reason for the decline is that the MSRP corpus is small and the complexity of the model increases, which makes the model easy to over-fit. The MG-CMF model accuracy is higher than that of the traditional feature TF-IDF model 3.29%, and the F_1 score is higher than 4.58%. It is shown that the capture of matching feature information from multiple granularities is effective for text matching task.

Comparing other deep learning models, observe the experimental results of the models 3, 4, 5, 6 in the table. The accuracy is higher than baseline, but it is slightly lower than the traditional feature model TF-IDF. Model 7 in the table is better than models 3, 4, 5, 6, and the reason is that MultiGranCNN extracts the information of text from multiple granularities, enriching the details of the text.

The best experimental results of PI task are uRAE [18] model and the Multi-GranCNN model at present. The accuracy is 76.80% and 78.10%, respectively, and

Table 39.1 Results on MSRP

Model		Acc (%)	F_1 (%)
1. Baseline		66.50	79.90
2. TF-IDF		70.31	77.62
3. DSSM		70.09	80.96
4. CDSSM		69.80	80.42
5. ARC-I		69.60	80.27
6. ARC-II		69.90	80.91
7. MultiGranCNN without un-pretrained		72.50	81.40
8. MG-CMF	None-conv	69.79	80.52
	One-conv	72.23	81.42
	Two-conv	73.60	82.20
	Three-conv	71.96	81.26

the F_1 score is 83.60% and 84.40%, respectively. The main reason is that these models are pretrained on a larger corpus. The accuracy of the experimental results is 1.1% higher than that of the MultiGranCNN model, and the F_1 score is higher than 0.8% under the same corpus.

39.4.2 Answer Selection Task

To further validate the validity of the model, we apply the MG-CMF model to the AS task, we use the WIKIQA [19] corpus. The WIKIQA corpus includes 3047 questions and 29,258 sentences, of which 1473 sentences are marked as the answer sentence of the corresponding question. The AS task is to sort the relevance of the candidate answer to the question, and the evaluation indicator is mean average precision (MAP) and mean reciprocal rank (MRR).

For the three kinds of deformation of the MG-CMF model, this paper performed an experiment on the WIKIQA corpus. These three kinds of deformation are obtained by changing the granularity of the MG-CMF model under the same experimental conditions (see Fig. 39.7). The MG-CMF@1 model uses only the interaction of word granularity; the MG-CMF@2 model uses the interaction of word and short phrase granularity; and the MG-CMF@3 model uses the interaction of word, short phrase, and long phrase granularity. The diagram shows that the capture of matching information on multiple granularities is an effective method of text matching.

The results of the experiment list the results of the other models on AS task and compare them with the best experimental results in this paper (see Table 39.2). Some of the experimental results are from the published paper [2, 20]. The experimental

Fig. 39.7 Comparison of several patterns of MG-CMF Model

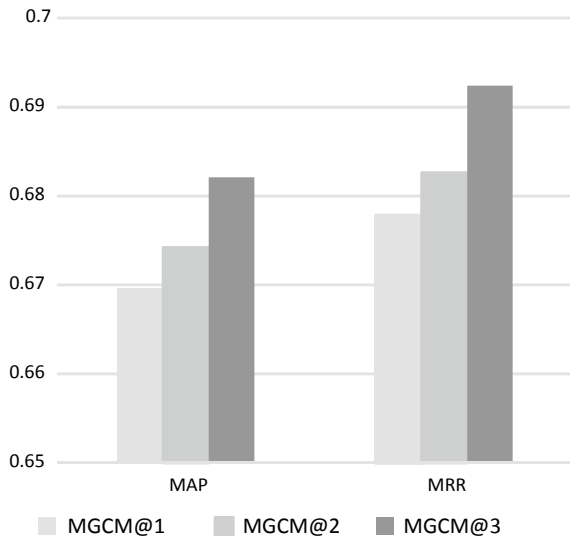


Table 39.2 Results on WikiQA

Model	MAP	MRR
WordCnt	0.4891	0.4924
WgtWordCnt	0.5099	0.5132
CNN-Cnt	0.6520	0.6652
Addition	0.5021	0.5069
Addition (+)	0.5888	0.5929
A-LSTM	0.5347	0.5483
A-LSTM (+)	0.6381	0.6537
MultiGranCNN	0.6629	0.6813
MG-CMF	0.6821	0.6924

results show that the MG-CMF model has a better computational effect than other models. Compared with MultiGranCNN, MAP and MRR are increased by 2.90% and 2.55%, respectively. The results show that the MG-CMF model has better semantic matching ability than the MultiGranCNN model, which verifies the effectiveness of the MG-CMF model.

39.5 Conclusion

In this paper, we transfer the attention from the text representation to capture text matching information. We capture the matching information at multiple granularities, which further enriches the details matching information of the text and reduces the loss of the matching information.

The experiments on PI and AS tasks show that capturing matching features at multiple granularities is better than extracting text information at multiple granularities.

References

1. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
2. Yin, W., Schütze, H., Xiang, B., et al.: ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs. *Computer Science* (2015)
3. Liang, P., Yan-Yan, L., Jun, X.U., et al.: A survey on deep text matching. *Chin. J. Comput.* **40**(04), 985–1003 (2017)
4. Kim, Y.: Convolutional neural networks for sentence classification. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1746–1751. Doha, Qatar (2014)
5. Hu, B., Lu, Z., Li, H., et al.: Convolutional neural network architectures for matching natural language sentences. In: *Proceedings of the 27th international Conference on Neural Information Processing Systems*, pp. 2042–2050. MIT Press, Cambridge (2014)

6. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolution neural network for modelling sentences. In: Proceedings of the Association for Computational Linguistics, pp. 655–665. Baltimore, Maryland (2014)
7. Yin, W., Schütze, T., Hinrich.: MultiGranCNN: an architecture for general matching of text chunks on multiple levels of granularity. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, pp. 63–73. Beijing, China (2015)
8. Pang, L., Lan, Y., Guo, J., et al.: Text matching as image recognition. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, pp. 2793–2799. Phoenix, USA (2016)
9. Kim, S., Kang, I., Kwak, N.: Semantic Sentence Matching with Densely-connected Recurrent and Co-attentive Information. (2018)
10. Huang, G., Liu, Z., Weinberger, K.Q., et al.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
11. Lai, H., Tao, Y., Wang, C., et al.: Bi-directional Attention Comparison for Semantic Sentence Matching. *Multimedia Tools and Applications* (2019)
12. Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed representations of words and phrases and their compositionality. *Adv. Neural. Inf. Process. Syst.* **26**, 3111–3119 (2013)
13. Salton, G.: *The SMART Retrieval System-Experiments in Automatic Document Processing*. Prentice Hall Inc., Englewood Cliffs (1971)
14. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vector space. *arXiv preprint arXiv: 1301. 3781* (2013)
15. Duchi, J.C., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12**, 2121–2159 (2011)
16. Dolan, B.: Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING). Association for Computational Linguistics (2004)
17. Huang, P.S., He, X., Gao, J., et al.: Learning deep structured semantic models for web search using click through data. In: Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management, pp. 2333–2338. Amazon, India (2013)
18. Socher, R., Huang, E.H., Pennington, J., et al.: Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Adv. Neural. Inf. Process. Syst.* **24**, 801–809 (2011)
19. Abbas, F., Malik, M.K., Rashid, M.U., et al.: WikiQA —A question answering system on Wikipedia using freebase, DBpedia and Infobox. In: Sixth International Conference on Innovative Computing Technology (INTECH). IEEE (2016)
20. Yi, Y., Wen-tau, Y., Christopher, M.: Wikiqa: a challenge dataset for open-domain question answering. In: Proceedings of EMNLP, pp. 2013–2018 (2015)

Chapter 40

Detecting Domain Name System Tunneling and Exfiltration from Domain Name System Traffic



Yi-Chung Tseng, Ming-Kung Sun, and Wei-An Chen

Abstract In this study, we present a framework to detect a Domain Name System (DNS) tunnel and DNS exfiltration network traffic by using both unsupervised and supervised learning algorithms. In general, considerable time is required to learn the structure of the data before clustering when using an unsupervised learning algorithm. Therefore, in this study, we leveraged the power of mathematical algorithms for calculating the optimal number of clusters and reducing the time required for understanding the data structure. Conversely, we used a supervised learning method to learn the data leakage behavior for detecting DNS exfiltration traffic. We used an open-source tool to generate testing data, and the experimental result proved the robustness of the proposed framework.

40.1 Introduction

The DNS [1] is an Internet service. A domain name is a characteristic structure that is easier to understand and remember than an Internet Protocol (IP) address. The DNS acts as a decentralized database that maps a domain name to an IP address. In addition to being used for providing Web services, Web hosts, and other online services, domain names are often used by attackers to steal personal information. For example, an attacker can attach personal information encryption to the domain name itself, such as “ojswczdnmuxg2zd4ge3q.malware.com,” through the Base64 encryption algorithm. The aforementioned domain name can be split into three blocks. The first block is “ox-wczdnmuxg2zd4ge3q;” the second block is “.malware;” and the third block is “.com.”

Y.-C. Tseng (✉) · M.-K. Sun · W.-A. Chen
Acer Cyber Security Incorporation, Taipei, Taiwan
e-mail: Bruce.Tseng@acercsi.com

M.-K. Sun
e-mail: Morgan.Sun@acercsi.com

W.-A. Chen
e-mail: Wayne.Chen@acercsi.com

The string in the first block represents the personal data encryption function hidden in the domain name itself by the attacker through an encryption algorithm. The second and third blocks specify the actual DNS server of the attacker. When an attacker desires to steal information, such as personal data which can be encrypted into the subdomain name, consequently, the domain name along with the target data is transmitted from the client server to an external DNS server through the DNS recursive technology, thus causing data leakage.

About the recursive DNS technique used for data leakage, a client sends a query to the local DNS server. If the server does not respond, the local DNS server continues to query the higher layers of the DNS server until the IP with the answer is found. The process is initiated from the root directory. For example, for www.google.com.tw, the “.tw” query is executed first. Then, “.com” is executed from the top level to the second level and so on.

Another method of conducting data leakage is by establishing a DNS tunnel. A hacker can use the tunnel to transmit the target data. This study proposes an architecture that can detect DNS tunneling and DNS exfiltration network traffic by leveraging the power of a modern machine learning algorithm. The experimental results indicate the simplicity, robustness, and scalability of the proposed approach.

The remainder of this paper is organized as follows. Section 2 introduces the relevant literature related to DNS tunneling and exfiltration. Section 3 presents an overview and the framework of the approach. Section 4 demonstrates the experimental results. Section 5 summarizes the conclusions of this study.

40.2 Related Work

40.2.1 DNS Tunnel

Establishment of a DNS Tunnel Anirban et al. [2] mentioned that when a client tries to establish a DNS tunnel with an external DNS server, the query that contains the TXT record is sent, which is a type of resource record in DNS. Once the server receives the request, it returns a response with the TXT records to the client. Thus, a DNS tunnel is established successfully. A malicious server can take advantage of this tunnel to establish a tunnel for starting a session or executing an instruction.

Detection of a DNS Tunnel Anirban et al. [2] used the *k*-means clustering algorithm to find a DNS tunnel. Binsalleeh et al. [3] characterize the malicious payload distribution tunnel in DNS. They proposed solution characterizes these tunnels based on the DNS query and response messages patterns. Farnham and Atlasis [4] presented an overview of the history and techniques used for DNS tunneling detection. Compared with the regular A or AAAA DNS queries that have a constant size range, tunneling traffic tends to have a considerably larger size range. Paxson et al. [5] use the implementation of Kolmogorov complexity to detect DNS tunnel. Dietrich et al. [6] used various features to cluster DNS traffic by using *k*-means clustering

with $k = 2$. Born [7] proposed a method to detect DNS tunnels by using meaningful words. Hind [8] proposed a neural network for detecting DNS tunnel but did not provide any information on the data sources, features used, model selected, or model performance.

40.2.2 DNS Exfiltration

Situational explanation of DNS Exfiltration Anirban et al. [2] mentioned that assumed a scenario in which a computer is infected by a malicious program that wants to steal and transmit information to the attacker server. In the first step, the malware encrypts the private data. In the second step, the encrypted data is attached to the attackers' domain. For example, if aGVs13IGb3Vu is the encrypted data and malware attaches this data to malware.com, the domain name aGVs13IGb3Vu.malware.com is obtained. In the third step, because the encoded domain is not part of the local cache, the domain is forwarded to the server of malware.com by using the DNS recursive technology. Once the attackers' DNS server receives the query, the attacker can extract the third-level domain and decode it. In the fourth step, the attacker can respond to the client, which appears benign. The following sections describe how DNS exfiltration can be detected using the proposed method.

Signature-based Detection of DNS Exfiltration Jawad et al. [9] mentioned traditional methods for detecting DNS exfiltration rely on signatures which are not sufficient. By registering a new domain name, an attacker can easily bypass the blacklist. Besides, the signature-based approach relies on rules checking and thresholds to trigger an alert and is struggling to discover the malware's pattern behavior. On the other hand, maintaining a blacklist is also inefficient.

Rule-based Detection of DNS Exfiltration Fawcett [10] described several encoding techniques for DNS exfiltration. These techniques rely on rule-based detection, such as detection according to the number of requests and responses, entropy of the hostname, percentage of numbers in the domain name, and number of non-existed domain.

Machine Learning-based Detection of DNS Exfiltration Anirban et al. [2] proposed machine learning models, by using logistic regression model to predict DNS exfiltration; they used eight features to describe the domain string and exfiltration domains as a negative set and benign domain as a positive set to train the model.

40.2.3 Unsupervised Learning

k -Means Clustering One of the most well-known unsupervised methods is the k -means clustering algorithm. The user randomly selects k points as the initial centroid,

where k is the user-specified parameter. Each point is then assigned to the cluster with the closest centroid. The centroid of each cluster is then updated by calculating the average of the data points for each cluster. The centroid is repeatedly assigned and updated until no improvement is obtained by changing the cluster or until the centroids are the same again.

Silhouette Method Rousseeuw [11] proposed the silhouette method, which is used for interpreting and verifying consistency within a data cluster. This technique provides a concise graphical representation of the classification of each object.

40.2.4 *Supervised Learning*

Extreme Gradient Boosting (XGBoost) XGBoost [12] is an optimized distributed gradient boosting library that is designed to be highly efficient, flexible, and portable. This library employs machine learning algorithms under the gradient boosting framework. XGBoost provides parallel tree boosting, which solves many data science problems rapidly and accurately.

40.3 Problem Formulation

We assumed a scenario in which an attacker hacks a client computer and installs a backdoor program. When attackers desire to send a command or steal personal information, they can use a DNS tunnel to establish a connection. Once a DNS tunnel is established, the encrypted private data is sent out through the DNS tunnel. The attackers can also use recursive DNS technology to send out the exfiltration data. Hence, we propose a detection method that leverages a machine learning algorithm to identify DNS tunneling and DNS exfiltration. Unsupervised and supervised learning are used in the proposed method. The proposed approach is introduced in the following section.

40.4 Overview of the Approach

The proposed architecture relies on analyzing DNS traffic and can identify DNS tunneling and DNS exfiltration. The proposed process is explained briefly in the following text. First, DNS traffic is collected. Second, feature engineering technology is used to extract features from the DNS query, including TXT and A records. Third, the silhouette method is used to calculate the optimal number of clusters with the k -means clustering algorithm by using the aforementioned features for determining

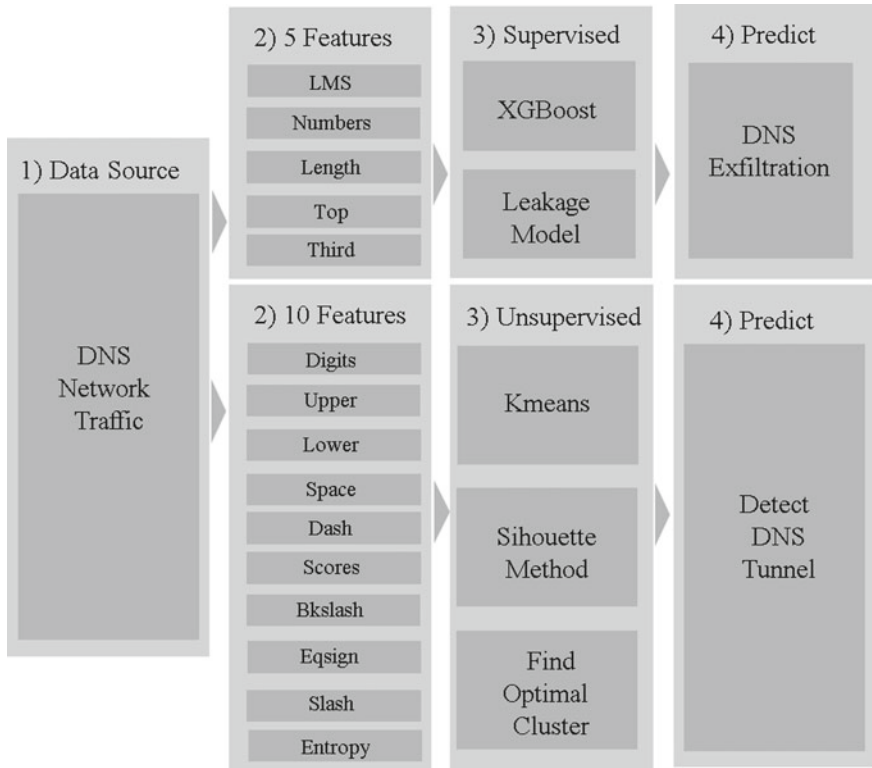


Fig. 40.1 Overview of the proposed approach for detecting DNS tunneling and DNS exfiltration

whether a DNS tunnel exists. We also trained a model that can identify exfiltration from a DNS query, as presented in Fig. 40.1.

40.4.1 DNS Query Collection

In the first step, the DNS network traffic is collected. Because there exist many types of data in DNS traffic, such as A, AAAA, PTR, and TXT, we only collect TXT, A, and AAAA records from DNS traffic.

40.4.2 Feature Engineering

In the second step, clustering is conducted on the TXT records by using feature engineering. Anirban et al. [2] used 12 features to describe the behavior of TXT

Table 40.1 Feature engineering for detecting DNS tunneling

Feature	Explanation
Number of digits	Calculate how many digits exist in the TXT record
Number of upper	Calculate how many upper cases exist in the TXT record
Number of lower	Calculate how many lower cases exist in the TXT record
Number of space	Calculate how many spaces exist in the TXT record
Number of dash	Calculate how many dash exist in the TXT records
Number of under line	Calculate how many under lines exist in TXT records
Number of slashes	Calculate how many slash exist in TXT records
Number of back-slash	Calculate how many back-slash exists in TXT records
Number of equal-sign	Calculate how many equal-signs exist in TXT records
Entropy	Calculate the Shannon entropy of TXT record

Table 40.2 Feature engineering for DNS exfiltration

Feature	Explanation
all_domain_length	Calculate the length of the entire domain name
third_domain_length	Calculate the length of the 3LD
domain_num_percentage	Calculate the proportion of numbers in the domain name
domain_LMS_percentage	Calculate the ratio of the longest and most meaningful string in the domain name to the overall string
top_domain_count	Calculate the same 1LD domain name as the total number of domain names

data. Therefore, in the present study, the 10 features presented in Fig. 40.1 were used. The meaning of each feature is provided in Table 40.1.

To train a model that can identify DNS exfiltration, feature engineering is required. We used five features to describe the behavior of a DNS query string. The meaning of each feature is specified in Table 40.2.

40.4.3 Determining the Optimal Number of Clusters that Can Find a DNS Tunnel

In the third step, an unsupervised algorithm is used to cluster the TXT records. Because the TXT records of DNS traffic differ with the company environment, the clustering value should be dynamically adjusted according to the DNS traffic. Therefore, the proposed work used the silhouette method to proactively find the optimal number of clusters. After clustering the data, we can observe whether each cluster has encoded strings. If there are encoded strings in a cluster, it can be concluded that there is a DNS tunnel in the DNS network.

40.4.4 *Training an XGBoost Model for Identifying DNS Exfiltration*

In the fourth step, we must train a model to identify DNS exfiltration by using a supervised learning algorithm. The preparation of a supervised model requires a labeled dataset. Therefore, we used a dataset that has only one field and contained a large number of leakage domains. This dataset [13] was provided by the Sydney University. Moreover, we used a credit card generator to generate a large number of credit card numbers; encrypt the card numbers through md5, base64, and other encryption algorithms; and then attach the encryption string into a domain. We also used a domain that was not leaked and was provided by Alexa [14]. The aforementioned datasets were integrated into a single dataset, including the leaked and benign domain datasets. Then, five features were used to characterize the behavior of the leaked and not leaked domains, as presented in Table 40.2. Finally, we used the XGBoost algorithm, which is a supervised learning model, to learn from the dataset for identifying DNS exfiltration.

40.5 Experiment

Detecting DNS tunneling by examining TXT records is a difficult problem primarily due to the high diversity of TXT records in real-world DNS traffic. Hence, we used a testing dataset obtained from an open source for detecting DNS tunneling. Moreover, we used the Data Exfiltration Toolkit (DET) [15] to generate a large number of DNS exfiltration samples for verifying whether our model could identify DNS exfiltration.

40.5.1 *DNS Tunnel Detection*

Data Collection The testing dataset [16] for DNS tunneling contained 1096 TXT records, which accounted for approximately 0.054308% of all DNS queries. Some TXT records were generated by incorporating DNScat [17] tunneling traffic, and the other records were regular DNS traffic.

***k*-Means Algorithm and Average Silhouette Method** We selected *k*-means clustering by using the aforementioned features to detect all the TXT queries that are encoded a string. The optimal number of clusters was calculated using the average silhouette method, as displayed in Fig. 40.2. In the figure, the *x*-axis represents the number of clusters and the *y*-axis represents the silhouette score. The higher the *y*-axis score, the better is the clustering result. In this study, the highest silhouette value was obtained by dividing the datasets into three clusters.

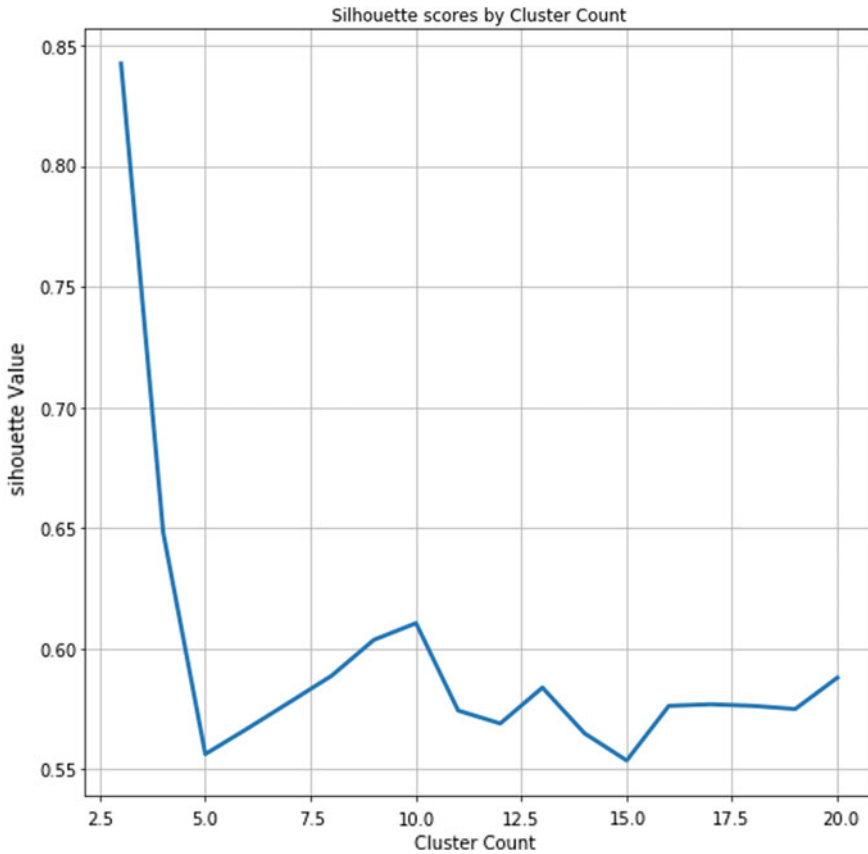


Fig. 40.2 Silhouette scores by cluster count

Cluster and TXT records As shown in Table 40.3, we decided that clusters 1 and 2 contained encoded TXT responses and that cluster 0 contained site verification traffic. Although clusters 1 and 2 provided the same type of encoded TXT record, the length of the encoded string in cluster 1 was longer than that in cluster 2. Moreover, a small number of encoded strings were classified into cluster 0. Overall, from the clustering results, we can quickly identify if a DNS tunnel exists.

40.5.2 DNS Exfiltration Detection

Data Collection In this experiment, we used a credit card generator [18] to generate a large number of credit card numbers, encrypted the card numbers through md5, base64, and other encryption algorithms; and then combined the data into a domain.

Table 40.3 Cluster of TXT records

Cluster	TXT records
0	254.229.168.192.spam.dnsbl.sorbs.net VERSION.BIND version.bind 251.229.168.192.spam.dnsbl.sorbs.net cf._dns-sd._udp.0.95.168.192.in-addr.arpa dnscat.690001a18087f230aaa1840010aaccff573
1	dnscat.4fc801a1803d607cb283df007ea280e2d578be0e25c47557d1f4db043faa.5d36c542b7b8248cf30b9c2c488b1669d80338f4849541 ad7d4f96ea5b76.5f4e43420d69c7ec8e7c2bdbc15068bc9117a4f69194f780c2a68b477d8f.e6fcd112c94374bc3c586ed675dde5cb7c058ad5bbe9fbc dnscat.c68801a180bd2c0a8262b000a2133c316f63766097374ef8fe820302d52e.4083f9058db6661c4c045f3e1a9d62830430c7 cf4b98e64b0fac3aabbd38e.e142207b5b81c6254cda8cc3343167e837e938d67dc45c2f8b26a2222c6.9c6ce247511c002050c3a7bfc11f9205ff12c9ed7b4d96dbe1a
2	dnscat.0ad001a1807f5322891563000c1038c3c7a525f6d975b68f3d73df039250.3db23e96eee3b7a4c2a261d88966c6491a731b3ff1f32fd1b3039f6270088.f5c730 dnscat.137903a180000000008187abbe71ed60866c8647ed9ede53317c6ca4bca9.9f8756ac08ba26d904ab5ddeb39f3ff291d5c 5e50ff3287dd350a7a6c4eb.58eff3d0c8b589dea47c563e1b

Table 40.4 Credit card number encryption

Credit card number	Encryption
4916658665745840	Mzc2MzgwNTA3NTY3NzAx5
30360128837301	30993683d51e835756d02f655af05ac
30368674657247	1e8299e0c7a1690ec3d6928f2a8366a
4916658665745840	275c15e507e168f5f71eb848cd56cfd3180c0a33
4556229341515026	275c15e507e168f5f71eb848ctq6cfd3180ec0a33

Table 40.4 lists the original credit card number and the corresponding base64 encryption data. Table 40.5 presents the domain name with the encrypted data. We combined 1,435,514 exfiltration domains as a negative set and 1,000,134 benign domains from Alexa as a positive set to train the model.

XGBoost Our feature space had five dimensions, as listed in Table 40.2. We selected the XGBoost algorithm because this algorithm is easy to deploy, effective, and exhibits superior performance. We used Scikit-learn [19] to deploy XGBoost and to validate and test our proposed method.

Self-verification For testing the effectiveness of the XGBoost model, we validated the metric `Roc_auc`, which is a performance measure, at various threshold settings (Table 40.6). Roc is the probability curve, and auc represents the degree or measure of separability. The precision score is obtained by calculating the ratio of all “correctly retrieved results (TP)” to all “actually retrieved (TP + FP).”

Effectiveness of the Model To demonstrate the effectiveness of the proposed model, we used the DET to generate a leakage domain along with the regular domain. As displayed in Table 40.7, five malicious domains were predicted using the developed model. The output value is a probability value. If the probability is more significant than 0.5, the domain is considered to have exfiltration. For

Table 40.5 Encrypted data attached to the domain name

Encryption	Domain name
Mzc2MzgwNTA3NTY3NzAx5	Mzc2MzgwNTA3NTY3NzAx5.malware.com
30993683d51e835756d02f655af05ac	30993683d51e835756d02f655af05ac.malware.com
1e8299e0c7a1690ec3d6928f2a8366a	1e8299e0c7a1690ec3d6928f2a8366a.malware.com
275c15e507e168f5f71eb848ctq6cfd3180ec	275c15e507e168f5f71eb848ctq6cfd3180ec0a33.malware.com

Table 40.6 Validation metrics

Metric	Our proposed method
roc_auc	1.0
precision	0.9999992885195514

Table 40.7 Domain and probability

Domain	Probability
59-124-10-43.hinet-ip.hinet.net	0.000015615
59-124-106-74.hinet-ip.hinet.net	0.0000823
130353161663731383536663564646236.google.com	0.63
init.ojswczdnmuxg2zd4ge3q.base64.systw.net	0.56

example, for `init.ojswczdnmuxg2zd4ge3q.base64`, the probability value corresponding to `.systw.net` was 0.56. Thus, this case is considered to have exfiltration. Conversely, the probability value corresponding to `59-124-106-74.hinet-ip.hinet.net` was 0.00008032. Thus, this case does not have exfiltration.

40.6 Conclusion

In this study, we present a framework to detect DNS tunneling and DNS exfiltration through the DNS network traffic. This framework uses unsupervised learning and the silhouette method to determine the optimal number of clusters for identifying DNS tunneling. We also used a supervised learning algorithm to train the XGBoost model for identifying whether any information leakage occurred in the traffic. Finally, to verify the effectiveness of our architecture, we used open-source DNS network traffic that contained tunneling and exfiltration. The proposed framework can accurately detect tunneling and exfiltration and prove the robustness, simplicity, and scalability of our method.

References

1. DNS: https://en.wikipedia.org/wiki/Domain_Name_System
2. Anirban, D., Min-Yi, S., Madhu, S.: Detection of exfiltration and tunneling over DNS. In: 16th IEEE International Conference on Machine Learning and Applications (2017)
3. Binsalleeh, H., Kara, A.M., Madhu, S., Debbabi, M.: DNS noise: characterization of covert channels in DNS. In: Mobility and Security (NTMS) (2014)
4. Farnham, G., Atlasis, A.: Detecting DNS Tunneling. SANS Institute InfoSec Reading Room (2014)
5. Paxson, V., Christodorescu, M., Javed, M., Rao, J.R., Sailer, R., Schales, D.L., Stoecklin, M.P., Thomas, K., Venema, W., Weaver, N.: Practical comprehensive bounds on surreptitious communication over DNS. In: USENIX Security (2013)
6. Dietrich, C.J., Rossow, C., Freiling, F.C., Bos, H., Van Steen, M., Pohlmann, N.: On bot-nets that use DNS for command and control (2011)
7. Born, K.: DNS tunnel detection using character frequency analysis (2010)
8. Hind, J.: ExFILD: Catching DNS tunnels with ai (2009)

9. Jawad, A., Hassan, H., Qasim, R., Craig, R., Vijay, S., Lee.: Real-Time Detection of DNS Exfiltration and Tunneling from Enterprise Networks. Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia (2019)
10. Fawcett, T.: ExFILD: a tool for the detection of data exfiltration using entropy and encryption characteristics of network traffic (2010)
11. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Comput Appl Math* (1987)
12. XGBoost: A Scalable Tree Boosting System
13. Datasets: http://downloads.majestic.com/majestic_million.csv
14. Alexa: <http://s3.amazonaws.com/alexa-static/top-1m.csv.zip>
15. Data Exfiltration Toolkit: <https://github.com/qasimraz/DET>
16. TXT Record: <https://github.com/chuayupeng/dns-tunnelling-detection/tree/master/pcap>
17. DNScat: <http://bit.ly/1PhF8Qd> (2004)
18. Credit card generator: <https://www.fakepersongenerator.com/credit-card-generator>
19. Scikit-learn: <https://scikit-learn.org/stable/>

Chapter 41

Activity Recognition Based on Latent Knowledge Mining in Smart Home



Yu Tong, Rong Chen, and Bo Yu

Abstract Activity recognition in smart home is one pattern recognition problem. Many activity recognition algorithms have appeared so far to recognize activities in smart home. Past researches have proved that dynamic and deep knowledge mining algorithms will help improve the accuracy. But because of the uncertainty of sensors and the complexity of the user activities, existing activity recognition methods still have a lot of room for improvement. Considering there is some latent knowledge existed in sensors or user activities, this paper proposes to recognize activities by exploring latent knowledge. Firstly, this paper improves activity recognition by extracting latent knowledge between sensors and activities, thereby proposed one feature preprocessing method. Then, it proves one new multi-resident activity recognition method based on latent knowledge in multi-resident activities. Simulations conclude that extracting latent knowledge can greatly enhance activity recognition.

41.1 Introduction

Non-invasive activity recognition (NAR) is an ambient intelligence technology which can recognize activities based on non-invasive sensors without affecting the living conditions of residents. NAR has an important application in the field of smart home that can understand individual behavior, group behavior, and the interaction between people and the environment. Over the past decade, most of the previous NAR methods are based on pattern recognition methods and have evolved from static algorithms to

Y. Tong (✉)

School of Computer Science and Technology, Hefei Normal University, 230601 Hefei, China
e-mail: tongyu24@126.com

R. Chen

College of Information Science and Technology, Dalian Maritime University, 116026 Dalian, China

B. Yu

School of Civil Engineering, Hefei University of Technology, 230009 Hefei, China

dynamic algorithms, from simple feature representation to multi-level deep knowledge mining algorithms [1], such as support vector machine [2], Naive Bayes [3], hidden Markov models [4], latent-dynamic conditional random fields [5] and deep learning [6]. Past research has proved dynamic activity recognition algorithms and deep knowledge mining algorithms will help improve the accuracy of NAR [5, 7]. Based on these considerations, this paper will enhance NAR based on dynamic activity recognition algorithms by mining latent knowledge that exists in sensors and activities.

On one hand, with low-cost sensors and wireless sensor networks development, various passive sensors have recently been used to recognize activities [8, 9] in non-invasive smart home. For example, motion sensors which are installed in the floor can capture human motion and the RFID tags which are attached to the object can capture human-to-environment interaction. However, observed feature dimension increases with the increase of sensors and causes the higher computational complexity. The more features are used, the higher computational complexity it will cause, while modeling fewer features is often insufficient to ensure recognition accuracy. Besides, activity observation feature which composed of sensor observation is often abundant and sometime redundant. Thus, the feature selection directly contributes to the performance of the recognition model. Principal component analysis (PCA) [10] is tested for feature generation, but the algorithms need to choose appropriate principal component number and number selection is an impact on the result. Fortunately, there is a lot of latent knowledge in sensor networks, such as multiple sensors that are often related to only one activity. It will help NAR if we can mine the latent knowledge between sensors and activities.

On the other hand, the activities in activity recognition are not only the activities of a single person, but also the activities of multiple residents. Multi-resident activity recognition (MRAR) is more difficult due to user activity interfering with each other. In a smart home with non-obtrusive sensors, MRAR often uses data association [4, 11] which associates sensor data to the person who triggered the sensor or changed the value. To improve the MRAR accuracy, dynamic Bayesian networks such as CHMM and FCRF often used to model interacting process [12, 13]. However, data associations are often unknown and hard to obtain in ubiquitous sensor environment. Beside, for multiple residents in smart home with non-obtrusive sensors, who triggered the sensor is often ambiguous and there are not strong underlying data associations to use. If the data association is incorrect, the MRAR will be correspondingly inaccurate. So, it would be interesting to find a method for MRAR that does not rely on data association. Fortunately, there is some latent knowledge which is often invariant in multi-resident environment. For instance, there are some global features and trends, playing chess collaboratively, only one person can use computer at the same time since there is only one computer. The latent knowledge is often easy to represent in multi-resident environment. If we can mine the latent knowledge well, multi-resident activity recognition will be improved.

The paper is organized like this, in Sect. 41.2, it will introduce one activity recognition method by extracting latent knowledge between sensors and activities. Then, Sect. 41.3 will prove one new multi-resident activity recognition method based on

latent knowledge in multi-resident activities. Section 41.4 is validation. Finally, the article concludes with some conclusions.

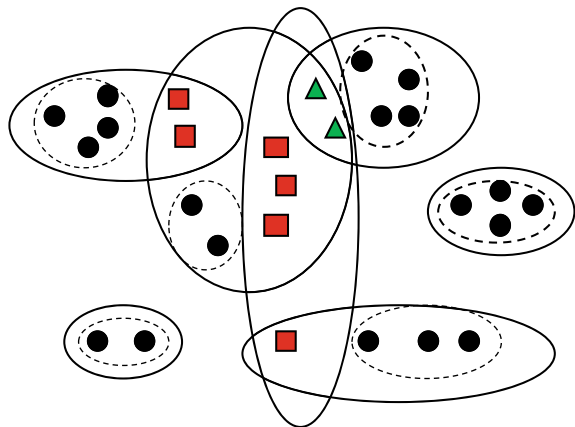
41.2 Latent Knowledge Between Sensors and Activities

Motivated by the relationship between sensors and activities, this section will combine the multiple features that related only one activity as one feature and use CRF to recognize activities in smart homes. To describe our method, we start from analyzing the relationship between sensor data and activities.

The activity observation feature vector at time t is often denoted as $\mathbf{x}_t = (x_t^1, x_t^2, \dots, x_t^N)$, where N is the dimension of observation feature. By considering one sensor as one observation feature, the dimension of observation feature vector will equal to the total sensor number. However, it is common that some sensors are related to only one activity and some sensor states are related to more than one activity. Figure 41.1 is the relationship between sensors and activities. In the figure, the sensors that related only one activity are denoted as “●,” the sensors that related two activities are denoted as “■,” and the sensors that related more than two activities are denoted as “▲.” When the state of sensor data that relates only one activity and does not relate other activities changed, it is easy to deduce that the related activity rather than other activities is being carried out. In addition, it is also common that several sensors related only one activity and does not relate other activities (“●” that in dashed circle). When one or several states of those sensors changed, we can deduce the related activity is being carried out.

If we regard the relationship between sensors and activities as latent knowledge and the sensors observation that related only one activity and do not relate other activities as one combined observation feature, we can deduce the observation feature vector $(x_t^1, x_t^2, \dots, x_t^N)$ to $(x_t^1, x_t^2, \dots, x_t^L)$, where N and L is the observation feature

Fig. 41.1 Relationship between sensors and activities



dimension before and after feature combining. The feature combining method is shown in Algorithm 1.

Algorithm 1. Feature combining method

Input: Observation feature x_0 :

$$\{\mathbf{x}_t = (x_t^1, x_t^2, \dots, x_t^N), t = 1, \dots, T_0\},$$

activity labels y_0 :

$$\{y_t, t = 1, \dots, T_0\}$$

Output: Combined feature X_0 :

$$\{\mathbf{x}_t = (x_t^1, x_t^2, \dots, x_t^L), t = 1, \dots, T_0\}$$

1. Find the number of activities $N_a = \max\{y_0\}$ and the number of sensors $N_s = N$;
 2. Find the sensors that related to activity i and put them in sensor set $S_i = \{s_i\}$, $i = 1, \dots, N_a$;
 3. For every s_j in S_i , put it to set C_i if s_j does not appear in other set S_j , $j \neq i$, $i, j = 1, \dots, N_a$;
 4. Combined the sensors in C_i as one combined feature;
 5. For every observation feature \mathbf{x}_t in x_0 , update the features that corresponded sensors in C_i with the combined feature, $i = 1, \dots, N_a$;
 6. Denote the updated x_0 as X_0 ;
 7. Return X_0 .
-

41.3 Latent Knowledge in Multi-resident Activities

MRAR is to infer multi-resident activities from observations. Multi-resident activity sequence is often denoted as $\{y_1, y_2, \dots, y_T\}$ and observation is often denoted as $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$. For $t = 1, 2, \dots, T$, y_t represents multi-resident activities at time t , and \mathbf{x}_t represents sensor observation vectors at time t . Both y_t and \mathbf{x}_t are multi-dimensional variables, where the dimension of y_t is the number of residents, and the dimension of \mathbf{x}_t is the number of observation feature. MRAR with machine learning method often needs some empirical data to train a recognition model, where empirical data are often used as training samples $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_{T_0}, y_{T_0})\}$. To better illustrate the problem, this paper will give a multi-resident scenario below.

Scenario: Two residents (ID = 1 and ID = 2) randomly perform three daily activities in one smart home. The three activities are labeled 1, 2, 3, and 0 if the user does not perform any activity. Assume that $A = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_7, y_7)\}$ is the collect empirical data when the two residents perform activities, where $y_1 = (0, 0)$, $y_2 = (1, 0)$, $y_3 = (1, 0)$, $y_4 = (3, 0)$, $y_5 = (1, 1)$, $y_6 = (1, 1)$, and $y_7 = (3, 3)$. In this case, $y_t = (y_t^1, y_t^2)$ is two-dimension, where y_t^i , $i = 1, 2$ represent the activity ID that the i th resident performed at time t .

For the four activity labels (1, 2, 3, and 0) for two residents, theoretically we can get $4 \times 4 = 16$ different vectors $\{(0,0), (0,1), \dots, (3,3)\}$. However, due to resident preferences, some exclusive and independent activities occurred, some states we cannot observe in fact. So, only seven label vectors are obtained.

To represent prior knowledge, some terms will be introduced below.

A single label y often has multiple possible values, which we often denote as state and all the possible value sets as state set. MRAR can be seen as a multi-label state labeling problem. For m residents, there are many possible values for $(y_t^1, y_t^2, \dots, y_t^m)$, since different residents may perform different activities.

Here, we use the **State Event** $(y_t^1, y_t^2, \dots, y_t^m)$ to represent the activities of multiple users at the same time, use **State Event Set A** to represent various values of State Events, and use **State Event Matrix M** to represent the values of State Events at time 1 to T . Then, State Event Set denotes as

$$A = \{(y_1^1, y_1^2, \dots, y_1^m), (y_2^1, y_2^2, \dots, y_2^m), \dots, (y_K^1, y_K^2, \dots, y_K^m)\}$$

where K is the State Events number.

State Event Matrix is given by

$$M = \begin{bmatrix} y_1^1 & y_1^2 & \dots & y_1^m \\ y_2^1 & y_2^2 & \dots & y_2^m \\ \vdots & \vdots & \ddots & \vdots \\ y_T^1 & y_T^2 & \dots & y_T^m \end{bmatrix}$$

which includes T State Events.

For above multi-resident scenario, State Event Set can be denoted as

$$A1 = \{(0, 0), (1, 0), (3, 0), (1, 1), (3, 3)\}$$

The State Event Matrix can be denoted as

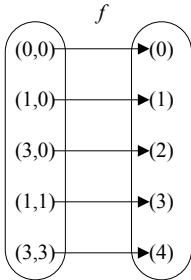
$$M1 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 3 & 0 \\ 1 & 1 \\ 1 & 1 \\ 3 & 3 \end{bmatrix}$$

Note that two State Events in M may be the same, but any two State Events in A are different, and all State Events in M could be found in A . From $M1$ we can see that the State Event 2 and 3 are the same, State Event 5 and 6 are the same, and all State Event can be found in $A1$.

Represented $(y_t^1, y_t^2, \dots, y_t^m)$ with one uniquely combined label C , it can get **combined label states set B** $= \{0, 1, \dots, K - 1\}$. The map between State Event $(y_k^1, y_k^2, \dots, y_k^m)$ and combined label state $C_k \in B$ is defined as

$$(y_k^1, y_k^2, \dots, y_k^m) \longrightarrow f C_k$$

For the State Event Set $\mathbf{A1} = \{(0, 0), (1, 0), (3, 0), (1, 1), (3, 3)\}$, there a recombined label states set $\mathbf{B1} = \{0, 1, 2, 3, 4\}$. The mapping is defined as



Similarly, there are mapping between $\mathbf{M1}$ and $\mathbf{B1}$ as follows

$$\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 3 & 0 \\ 1 & 1 \\ 1 & 1 \\ 3 & 3 \end{bmatrix} \xrightarrow{f} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 2 \\ 3 \\ 3 \\ 4 \end{bmatrix}$$

The states of single activity $y_t^i, i \in \{0, 1, \dots, K - 1\}$ can be obtained by inverse mapping. In the two-resident scenario, for $C = 1$, it can get $y_1 = 1, y_2 = 0$ by inverse mapping, and for $C = 3$, it can get $y_1 = 1, y_2 = 1$ by inverse mapping.

The algorithm of extracting latent knowledge in multi-resident activities is given in Figs. 41.2 and 41.3. The former is the model building flowchart, with which we can get State Event Set A , mapping f , combined label states set B , and combined label recognition model. The latter is activity recognizing flowchart from which it can see that new testing multi-resident activities are recognized with two steps: firstly, recognize the states of combined label C , then inverse map C to State Event by f^{-1} . Finally, figure out multi-resident activities based on the State Event.

It can be seen that the extracting latent knowledge algorithm did not use data association when recognizing multi-resident activities. But, if there is a need (i.e., tracking the resident), it can also find out data association. For $C = 1$, if figure out $y_1 = 1, y_2 = 0$, we say the data is get by the first resident, since $y_2 = 0$ represents the second resident does not carry out any activity and considered not trigger any sensors.

The algorithm can also handle some uncertain multi-resident activity patterns. For the two residents activity label $(A1, A2)$, where $A1$ is the activity that the first resident performed and $A2$ is the activity that the second resident performed. If $A1$ or $A2$ is equal '0,' it means the resident is performing one unknown activity and can be any one activity. For two residents with N total activities, there are

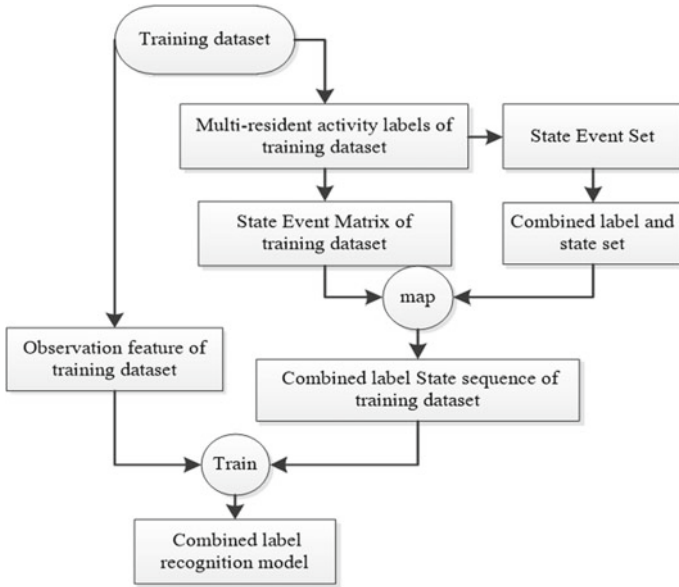
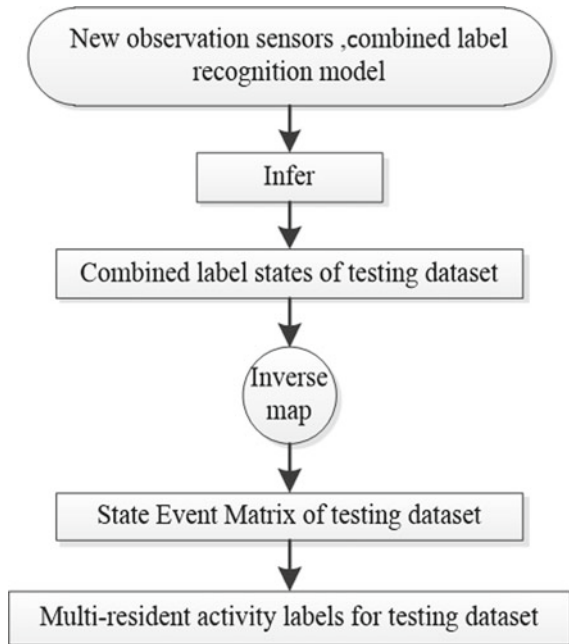


Fig. 41.2 Model building flowchart

Fig. 41.3 Activity recognizing flowchart



$$\begin{aligned}
 (0, A2) &= (1, A2) \vee (2, A2), \dots, \vee(N, A2) \\
 (A1, 0) &= (A1, 1) \vee (A1, 2), \dots, \vee(A1, N) \\
 (0, 0) &= (0, A2) \vee (A1, 0)
 \end{aligned}$$

The unknown state is actually a union of all possible activities, thus the algorithm can handle some uncertainties and can improve activity recognition.

41.4 Validation

41.4.1 Validation 1

To validate our feature combining method, two experiments will be given. For every experiment, we will introduce the datasets, their activities, and sensor features, and then carry out our experiments. To measure the percentage of correctly classified testing samples, we define the recognition accuracy of the class as

$$\text{Accuracy} = \frac{\sum_{n=1}^N [\text{inferred}(n) = \text{true}(n)]}{N} \quad (1)$$

where N is total testing samples.

In addition, to verify the recognition of a single class, we also give the recognition accuracy of individual activities as

$$\frac{\text{inferred}_c(n) = \text{true}_c(n)}{N_c} \quad (2)$$

where N_c is the total samples that contained in class c .

Experiment 1

The first experiment is based on the ‘‘ADL adlnormal’’ dataset that is collected in WSU Apartment Test bed [14]. There are five daily activities in the dataset. The apartment is installed with various non-invasive sensors.

The raw sensor number, cleaned sensor number, and the finally sensor number after sensor combining are shown in Table 41.1. As it was shown, the sensor number decreases obviously after sensor cleaning whereas the finally sensor number after

Table 41.1 Sensor number changes after cleaning and combining

Feature	Sensor number
All feature	39
Clean feature	23
Finally feature	19

Table 41.2 Involved sensor number changes before and after sensor combining for five individual activities

	A1	A2	A3	A4	A5
Before	13	14	16	15	17
After	9	14	16	15	17

sensor combining does not decrease much. The involved sensor numbers before and after sensor combining for the five activities can be seen in Table 41.2.

After the sensors cleaning and combining, we take every sensor as a feature and recognize activities in this dataset using CRF with threefold cross-validations. To validate our sensors combining solution, we compared the results with the result before sensor cleaning and sensor combining. Table 41.3 is the recognition accuracy and the total time that used for both training and testing with raw sensor and finally sensor. From the table, we can see that the recognition accuracy is increased after sensor cleaning and sensor combining whereas the time used is reduced.

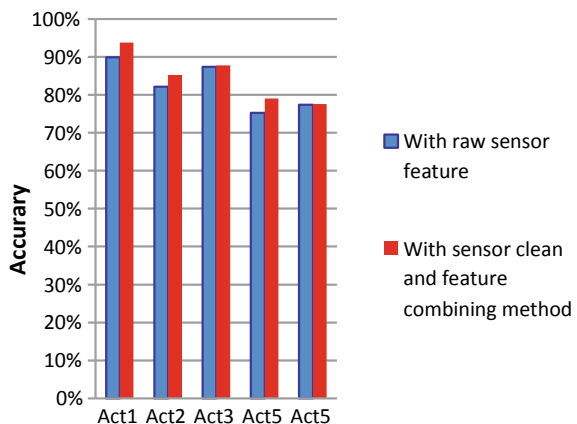
The recognition accuracies for five individual activities with the raw sensor and with the sensor after cleaning and combining are shown in Fig. 41.4. We can see that the recognition accuracies of all the activities are increased after sensor cleaning and sensor combining for that our method not only can reduce parameter in count, but also can avoid the error caused by redundant information.

Experiment 2 The second experiment focuses on routine morning activities collected in kitchen outfitted with 60 RFID tags [8]. In the kitchen, 11 routine morning

Table 41.3 Recognition accuracy and the total time changes

	Accuracy	Time
Raw sensor	0.8308	77.1716
Senor after combining	0.8460	58.1504

Fig. 41.4 Recognition accuracies for five individual activities



activities are performed by user in different ways and use the RFID tags to collect sensor data. The objects that attached tags include bowl, coffee container, cupboard, dishwasher, door, drawer, egg carton, hand soap, kettle, cooking spoon, stove control, telephone, and so on.

Before combining the sensors, we first clean the uninvolved sensors, since some sensors may not involve in all the activities. The raw sensor number, cleaned sensor number, and the finally sensor number after combining are shown in Table 41.4. From the table, we can see that all of the 60 sensors are involved in the activities and the total sensor number decreases obviously after combining.

Also, we give the involved sensor number before and after combining for 11 individual activities in Table 41.5, where $A_i, i = 1, \dots, 11$, presents the i th activity. The table shows that the sensors in activities 4, 5, 6, 8, 10 are combined, and thus the involved sensor numbers are decreased.

After sensors combining, we take every sensor as a feature. To validate the algorithm of extracting latent knowledge between sensors and activities, we recognize activities based on CRF with leave-one-out cross-validation. Also, we compare the results with PCA method that extracting 35 principal components as features. Table 41.6 is the recognition accuracy and the total time that used for both training and testing with different method. From the table, we can see that the recognition accuracy is increased after sensor combining whereas the time used is reduced. Although based on the feature with same dimension, PCA gets worse result than the algorithm

Table 41.4 Sensor number changes after cleaning and combining

Sensor type	Sensor number
Raw sensor number	60
Sensor number after cleaning	60
Sensor number after combining	35

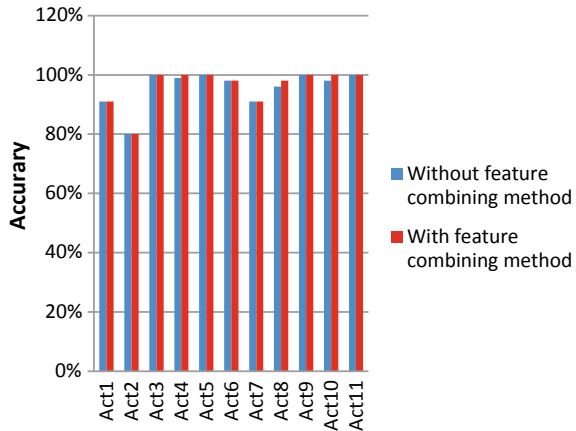
Table 41.5 Involved sensor number before and after combining for 11 individual activities

Activity	Before	After
A1	13	13
A2	10	10
A3	1	1
A4	17	7
A5	6	4
A6	15	9
A7	6	6
A8	13	10
A9	14	14
A10	5	1
A11	1	1

Table 41.6 Recognition accuracy and the total time changes

	Accuracy	Time (s)
Raw feature	0.9355	1245.75
Combined feature with 35 feature	0.9388	1043.30
PCA with 35 feature	0.9228	1234.03

Fig. 41.5 Recognition accuracies for 11 individual activities



that extracting latent knowledge between sensors and activities with much time that used for both training and testing.

The recognition accuracies for 11 individual activities are shown in Fig. 41.5. We can see that the recognition accuracies of activity 4, 8, 10 are increased after sensor combining. This is because our method not only can reduce parameter in count, but also can avoid the error caused by redundant information.

41.4.2 Validation 2

We will validate our algorithm exploiting latent knowledge of multi-resident activities based on multi-resident activities dataset [4] collected in the CASAS project. In the dataset, there are two residents and 15 activities.

The multi-resident activity State Event and their frequency F are shown in Table 41.7.

As Table 41.7 shows that some State Events occur frequently, while some appear rarely, does not happen actually. For two residents with 16 (activity 0 represent the resident performed unknown activity), there are $16 \times 16 = 256$ State Events theoretically, but in this case, there are only 27 State Events, since some State Event do not occur at all actually. To validate our algorithm, one experiment will be given.

Table 41.7 Occurrence counts of different State Events

(A1, A2)	F	(A1, A2)	F	(A1, A2)	F
(0, 0)	3	(0, 15)	748	(10, 11)	284
(0, 2)	1568	(1, 0)	1175	(10, 15)	2
(0, 3)	668	(4, 0)	864	(12, 0)	1179
(0, 4)	1	(6, 0)	1505	(12, 13)	272
(0, 5)	545	(6, 7)	350	(13, 0)	4
(0, 7)	1529	(9, 0)	866	(13, 13)	865
(0, 8)	432	(9, 8)	280	(14, 0)	387
(0, 11)	891	(10, 0)	660	(14, 15)	845
(0, 13)	1309	(10, 1)	1	(15, 0)	1

Experiment 3 This experiment is carried out with three fold cross-validations. In the training stage, firstly, it will build mapping f and inverse mapping f^{-1} , and map State Event Matrix of training data as combined label state sequence. Then, dynamic activity recognition algorithm, such as HMM, CRF, and latent-dynamic conditional random fields (LDCRF) [4], is trained with observation sequences and combined label states sequence.

In the testing stage, we estimate combined label state firstly with the trained model and observation sequences in the test dataset. The average accuracy of combined label state for HMM with latent knowledge (LK-HMM) and CRF with latent knowledge (LK-CRF), and LDCRF with latent knowledge (LK-LDCRF) are 65.46, 67.61, and 63.87% correspondingly.

It is important to note that the above is not the ultimate accuracy of MRAR. To get multi-resident activity of test dataset, we need to map combined label states to State Event Matrix with f^{-1} . Figure 41.6 is the average MRAR accuracies of five models. From it, we can see that LK-HMM gets 75.77%, LK-CRF gets 75.38%, and LK-LDCRF gets 72.69% which all get higher average MRAR accuracies than single

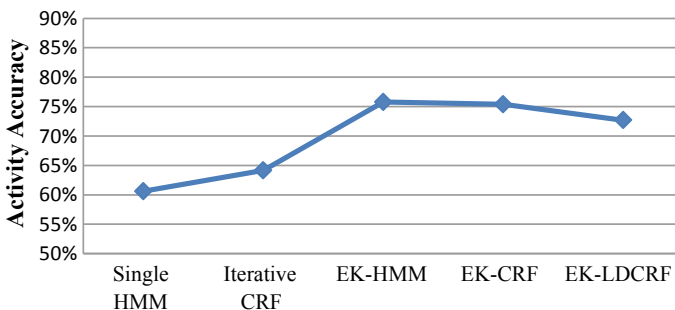


Fig. 41.6 Average recognition accuracies of five models in recognizing activities for multi-residents

HMM and iterative CRF which did not give data association when MRAR. Thus, mining latent knowledge can help MRAR.

There are some reasons for LK-HMM, LK-CRF, and LK-LDCRF to outperform single HMM and iterative CRF. Firstly, when single HMM model is implemented for both residents, it cannot well represent transitions between activities and multi-resident at the same time. Secondly, although iterative CRF does not need to give data associations, it still needs to learn data associations for MRAR. When data association is learned badly, MRAR accuracy will be low. In addition, LK-HMM, LK-CRF, and LK-LDCRF can mine knowledge in the multi-resident environment and can capture global features and trends of multi-resident activities.

In the LK-LDCRF case, it gets lower accuracy than LK-CRF which gets higher accuracy than CRF in single-user activity recognition [4]. This is because there are many combined label states with different internal structure, fixed hidden state for LK-LDCRF is difficult to adapt to all combined label states. Thus, the number of hidden states is difficult to determine, and it is not easy to determine. If we chose not suitable hidden states number, LK-LDCRF will get lower accuracy. In future, we will study the chosen suitable hidden states number for LDCRF in multi-resident environment and compare the result to our method.

41.5 Conclusions

This paper recognizes activities with some latent knowledge that exists in training samples. Firstly, this paper gives one new pretreatment method for activity recognition by extracting latent knowledge between sensors and activities, and then it proves one new multi-resident activity recognition algorithm by extracting latent knowledge in multi-resident activities. From the simulation, we conclude that extracting latent knowledge can greatly enhance activity recognition.

Acknowledgements This work was supported by the Anhui Provincial Natural Science Research Project of Colleges and Universities (No. KJ2017A933), the National Natural Science Foundation of China (No. 11872166) and China Postdoctoral Science Foundation (No. 2016M592042).

References

1. Inoue, M., Inoue, S., Nishida, T.: Deep recurrent neural network for mobile human activity recognition with high throughput. *Artif Life Robot* **23**(2), 173–185 (2018)
2. Kadouche R, Pigot H, Abdulrazak B et al (2011) User's behavior classification model for smart houses occupant prediction. In: *Activity recognition in pervasive intelligent environments*. Atlantis Press, pp 149–164
3. van Kasteren T, Krose B (2007) Bayesian activity recognition in residence for elders

4. Singla, G., Cook, D.J., Schmitter-Edgecombe, M.: Recognizing independent and joint activities among multiple residents in smart environments. *J Ambient Intell Humaniz Comput* **1**(1), 57–63 (2010)
5. Tong, Y., Chen, R.: Latent-dynamic conditional random fields for recognizing activities in smart homes. *J Ambient Intell Smart Environ* **6**(1), 39–55 (2014)
6. Wang, J., Chen, Y., Hao, S., et al.: Deep learning for sensor-based activity recognition: a survey. *Pattern Recogn Lett* **119**, 3–11 (2019)
7. Ma, C.Y., Chen, M.H., Kira, Z., et al.: Ts-lstm and temporal-inception: exploiting spatiotemporal dynamics for activity recognition. *Sig Process Image Commun* **71**, 76–87 (2019)
8. Patterson DJ, Fox D, Kautz H et al (2005) Fine-grained activity recognition by aggregating abstract object usage. In: Ninth IEEE international symposium on wearable computers (ISWC'05). IEEE, pp 44–51
9. Kasteren TV, Noulas A et al (2008) Accurate activity recognition in a home setting. In: Proceedings of the 10th international conference on Ubiquitous computing, Seoul, Korea
10. Mantyjarvi J, Himberg J, Seppanen T (2001) Recognizing human motion with multiple acceleration sensors. In: 2001 IEEE international conference on systems, man and cybernetics. e-Systems and e-Man for cybernetics in cyberspace (Cat. No. 01CH37236). IEEE, vol 2, pp 747–752
11. Hsu KC, Chiang YT et al (2010) Strategies for inference mechanism of conditional random fields for multi-resident activity recognition in a smart home. In: The twenty third international conference on industrial, engineering & other applications of applied intelligent systems (IEA-AIE), Córdoba, Spain, pp 417–426
12. Chiang YT, Hsu KC (2010) Interaction models for multi-resident activity recognition in a smart home. *Intelligent robots and systems (IROS)*, pp 3753–3758
13. Wang, L., Gu, T., et al.: Recognizing multi-user activities using wearable sensors in a smart home. *Pervasive Mob Comput* **7**, 287–298 (2011)
14. Cook, D.J., Schmitter-Edgecombe, M.: Assessing the quality of activities in a smart environment. *Methods Inf Med* **48**(05), 480–485 (2009)

Chapter 42

Pattern Retrieval on the Game of Go



Shi-Jim Yen and Yu-Chie Ho

Abstract It is said that the game of Go is the drosophila of artificial intelligence and machine learning. A series of AlphaGo and AlphaZero programs created history. For assisting human learning, we believe Go can also play the role of the drosophila. Interesting games and specific capturing races problems in Go game are recorded in Smart Game Format files. Valuable information takes place in these files. A Go information retrieval system would be helpful for assisting human learning Go. This article proposes a sequential algorithm to handle the Go pattern searching process in the Go game records, but the search time increases linearly with the size of game records. Thus, we make the index structure in the Go game database. The index structure also integrates methods of information retrieval and human domain knowledge of the Go game, four applications of proposed structure based on user needs. The index structure can improve the speed of pattern retrieval on the game of Go.

42.1 Introduction

The game of Go consists of a board, black, and white stones. The board is made of 19 vertical and horizontal lines each, and the players, each holding either black or white pieces, put their pieces on the intersections of the board alternately. The goal of the game is to surround the territories of the board using one's own stones or the side of the board. When the game is over, the player that surrounds the larger territory wins. Although it has simple rules, the player will face a lot of challenges during the game. The game of Go is the drosophila of artificial intelligence. A series of AlphaGo and AlphaZero programs created a new history for artificial intelligence and machine learning [1, 2]. Table 42.1 shows the space state complexity and the game tree complexity of Chinese chess, Shogi, and Go. Those games are the most

S.-J. Yen (✉) · Y.-C. Ho

Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan, R.O.C.

e-mail: sjyen@gms.ndhu.edu.tw

Y.-C. Ho

e-mail: yccho@gms.ndhu.edu.tw

© Springer Nature Singapore Pte Ltd. 2021

S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation, Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_42

587

Table 42.1 Complexity of the popular games

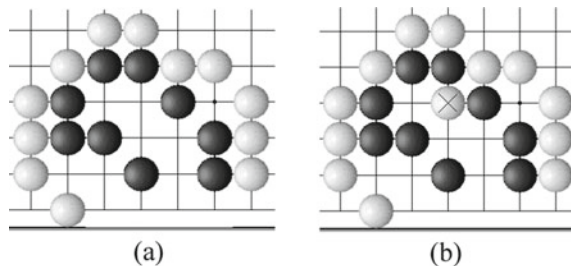
Game	Space states complexity ($\log n$)	Game tree complexity ($\log n$)
Western chess	50	123
Chinese chess	48	150
Japanese Shogi	71	226
The game of Go	160	400

publicly known board games in the world. The complexity of Chinese chess was estimated by Allis [3] and Yen et al. [4]. The complexities of other games were estimated by Bouzy and Cazenave [5] and Iida et al. [6]. Observably, the complexity of Go is greater than other games.

Many professional Go game records, capturing races, cutting, connecting, and other tsumego are saved as files through the Smart Game Format. A lot of important information is saved in these game record files, including the sequence and positions of stones, the names of the players, final results, and comments. Information retrieval in Go game records is very useful. The Go teacher can help the student to learn and analyze the improvements based on the information in the game records [7]. A Go player may be interested in how to react to a certain game situation, for example, an uncommon tsumego, illustrated in Fig. 42.1a. The information retrieval system can quickly help the player to find the answer in the tsumego game records database, illustrated in Fig. 42.1b. In addition, if one wants to know the win rate of open game strategies, the answer can be found in the Go information retrieval system. As for the researchers of computer Go, the retrieval of game records can help to correct the content of the Go pattern database or help to debug Go programs.

When using the retrieval information in the game records, the most difficult part is to find a Go pattern among a lot of game records. Deep learning policy network could output a move for a game board, but it cannot output all the game records containing some pattern [8–10]. This problem can be included in the pattern matching problems. A Go pattern is a certain arrangement of stones. The size of a pattern can start from 3×3 to 19×19 . The state of each position in the pattern can be white, black, or no stone. With the image operations rotation, reflection, and color changing, one pattern has 16 situations on a board, as shown in Fig. 42.2.

Fig. 42.1 **a** A tsumego problem. **b** Key point for white to capture black stones



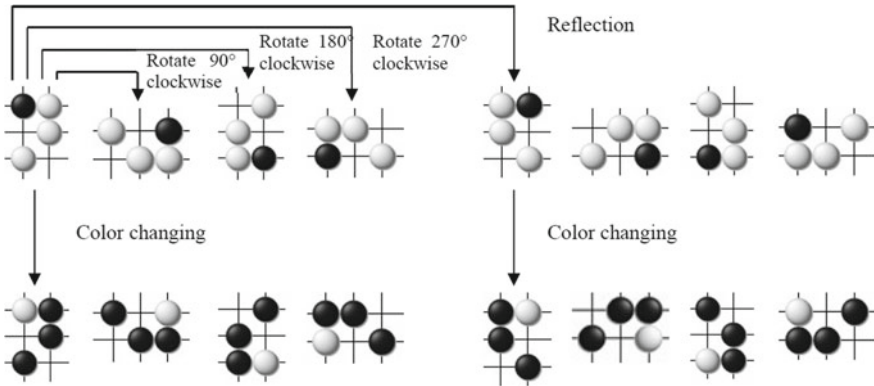


Fig. 42.2 Sixteen variances of a pattern

Pattern matching is important in the game of Go [11–13]. It helps to adjust the strength of computer Go programs [14, 15]. Predict life and death and score final positions also need Go pattern matching [16, 17]. This article will discuss how to find those game records containing the desired query pattern in a lot of game records. The input is a query pattern, and outputs are the game records which contain this query pattern. From those outputs, we can get the game record information, such as name of the players, move sequences, players’ strength, handicap numbers, and results.

The organization of the article is as follows: The proposed framework is in Sect. 42.2. Section 42.3 describes how to process the raw Go game records. Section 42.4 gives a pattern matching algorithm. Section 42.5 gives the index structure to speed up the matching. Section 42.6 proposes four applications of our index structure. Finally, Sect. 42.7 gives a conclusion.

42.2 Information Retrieval System

Our final goal is to develop an efficient retrieval system. Figure 42.3a shows its framework. The ability to receive both pattern query and text query will be an important advantage of the system because it allows users to search for information with more directions. For instance, the user may want to know the win rate when plays with a specific opening pattern, (SGF records record the names of the players and the results of the current game) by sending a query content with name of the player, result, and the opening pattern, the game records will be received, and the win rate can be computed.

When the query is with a logical operation like *and*, *or*, *not* ..., etc., the proposed structure makes these operations be done more quickly. We could also set up some

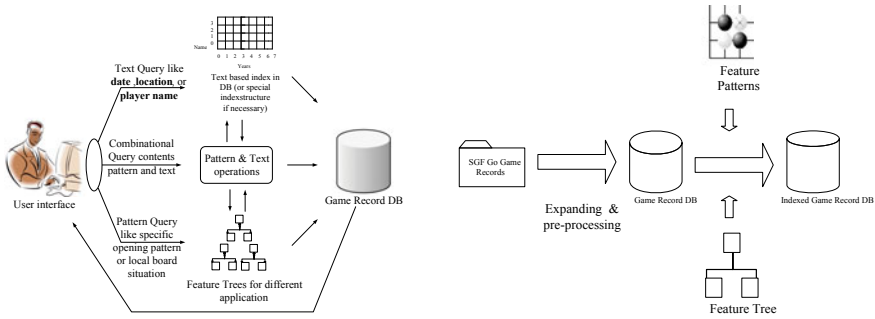


Fig. 42.3 a The information retrieval system for Go. b The index structure

ranking rules by information in the game record. For example, games played by professional players will get higher weights to be shown on the top of the list.

The procedure and main components of our idea were shown in Fig. 42.3a. It is based on a sequential matching algorithm. It will be used during the constructing state and the final searching state. Besides the searching algorithm, there are two main factors that affect the performance of the system: feature patterns and feature trees (Fig. 42.3b). The sequential matching algorithm will be proposed in Sect. 42.4. Feature patterns and text and feature trees will be discussed in Sect. 42.5.

42.3 Converting and Expending of the SGF Go Game Records

The raw game record format used in this paper is the most popular game record format in current days: Smart Game Format (SGF). SGF was developed by Auder Kierulf in 1987 [18]. It has been widely used to record Go games. Many digitized Go books are also saved in SGF. SGF is a text format, recording the game board information between two players. Records include the sequence and positions of moves, the names of the players, results, comments, and so on. We must convert and expend SGF Go game records into the format we need and store it in the database. We pick small parts of an SGF example and show the converting processes.

An SGF record could be divided into two parts. For the following example, let us mark them by S_1 and S_2 , respectively. S_1 records game information like players, location, time of this game, and handicap information. For example, SZ [19] means the size of the game is 19×19 . PW[guojuan] means the player of white is guojuan. RE[W+Resign] means the game was won by white with black resign. In the pattern matching process, we will skip most of them, only keep handicap information which is typed in boldface, e.g., AB[**dd**] [pd] [jj] [dp] [pp], and add them to the top of the S_2 .

S_1 :

(; GM [1] FF [4] SZ [19] PW[guojuan] WR[5p] PB[wolf] BR[1k] DT[2000-7-19] PC[The Kiseido Go Server (KGS) at <http://kgs.kiseido.com/>] KM[0.50] RE[W+Resign] HA [5] RU[Japanese] CA[UTF-8] TM[1800] OT[5 30 byo-yomi] **AB[dd] [pd] [jj] [dp] [pp]**

S_2 :

; W[mp]; B[pn]; W[fq]; B[cn]; W[fo]; B[nq]; W[mq]; B[dq]; W[ql]; B[qm]; W[pl]; B[nm]; W[pg]; B[nl]; W[np]; B[oq]; W[on]; B[om]; W[pm]; B[qn]; W[cf]; B[fc]; W[oo]; B[lm]; W[po]; B[qo]; W[qp]; B[qq]; W[rp]; B[rq]; W[ro]; B[op]; W[mn]; B[sp]; W[rm]; **B[so]**; W[rn]; B[mm]; W[mr]; B[pr]; W[ef]; B[ko]; W[jp]; B[jo]; W[ip]; B[dk]; W[jm]; B[km]; W[jl]; B[lk]; W[kj]; B[kk]; W[jk]; B[ij]; W[ki]; B[in]; W[hk]; B[hj]; W[hm]; B[hn]; W[gj]; B[ih]; W[gn]; B[mi]; W[kg]; B[mg]; W[if]; B[gi]; W[fj]; B[gg]; W[gf]; B[hg]; W[hf]; B[eh]; W[dj]; B[fg]; W[ff]; B[ei]; W[ej]; B[fi]; W[ch]; B[pi]; W[qd]; B[pe]; W[qe]; B[qh]; W[qg]; B[qc]; W[rc]; B[qb]; W[ph]; B[oi]; W[qi]; B[qj]; W[rh];)

Then we have full information we need in the pattern matching—all the stones and their positions on the board. However, if we use it directly, some mistakes might occur because of the original SGF data did not record information of removed stones. In the rule of Go, if the liberty of a stone is equal to zero, then the stone must be removed from the game board. In Fig. 42.4, after some stones are removed, their positions might look empty, but their positions will not be deleted from SGF record.

To avoid these mistakes, the SGF game records must expand to many game boards. Expanding a game record could be considered as expanding one animation to hundreds of pictures. Every game board will be transformed into an independent record and skip the records of removed stones. For example, S_2 can be transformed to S_3 . S_3 will become 70 independent records (depends on the number of moves), and each of them means a board situation after a new move. Please notice that after turn 36 B[so], the four black stones B[pn], B[qm], B[qn], and B[qo] were taken, as the four strikethroughs in S_3 .

S_3 :

1. **AB[dd] [pd] [jj] [dp] [pp]** W[mp]
2. **AB[dd] [pd] [jj] [dp] [pp]** W[mp]; B[pn]
- ...

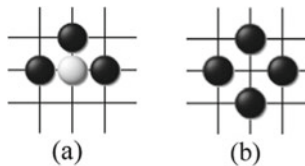


Fig. 42.4 Removing stones from the game board. **a** A white stone only has one liberty. **b** The lower black move causes the white stone to lose its liberty, resulting in the white stone being removed from the board

- 35. **AB[dd][pd][jj][dp][pp]** W[mp] B[pn]; W[fq]; B[cn]; W[fo]; B[nq]; W[mq]; B[dq]; W[ql]; B[qm]; W[pl]; B[nm]; W[pg]; B[nc]; W[np]; B[oq]; W[on]; B[om]; W[pm]; B[qn]; W[cf]; B[fc]; W[oo]; B[lm]; W[po]; B[qo]; W[qp]; B[qq]; W[rp]; B[rq]; W[ro]; B[op]; W[mn]; B[sp]; W[rm];
- 36. **AB[dd][pd][jj][dp][pp]** W[mp] B[pn]; W[fq]; B[cn]; W[fo]; B[nq]; W[mq]; B[dq]; W[ql]; B[qm]; W[pl]; B[nm]; W[pg]; B[nc]; W[np]; B[oq]; W[on]; B[om]; W[pm]; B[qn]; W[cf]; B[fc]; W[oo]; B[lm]; W[po]; B[qo]; W[qp]; B[qq]; W[rp]; B[rq]; W[ro]; B[op]; W[mn]; B[sp]; W[rm]; **B[so];**
- 37. ...

42.4 A Sequential Go Pattern Matching Algorithm

The game board and the $n \times m$ query pattern are a 19×19 and $n \times m$ arrays, respectively. Each item of the array is represented by 4 bits. Figure 42.5 shows a 4×4 query pattern. We use 4 bits to represent four Boolean state values of a position, respectively. Bit 1 is for empty on the position; bit 2 is for a black stone; bit 3 is for a white stone, and bit 4 is for the position is on the board border. Every number in the array is generated from 4 bits. For example, a number “3” in the array is represented by “0011,” and this position could be “black stone” or “empty”.

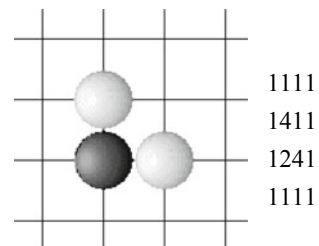
Knuth–Morris–Pratt algorithm helps us to skip unnecessary matchings. Regarding a two-dimensional pattern as a set of one-dimensional patterns, we proceed along with the following steps:

Step 1. Choose one row from the query pattern. It is one-dimensional pattern and will be suitable for the one-dimensional Knuth–Morris–Pratt algorithm. The row in the set will be ranked by their:

1. **State:** Choose the fixed state row. (Patterns may contain elements with the unfixed state. If each row pattern contains an unfixed state, then we must generate all possible row patterns.
2. **Length:** Choose the longest row pattern. If there are more than two longest row patterns have the same lengths, then choose the one that has the highest repetition number.

Step 2. Search the chose row pattern by Knuth–Morris–Pratt algorithm.

Fig. 42.5 Representation of a pattern



Step 3. If the row pattern is found, then check each item in the query pattern with the corresponding area by bit mapping: Let Q be the query pattern and A be the corresponding area. Calculate
 Let $Z = Q \text{ AND } A$
 If $Z = A$, then matched
 Else fail

Figure 42.6 shows an example. We saw some patterns on the board match the target pattern, but their surroundings do not. These patterns will be skipped immediately unless the whole query pattern is matched.

Because new patterns may only appear around the last move. If the size of the query pattern is $n \times n$, we would only scan the $(2n - 1) \times (2n - 1)$ local area. Figure 42.7 shows an example.

When the last move removes some stones, it might bring new patterns as well. Figure 42.8 shows an example. In Fig. 42.8, two black stones under the last white move marked by “x” were removed, the total variation on the board is 3×1 , so the searching area is square A . The size of square A is $(2 \times (3 - 1) + 3) \times (2 \times (3 - 1) + 1) = 7 \times 5$ when size of query pattern is 3×3 .

If the query pattern is $n \times n$, time complexity for a board is $O((2n - 1) \times (2n - 1))$. Some query pattern may be on the edge. Figure 42.9 shows an example. When searching for an edge pattern, the time complexity will be $O(2n - 1)$ (only scan along with an edge). In case some query pattern is on the corner, the time complexity is a constant. Figure 42.10 shows an example. The time complexity is a constant.

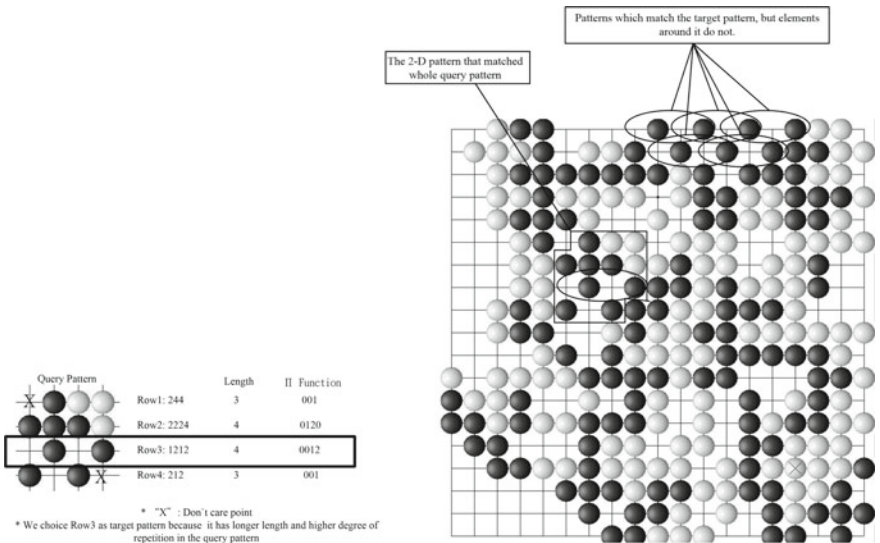


Fig. 42.6 Example for the sequential matching algorithm

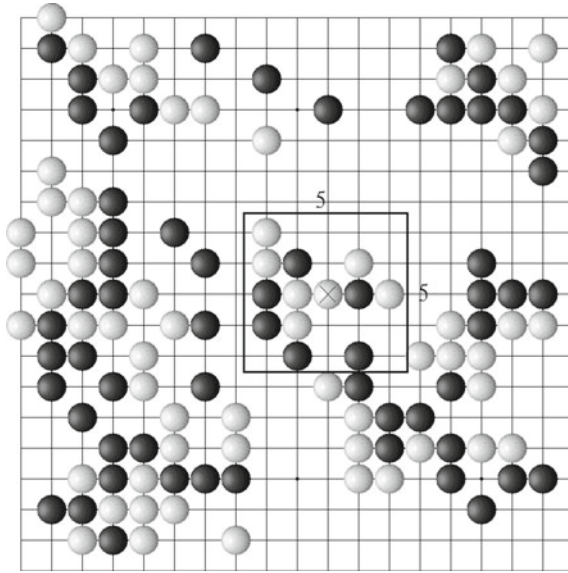


Fig. 42.7 5×5 local game board around the last move (query pattern is 3×3)

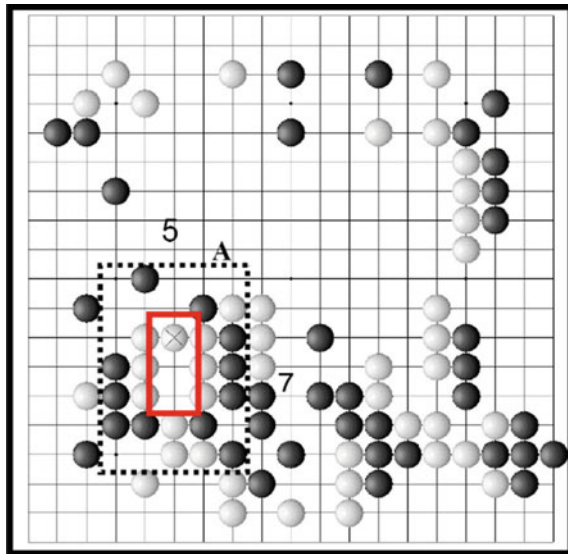


Fig. 42.8 Two black stones under the last white move marked by “X” were removed

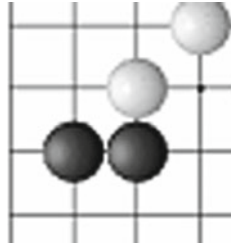


Fig. 42.9 Edge pattern

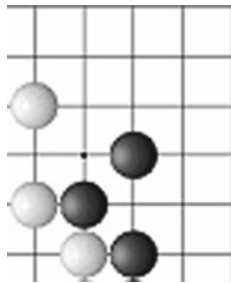


Fig. 42.10 Corner pattern

For the desired query pattern, we must consider color changing, rotation, and reflection for the query pattern (see Fig. 42.2). There are 16 related patterns for the query pattern.

The experimental result for the algorithm is as follows. The total game record is 6000, and the number of total game boards is 1, 202,201. If the query pattern is not an edge pattern, the average time is between 10 and 20 s. Figure 42.11 shows the time of $(3 \times 3, 4 \times 4, \dots, 18 \times 18, 19 \times 19)$ sizes of query patterns. According to the result, it may not satisfy the human users. The following section will describe how to use feature patterns and index structures to increase the speed.

42.5 Index Structure on the Go Game Database

To construct our index structure on the database, we use features as the indexing key. We use about 400 features which appear frequently for Go games. These features are from JIMMY Go program [19]. Some features are *jump*, *knight move*, or *diagonal* basic patterns, but most of them are complicated and have their own characteristics and meanings. These features are good indexing keys, and we can classify the database systematically. The feature patterns are integrated with a popular index structure: an *inverted list*. We use the Go pattern matching algorithm to find the

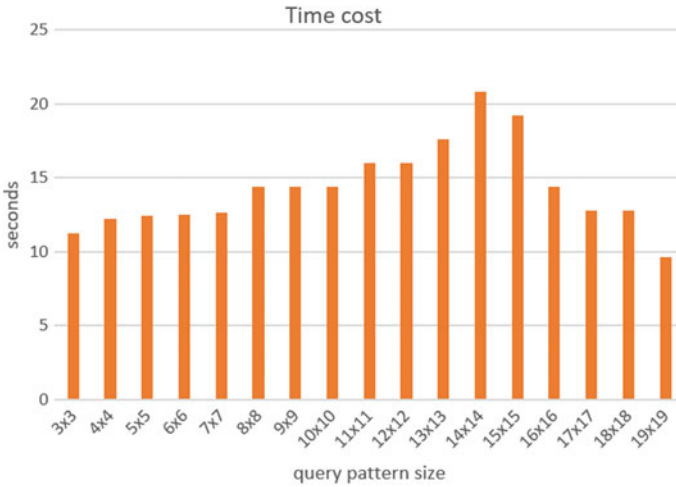


Fig. 42.11 Searching time on different query pattern sizes

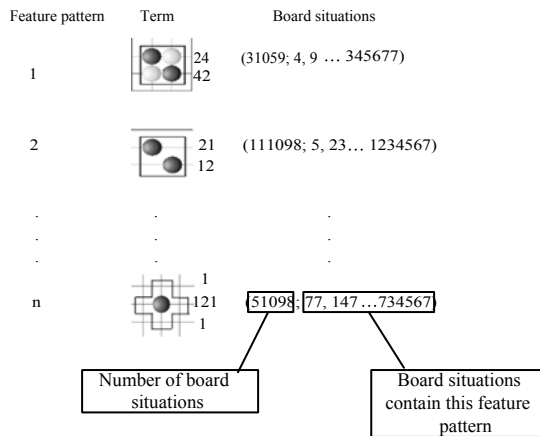


Fig. 42.12 Building inverted file using feature patterns

game boards contain the feature and build an inverted file. Figure 42.12 shows an example.

42.6 Four Applications of the Proposed Structure

We will consider more applications of an index structure in Go game records on the view of users. This topic will be discussed in four directions: **Opening**, **Joseki (A set of sequences)**, **Tesuji (A clever move)**, and **Endgame**. The following approaches

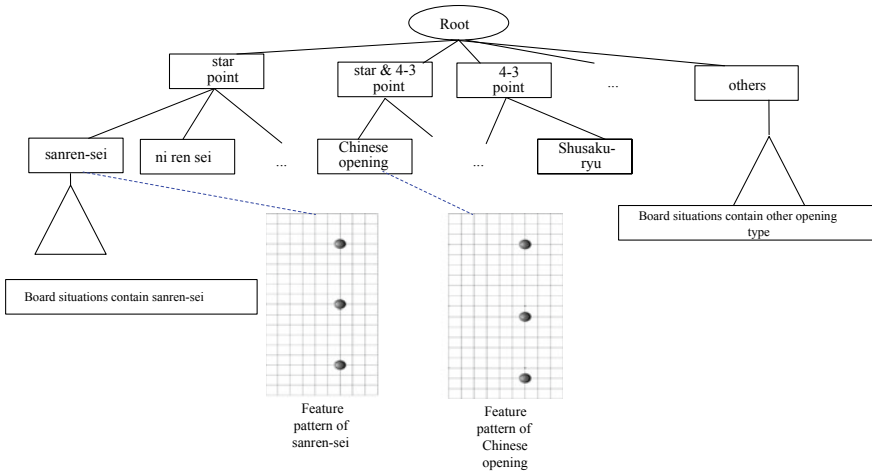


Fig. 42.13 Opening tree

are represented by a tree structure, but they are all based on the index structure in an inverted list.

42.6.1 Opening

At the opening stage, most users want to know the variations after some type of opening, such as sanrensei or Chinese opening. A type of opening can be viewed as a special Go pattern. There are two rules to be followed while constructing an opening pattern index.

1. We define the “opening stage” as “the first 30 moves of a game.” Only the first 30 game boards of every game record are considered while constructing an opening tree. Reducing a few target game boards is important for the system.
2. In the Go game, we usually care about what kind of opening pattern black player uses, while constructing the opening tree, we also follow this habit.

We consider every type of opening move sequence as a 19 × 10 feature pattern. An opening index tree and two examples of feature pattern were shown in Fig. 42.13.

42.6.2 Joseki

A joseki index tree is similar to an opening index tree. The only difference is the matching region and size of the feature patterns. Joseki moves always appear on the corners of the board. This characteristic can be used to limit the matching region in

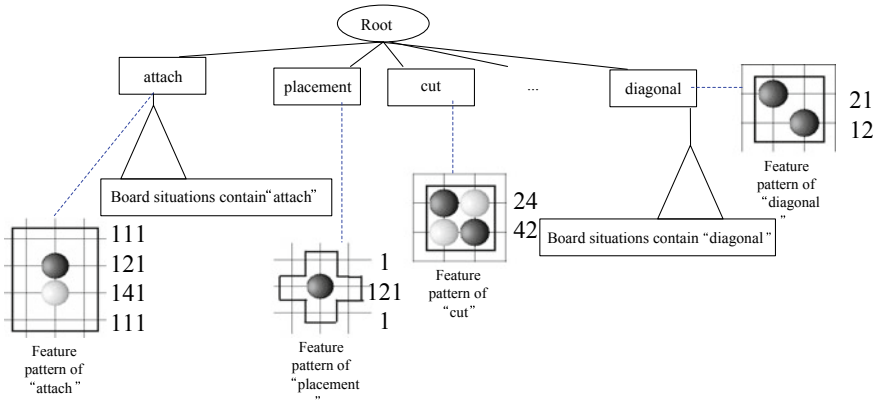


Fig. 42.14 Index tesuji tree

the four corners of the board. This constraint speeds up not only the construction of joseki tree but also the user query. We consider a joseki move as a 10 × 10 feature pattern and try to match them only on the corners of every game board.

42.6.3 Tesuji Tree

The issues of the tesuji and endgame are quite different from joseki, and the main difference is that we only want to search for the first game board that queried pattern appears. The game boards before or after them will not be considered. That is because tesuji means “a clever move” in Go game. This restriction has brought a benefit on pattern matching because we only have to collect these game boards in which feature pattern appears at their last move. This will reduce many game boards in the tesuji tree and make the matching more efficient. A tesuji index tree was shown in Fig. 42.14.

42.6.4 Endgame

The endgame index tree construction is similar to tesuji index tree. But there are more constraints in it. Most of endgame moves happened at line 2 and line 3 of the board. In Go game, there are some endgame moves happened at the center of the board, but these endgame moves are abstract. We cannot consider them in our approach. That means we will only search for the patterns appear at line 2 and line 3.

Endgame move happened at the borders of black stones and white stones. We have to check it also. We could get the information of region by calculating the effect of every black and white stone. Every black stone is given a positive “energy” value which decreases with distance; on the other hand, every white stone is given

a negative energy. We can get a general region of black stone and white stone after some operations.

42.7 Conclusion

We propose a Go pattern matching algorithm. When query patterns are smaller than 10×10 , for 1.2 million 19×19 Go boards, those records for this query pattern can be found in 15 s. We also propose a simple and efficient index structure. The index is built based on the feature patterns of the pattern database of the computer Go program JIMMY. All the significant Go patterns contain at least one of the feature patterns. This index structure can improve the speed of the pattern searching to one second, which is a satisfactory time for a query pattern. Besides, we propose four applications on the index structure, which can help users to find desired information quickly. The proposed methods are useful in a Go information retrieval system. Text query in game records is easy to implement. It can look like as text information retrieval. With proposed methods, we construct a Go information retrieval system. The system can handle both pattern and text queries. This system is useful for Go players to find the Go information in Go game records.

Acknowledgements The authors would like to thank anonymous referees for their valuable comments in improving the overall quality of this paper. This work was supported in part by the Ministry of Science and Technology of Taiwan under contract 108-2634-F-259-001 through Pervasive Artificial Intelligence Research (PAIR) Laboratories, Taiwan.

References

1. Silver, D., et al.: Mastering the game of Go with deep neural networks and tree search. *Nature* **529**(7587), 484–489 (2016)
2. Silver, D., et al.: Mastering the game of go without human knowledge. *Nature* **550**(7676), 354–359 (2017)
3. Allis, L.V.: Searching for solutions in games and artificial intelligence. Ph.D. thesis, University of Limburg, Maastricht, The Netherlands. ISBN 90-9007488-0 (1994)
4. Yen, S.-J., Chen, J.-C., Yang, T.-N., Hsu, S.-C.: Computer Chinese chess. *ICGA J.* **27**(1), 3–18 (2004)
5. Bouzy, B., Cazenave, T.: Computer Go: an AI oriented survey. *Artif. Intell.* **132**, 39–103 (2001)
6. Iida, H., Sakuta, M., Rollason, J.: Computer shogi. *Artif. Intell.* **134**, 121–144 (2002)
7. Yen, S.J., Chen, Y.L., Lin, H.I.: Scaffolding learning for the Novice Players of Go. In: 2019 International Conference of Innovative Technologies and Learning (ICITL 2019), LNCS 11937 (2019)
8. Tian, Y., Zhu, Y.: Better computer go player with neural network and long-term prediction. In: International Conference on Learning Representations (2016)
9. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. *CoRR* **1409**, 4842 (2014)

10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Adv. Neural. Inf. Process. Syst.* **25**, 1097–1105 (2012)
11. Boon, M.: A pattern matcher for Goliath. *Comput. Go* **13**, 12–24 (1990)
12. Mueller, M.: Pattern matching in explorer. In: *Proceedings of the Game Playing System Workshop*, pp. 1–3, Tokyo, Japan (1991)
13. Nakamura, T., Kajiyama, T.: Automatic Acquisition of Move Sequence Patterns from Encoded Strings of Go Moves. Department of Artificial Intelligence, Kyushu Institute of Technology (2002)
14. Wu, I.-C., Wu, T.-R., Liu, A.-J., Guei, H., Wei, T.: On strength adjustment for MCTS-based programs. In: *Thirty-Third AAAI Conference on Artificial Intelligence* (2019)
15. Ikeda, K., Viennot, S.: Production of various strategies and position control for Monte-Carlo go entertaining human players. In: *2013 IEEE Conference on Computational Intelligence in Games (CIG)* (2013)
16. Werf, E.C.D., van der Herik, H.J., van den Uiterwijk, J.W.H.M.: Learning to score final positions in the game of Go. In *Advances in Computer Games: Many Games, Many Challenges*, pp. 143–158 (2003)
17. Werf, E.C.D., van der Winands, M.H.M., Herik, H., Jaap van den., Uiterwijk, J.W.H.M.: Learning to predict life and death from go game records. *Inf. Sci.* **175**(4), 258–272 (2005)
18. Kierulf, A.: Smart game board: a workbench for game-playing programs, with go and othello as case studies. Ph.D. thesis, ETH Zürich. Thesis advisor: Prof. J. Nievergelt (1990)
19. Yen, S.-J.: Design and implementation of a computer GO program JIMMY. Ph.d. thesis, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan (1999)